ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Genova

INFN/AE-04/08
21 Giugno 2004

A GOODNESS-OF-FIT STATISTICAL TOOLKIT

G.A.P. Cirrone[1], S. Donadio[2], S. Guatelli[2], A. Mantero[2], B. Mascialino[2], S. Parlati[3], A. Pfeiffer[4], M.G. Pia[2], A. Ribon[4], P. Viarengo[5]

[1)]*INFN, LNS, Sezione di Catania,Univ. di Catania, I-95100 Catania, Italy*
[2)]*INFN, Sezione di Genova, Dipartimento di Fisica, Univ .di Genova, I-16131Genova,Italy*
[3)]*INFN, LNGS, Sezione de L'Aquila,I-67010 L'Aquila, Italy*
[4)]*CERN,CH-1122, Geneve, Switzerland*
[5)]*IST, Istituto Nazionale per la Ricerca sul Cancro, I-16132 Genova, Italy*

Abstract

Statistical methods play a significant role through out the life-cycle of physics experiments, being an essential component of physics analysis. The present project in progress aims to developan object-oriented software Toolkit for statistical data analysis. The Toolkit contains a variety of Goodness-of-Fittests, from Chi-squared to Kolmogorov-Smirnov, to less known, but generally much more powerful tests such as Anderson-Darling, Goodman, Fisz-Cramer-vonMises, Kuiper. Thanks to the component-baseddesign 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. In this paper we describe the statistical details of the algorithms and the computational features of the Toolkit. With the aim of showing the consistency between the code and the mathematical features of the algorithms, we describe the results we obtained reproducing by means of the Toolkit a couple of Goodness-of-Fittesting examples of relevance in statistics literature.

PACS: 11.30.Er; 13.20.Eb; 13.20Jf; 29.40.Gx; 29.40.Vj

Paper published on IEEE - Transactions on Nuclear Science (2004)

Published by ***SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 Introduction

Statistical comparison of distributions is an essential section of data analysis in any field and in particular in high energy physics experiments. In spite of this, only a few basic tools for statistical analysis were available in the public domain FORTRAN libraries for high energy physics, such as CERN Libraries (CERNLIB) [1] and PAW [2]. Nowadays the situation is unchanged even among the libraries for data analysis of the new generation, such as for istance Anaphe [3], JAS [4], Open Scientist [5] and Root [6].
For these reasons a new project was launched to build an open-source and up-to-date object-oriented Toolkit for high energy physics experiments and other applications.
In this paper we will focus our attention on a specific component of the statistical Toolkit that we developed. The Statistical Comparison component of the Toolkit provides some Goodness-of-Fit (**GoF**) algorithms for the comparison of data distributions in a variety of use cases typical for physics experiments, as:

- regression testing (in various phases of the software life-cycle),
- validation of simulation through comparison with experimental data,
- comparison of expected versus reconstructed distributions,
- comparison of data from different sources, such as different sets of experimental data, or experimental with respect to theoretical distributions.

The **GoF** Toolkit gathers together some of the most important **GoF** tests available in literature. Its aim is to provide a wide set of algorithms in order to test the compatibility of the distributions of two variables.

2 The Goodness of Fit Statistical Toolkit

The **GoF** 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.
Specifically, consider two real-valued random variables X and Y, and let $(x_1, \ldots, x_n)$ be a sample of independent and identically distributed observations with cumulative distributions (*cdf*) $F(x) = P[X \leq x]$, and $G(x) = P[Y \leq x]$. If X is a discrete random variable, then the *cdf* of X will be discontinuous at the points x_i and constant in between, if X is a

continuous random variable, the *cdf* of X will be continuous. If we consider that F and G may be continuous or discrete, the **GoF** tests consist in testing the null hypothesis

$$H_0 : F = G \tag{1}$$

against the alternative hypothesis

$$H_1 : F \neq G. \tag{2}$$

Under the assumption H_0, the fraction of wrongly rejected experiments is typically fixed to a few percent. In all the tests presented below, H_0 is rejected when a pre-defined test statistics is large with respect to some conventional value.
The tests are classified in distribution-dependent (parametric tests) and distribution-free tests (non parametric tests). In the case of weak, very general assumptions, the latter are in principle preferable, because they can be adapted to arbitrary distributions. Non parametric tests can be easily performed when the comparison concerns one dimensional distributions, although some of them can be extended to two or three dimensions.
In most high energy physics experiments, H_0 is a non parametric hypothesis, so testing H_0 requires a distribution-free procedure.
In the following section we describe in detail the statistical features of the **GoF** tests. It must be pointed out that it is important to have almost all the possible **GoF** tests together in the **GoF** Toolkit, as different physics problems could demand different statistics solution, depending on the distributions type or shape. In fact, we distinguish between tests suitable for binned data and binning free tests for unbinned data.
We remark that **GoF** tests are primarily included in classical statistical inferential methods, because there is no unanimous consensus on the correct approach to Bayesian model checking. However, **GoF** tests are closely related to hypothesis testing of models, and this argument belongs peculiarly to Bayesian statistics since many studies demonstrate that p-values should better be replaced by a Bayes-factor [7], [8]. Future extensions and developments of the **GoF** Toolkit may include some treatment with Bayesian methods.

3 Non parametric tests

3.1 One Sample tests: Comparing data to a theoretical distribution

3.1.1 Chi-squared test

The **Chi-squared** test verifies the adaptation of the experimental data to a certain theoretical distribution.

If B represents the number of bins, the test statistics X_T^2 (T means test) is

$$X_T^2 = \sum_{i=1}^{B} \frac{(o_i - e_i)^2}{\sigma_i^2} \tag{3}$$

where o_i is the i-th observed data, e_i the expected value and σ_i^2 the variance. The expected value of X_T^2 is

$$E(X_T^2) = B. \tag{4}$$

In the Gaussian approximation, the test statistics follows a χ^2 distribution with B degrees of freedom.

If we have a histogram, where the random variable N_i (*count* in each class) is a total of N events distributed according to a multinomial distribution among the B bins with probabilities p_i, we use **Pearson**'s definition:

$$X_T^2 = \sum_{i=1}^{B} \frac{(N_i - Np_i)^2}{Np_i} \tag{5}$$

which again follows asymptotically a χ^2 distribution, this time with $(B-1)$ degrees of freedom. The reduced number of degrees of freedom (*ndf*) is due to the constraint $\sum_i N_i = N$.

If some parameters of the model (e.g. Binomial, Poisson, Normal, ...) are estimated from data, we may simply reduce the *ndf*.

When Pearson's χ^2 test is applied to unbinned distributions, the analyst must group data into classes before performing the comparison, but the counting of theoretical frequencies in each class must be sufficiently large (*i.e.* $N_i \geq 5$).

3.1.2 Tests based on maximum distance

A common alternative to the Chi-squared tests includes tests based on the Empirical Distribution Function (*edf*) definition. Such tests are mostly used for continuous distributions and are in general more powerful than the Chi-squared test [9], as they do not require regrouping.

Consider the order statistics

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)} \tag{6}$$

and the *edf* $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$, where $I(\cdot)$ is the indicator function, jumping from $\frac{(i-1)}{n}$ to $\frac{i}{n}$ at the point $X_{(i)}$ and being constant except for the jumps at the order statistics. $F_n(x)$ must be compared with the distribution $F(x)$, defined in the null hypothesis H_0. The maximum positive (negative) deviation of $F_n(x)$ from $F(x)$, $D_{n+}(D_{n-})$,

can be used as a test statistics. The one sided **Kolmogorov** [10] test statistics is given by the maximum difference

$$D_{n+} = sup_x(|F_n(x) - F(x)|). \tag{7}$$

Kolmogorovs test is used to determine whether an empirical distribution differs from a theoretical distribution. The probability distribution of the test statistics (7), given that the null hypothesis of equality of distributions is true, does not depend on what the hypothesized distribution F is, as long as it is continuous. This is the most interesting case, dealing with non parametric testing. It must be stressed that the one sample Kolmogorov test is not very useful in practice because it requires a simple null hypothesis, as the distribution must be *completely specified* with *all parameters known*. In the case that some parameters are estimated from data, the critical region of test is no longer valid. In fact, for composite hypothesis, the expression

$$sup_x|F_n(x) - F(x, \theta)| \tag{8}$$

is no more a test statistics because it depends on unspecified parameters. When the parameters are estimated from the data, for the Kolmogorov test we obtain an estimated $\hat{D}_{n+}$, whose sampling distribution is different from the one of D_{n+}. This test is conservative (i.e. tends to privilege the acceptance of the null hypothesis) if we use the usual critical values. However, from a theoretical point of view, it is possible to approximate the null distribution of the test statistics by means of bootstrap methods: the idea of using resampling techniques to obtain critical values for Kolmogorov type statistics has been used by Romano [12] and Praestgaard [13] among the others. Some limitations of this test are concerned with the fact that it tends to be more sensitive near the center of the distribution with respect to the tails.

In **Kuipers** test we use the maximum deviation above and below of the two distribution functions

$$V_n = D_{n+} + D_{n-} \tag{9}$$

as a test statistics. Kuipers test is useful for unbinned cyclic observations as well as for cyclic transformation of the independent variable because the value of V_n does not depend on the choice of origin [14]. Thanks to its definition, the test is as sensitive to the tails as to the median of the distributions [15].

3.1.3 Tests based on quadratic distance

The Anderson-Darling family of tests measures the integrated quadratic deviation of the two *edfs* suitably weighted by a weighting function $\psi(F(x))$:

$$Q_n = n \int_{-\infty}^{\infty} [F(x) - F_n(x)]^2 \psi(F(x)) dF(x). \tag{10}$$

With

$$\psi_{CvM}(F(x)) = 1 \tag{11}$$

we get the **Cramer-von Mises** test, that can be performed on unbinned data and is satisfactory for symmetric and right-skewed distributions [16], [17].
If in equation (10) we use:

$$\psi_{AD}(F(x)) = [F(x)(1 - F(x))]^{-1} \tag{12}$$

we obtain the **Anderson-Darling** [AD] statistics A^2 [18]. This test can be performed both on binned and unbinned data and it gives more weight to the tails than Kolmogorovs test. As Kuipers test, the AD test provides equal sensitivity at the tails as at the median, but it does not guarantee cyclic invariance. A recent paper by Aksenov and Savageau [19] states that this test is suitable for any kind of distribution, independently on its particular skewness. However, the one sample AD test is only available for a few specific theoretical distributions.

3.2 Multi sample tests: Comparing two distributions

3.2.1 Tests based on maximum distance

The **Kolmogorov-Smirnov** [KS] test derives from Kolmogorovs test and represents an extension of the test statistics (7). This test was introduced by Smirnov [20], [21] and uses the test statistics

$$D_{n,m} = sup_x |F_n(x) - G_m(x)| \tag{13}$$

where $F_n(x)$ and $G_m(x)$ are two usual *edfs* associated with the X and Y random variables, respectively. This test is used to determine whether two data-samples *edfs* coming from the same theoretical distribution are different. Note that as for Kolmogorovs test it is not specified which is the common distribution, so KS is distribution-free under H_0 when the observations are independent and identically distributed with a continuous distribution. KS exact and limiting distribution is not standard (*i.e.* this test does not have usual limiting distribution) and tables for the exact distributions are only available for a limited number of sample sizes [22]; thus, this test is usually performed with the help of

tables based on asymptotic distributions. It must be noticed that the KS test is no longer distribution-free when F_n and G_m are discrete distributions.

Goodman [23] test statistics derives from KS test, representing the same test in a Chi-squared approximation. Under Goodmans approximation, the KS exact test statistics $D_{n,m}$ can be approximated with a Chi-squared test statistics, which follows a χ^2 distribution with 2 degrees of freedom

$$\chi^2 = 4D^2_{n,m}\frac{nm}{(n+m)}. \tag{14}$$

Kuipers statistics is still valid also when the comparison involves two samples; again its test statistics $V_{n,m}$ involves the signed deviations between the two *edfs*

$$V_{n,m} = sup_x[F_n(x) - G_m(x)] + sup_x[G_m(x) - F_n(x)]. \tag{15}$$

3.2.2 Tests based on quadratic distance

The two sample **Fisz-Cramer-von Mises** [FCM] [27] test derives from (10) with the condition (11). The test statistics is defined as

$$t_{n,m} = \frac{nm}{(n+m)^2}\sum_{i=1}^{n}[F_n(x_i) - G_m(x_i)]^2. \tag{16}$$

The FCM test is based on the assumption that the *edfs* F_n and G_m are unbinned. In case of binned distributions, this test is likely to be conservative. In both cases, the test can be applied on both symmetric and right-skewed distributions. The test is distribution-free under H_0 with F_n and G_m continuous. Again, the exact and limiting null distributions of FCM are not standard. The Anderson [24] and Burr [25] tables can be used for the exact distribution in the case of small sample sizes $(n + m \leq 17)$. Otherwise, a table of the asymptotic distribution is available from Anderson and Darling [26], in their paper concerning the Cramer-von Mises statistics; this one has been included in the **GoF** Toolkit.

Finally, we have the **k-sample** Anderson-Darling's test [28], deriving from (10) under the condition (12), which is an approximation for the case of k samples. The specific case of AD two-samples test statistics $A^2_{n,m}$ is contained in the **GoF** Toolkit. The AD test solves some of the limitation of the KS test. In addition, the AD test is more powerful than the KS one since it makes specific use of the underlying cumulative distribution.

3.3 Power of the statistical tests

We would like to remark that a test is considered powerful if the probability for accepting H_0 when H_0 is wrong is low. Actually, it must be stressed that with a non parametrical

set of tests a *proper* evaluation of their power cannot be quantified without specifying the alternatives. In general, the Chi-squared test is the least powerful one because of information loss due to data grouping (binning). On the other hand, all the tests based on the maximum 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.

4 The Design of the GoF Toolkit

The **GoF** Toolkit matches a sophisticated statistical data treatment with the most advanced computing techniques, such as object-oriented technology with the use of design patterns and generic programming. The system has been developed following a rigorous software process (*Unified Software Development Process* [29]), mapped onto the **ISO 15504** guidelines [30]. According to such an approach, the life-cycle of the software itself is iterative-incremental, every iteration representing an evolution, an improvement, an extension in comparison with the previous one. It must be underlined that this iterative software development allows an increasing understanding of the problem domain through successive refinements. The code of the **GoF** Toolkit has been implemented in C++.

4.1 Core Comparison Component

The project adopts a solid architectural approach, its design specifying the roles and the responsibilities of each component. The solid architecture permits at the same time both maintainability over a large time and extensibility, accommodating in this way future evolutions of the user requirements.
The architectural approach adopted is component-based: in this way it facilitates the reuse of the **GoF** Toolkit as well as its integration in other data analysis frameworks. The design makes use of both object-oriented techniques and of generic programming.
Fig. 1 represents the core component of the **GoF** Toolkit. Its main features can be 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 test algorithm.

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, that is the probability that the test statistic has a value at least as extreme as that observed, assuming the null hypothesis is true).
Fig. 2 details all the algorithms implemented up to now: it must be noticed that every algorithm is specialised for *only one* kind of distribution (binned or unbinned). Algorithms implemented for binned distributions are the Chi-squared, FCM and AD tests. Algorithms for unbinned data are Goodmans, Kolmogorovs, FCM, AD and Kuipers tests. The class *ComparisonAlgorithm* is templated on the specific distribution (binned or unbinned); its responsibility is the computation of the deviation, *i.e.* the distance, between two distributions, on the basis of the statistical test selected by the user.
The p-value computation is instead performed through an abstract class *StatisticsQualityChecker*. Thanks to the adoption of a *Strategy Pattern* [31], the algorithms are encapsulated into objects and thus made interchangeable. The *ComparisonAlgorithm* class maintains a reference to a *StatisticsQualityChecker*. When the user selects an algorithm to perform a specific **GoF** test, the *ComparisonAlgorithm* class forwards this responsibility to its *StatisticsQualityChecker* object.
Moreover, the object-oriented design allows for an easy extension of the **GoF** Toolkit to new algorithms. New statistical tests can be inserted in the **GoF** Toolkit without interfering with the existing code.

4.2 User Layer Component

The object-oriented techniques adopted together with the AIDA (*Abstract Interfaces for Data Analysis*) [32], that are recognised as a standard in high energy physics software domain, are able to shield the user from the complexity of both the design of the core component and the computational aspects of the mathematical algorithms implemented. All the user has to do is to choose the most appropriate algorithm and to run the comparison. In practice he/she activates the statistical test he/she wants to perform writing only *one* line of code, as evidenced in Fig. 3 in the specific case of a Chi-squared test. This implies that the user does not need to know the 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.
Moreover, the user is somehow guided in his analysis as this specific design of the **GoF** Toolkit prevents him/her from applying a wrong test to his/her distributions. As an example, if the user tries to apply the CVM algorithm to binned data, the **GoF** will not run the comparison, as the class *CramerVonMisesUnbinnedComparisonAlgorithm* is defined to work only on unbinned distributions. Thanks to this architecture, the user can access only

those algorithms whose applicability conditions fit the kind of distribution he/she deals with.

5 Code validation

On the basis of the rigorous software process that the project adopted, the **GoF** Toolkit code has undergone a test process. Unit tests covered each class of the design, such as the verification of all the distance and probability calculations.
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.
In this paper we show only a couple of examples that were extracted from two reference statistics books [33], [34]. This validation is simply 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.

5.1 Validation of GoF Tests for Binned Distributions

The first example of code validation comes from Piccolo's book [33] and concerns a sample consisting of 294 people. Every subject's birth and death day has been recorded. The aim of the study is to test whether the birth distribution *X* is the same as the death *Y* one. The hypothesis we are going to test is the following:

$$H_0 : F_X(\omega) = G_Y(\omega) \tag{17}$$

for every ω, against the alternative:

$$H_1 : F_X(\omega) \neq G_Y(\omega) \tag{18}$$

for at least one ω. Fig. 4 shows the *X* and *Y* original distributions.
Piccolo himself states that these two variables are continuous, but he performs the comparison after binning the data, grouping births and deaths in months. In spite of this manipulation, he applies to this example the KS test, according to the justification that *in practice* it is better to group data if one has to make the comparison *by hand*.
The KS test statistics he computes according to equation (13) is $D_{ref} = 0.08163$. He compares it with the KS critical value ($D_{crit}(0.05) = 0.11201$ if $n = 294$ and $m = 294$). According to this result he concludes that there is no evidence for rejecting the null hypothesis and, for this reason, the two distributions are identical from a statistical point of view.

Test	Distance	p-value
Chi-squared	$\chi^2 = 15.8$	$p = 0.20$
Anderson-Darling	$A^2 = 0.09$	$p > 0.05$
Fisz-Cramer-von Mises	$t = 2.54$	$p > 0.05$
Kolmogorov-Smirnov*	$D = 0.082$	$p = 1$
Goodman*	$\chi^2 = 3.91$	$p = 0.14$

Table 1: Statistical results for binned distributions tests code validation: the example comes from Piccolo's book [33]. Those tests evidenced with * have been included in the comparison only to compare their numerical result with Piccolo's solution. All the algorithms lead exactly to the results computed by Piccolo.

The statistical comparisons we performed running this example with the **GoF** Toolkit are shown in Table 5.1 for the Chi-squared, FCM and AD tests. Of course, the binned nature of data led us to run it with Chi-squared, FCM and AD tests. *Only* with the aim of checking if we could reproduce the same numerical result of Piccolo, we applied an *ad hoc* procedure to force the application of the KS test to binned data. The Goodman approximation of the KS test was performed as well.
In any test we obtain the same result of Piccolo and, in particular, the **GoF** Toolkit reproduces exactly the numerical result of the test statistics computed by the author ($D_{Toolkit} = 0.0816327$).

5.2 Validation of GoF Tests for Unbinned Distributions

The second example is specific for unbinned data and comes from Landenna's book [34], but deals with a physiology experiment. A collective made up by 20 subjects has been studied by Delse and Feather [35], with the aim of evaluating if a subject is able to control his own salivation process while he tries to increase or reduce it. Subjects were divided into two equal groups. The subjects of the first group had to try to modify their own salivation flux according to the direction (right or left) of a light source, while the subjects of the second one had to make the same things without any light stimulus. The aim of the study was to verify if the salivary flux distribution *X* of the sample subjected to the light stimulus (S) is different from the *Y* one coming from the group with no light stimulus (NS).
Again, the hypothesis we are going to test is the following:

$$H_0 : F_X(\omega) = G_Y(\omega) \tag{19}$$

Test	Distance	p-value
Goodman	$\chi^2 = 7.2$	$p = 0.03$
Kolmogorov-Smirnov	$D = 0.6$	$p = 0.03$
Fisz-Cramer-von Mises	$t = 0.485$	$p < 0.05$
Anderson-Darling	$A^2 = 37.9$	$p < 0.05$

Table 2: Statistical results of the unbinned distributions tests code validation: the example comes from Landenna's book [34] and it explains Delse and Feather's [35] physiology experiment. All the algorithms lead to the same results computed by Landenna.

for every ω, against the alternative:

$$H_1 : F_X(\omega) \neq G_Y(\omega) \tag{20}$$

for at least one ω.
Delse and Feather *edfs* are shown in Fig. 5. Landenna performs the statistical comparison between S and NS with two different algorithms: KS and FCM tests. The result of the KS test is that the maximum deviation between the S and NS distributions is $D_{ref} = 0.6$. Landenna compares this test statistics with the KS critical values available on statistical tables and he finds that, in the case of $n = 10$ and $m = 10$, $D_{crit}(0.05) = 0.6$. So there is evidence for rejecting the null hypothesis and this means that the two distributions are significantly different.
The FCM test leads the author to the same statistical result, the test statistics being equal to $t_{ref} = 0.485$, while the critical values come again from the tables included in his book: $t_{crit}(0.05) = 0.475$ if $n = 10$ and $m = 10$. Again, the test statistics falls in the null hypothesis rejection area and the author concludes that also this test states that the two distributions are significantly different.
The comparisons we produced by means of the **GoF** Toolkit involve the algorithms specialised for unbinned distributions: Goodman, KS, FCM and AD tests. **GoF** Toolkit statistics testing results are shown in Table 5.2. In all cases we obtain the same results as reported by Landenna and, in particular, with both KS and FCM tests we could reproduce exactly the same numerical results reported by the author.

6 Conclusion

The **GoF** Toolkit is an easy to use, up-to-date and versatile tool for data comparison in physics analysis. It is the first statistical software system providing such a variety of sophisticated and powerful algorithms in high energy physics.

The component-based design uses object-oriented techniques together with generic programming. The adoption of AIDA for the user layer decouples the usage of the **GoF** Toolkit from any concrete analysis system the user may have adopted in his/her analysis. The code is open-source and can be downloaded from the web together with user and software process documentation [36].
As it was underlined in the section describing the power of the **GoF** tests, none of the tests included in the system is optimum for *every* case. The user has the freedom to choose the **GoF** test most appropriate for his/her comparison among those available in the Toolkit. He/She is guided in this selection by the design of the Toolkit itself, that prevents the use of an incompatible type of **GoF** test for a given distribution.
Thanks to the great variety of its sophisticated and powerful statistical tests, the **GoF** Toolkit has been adopted by various projects, involving 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. Their fields of application are the following:

1. **General Purpose Software:** GEANT4 [37] adopts the **GoF** Toolkit for the microscopic validation of its physics models [38].

2. **Astrophysics:** The Bepi Colombo mission of the European Space Agency uses it for comparisons of experimental data from a test beam with simulations for the study of the detector design [39].

3. **Medical physics:** The CATANA project, treating patients affected by uveal melanoma with hadrontherapy, uses the **GoF** Toolkit in order to make the comparison of physical quantities of interest (as the Bragg peak or isodose distributions) for the optimisation of the design of the treatment beam line [40].

For all the features described, the **GoF** Toolkit represents a step forward in the quality of data analysis in high energy physics and could be easily used by other experimental software frameworks.

Acknowledgment

This work was partly supported by the European Space Agency under Contract No. 16339/02/NL/FM.
The authors would also like to thank Fred James, CERN, and Louis Lyons, University of Oxford, for the fruitful discussions concerning the themes outlined in this article.

References

[1] J. Shiers, “CERNLIB - CERN Program Library Short Writeups”, CERN Geneva, 1996.

[2] R. Bock, R. Brun, O. Couet, J.C. Marin, R. Nierhaus, L. Pape, et al., “PAW-Towards a Physics Analysis Workstation”, *Computer Physics Communications*, vol. 45, pp. 181-190, 1987.

[3] O. Couet, B. Ferrero-Merlino, Z. Molnár, J.T.Moscicki, A.Pfeiffer, M.Sang, “ANAPHE-OO Libraries and Tools for Data Analysis”, Proceedings of the CHEP 2001 Conference, Beijing, China, Spetember 2001.

[4] T. Johnson, V. Serbo, M. Turri, M. Donszelmann, J. Perl, “JAS3-A general purpose Data Analysis Framework for High Energy Physics and beyond”, Proceedings of the IEEE Nuclear Science Symposium, Portland, Oregon, October 2003.

[5] http://www.lal.in2p3.fr/OpenScientist

[6] R. Brun, F. Rademakers “ROOT, An object-oriented Data Analysis Framework”, *NIM - Section A*, vol.389, pp.81-86, 1997.

[7] J.O. Berger, T. Sellke, “Testing a point null hypothesis: the irreconciliability of p-values and evidence”, *Journal of American Statistics Associated*, vol.82, pp. 112-122, 1987.

[8] L. Piccinato, “Il fattore di Bayes come strumento pratico di statistica applicata”, *Università degli studi di Roma La Sapienza*, series B, vol.1, 1997.

[9] M.A. Stephens, “Introduction to Kolmogorov (1933) on the empirical determination of a distribution”, S. Kotz, N.L. Johnson, “Breakthrought in Statistics”, vol. 2, Ed. New York: Springer Verlag, 1992.

[10] A.N. Kolmogorov, “Sulla determinazione empirica di una legge di distribuzione”, *Giornale dell’istituto italiano degli attuari*, vol.4, pp. 83-91, 1933.

[11] W.J. Conover, “Practical Nonparametric Statistics”, Ed. New York: John Wiley & Sons, 1991.

[12] J.P. Romano, “A Bootstrap Revival of Some Nonparamentric Distance Tests”, *Journal of the American Statistical Association*, vol. 83, pp. 698-708, 1988.

[13] J.T. Praestgaard, "Permutation and Bootstrap Kolmogorov-Smirnov Tests for the Equality of Two Distributions", *Scandinavian Journal of Statistics*, vol. 22, pp. 305-322, 1995.

[14] N.H. Kuiper, "Tests concerning random points on a circle", *Proc. Koninkl. Neder. Akad. van Wettensch. A*, vol.63, pp. 38-47, 1960.

[15] http://encyclopedia.thefreedictionary.com/Kuiper's

[16] H. Cramer, "On the composition of elementary errors", *Second paper: statistical applications, Skand. Aktuarietidskrift*, vol.11, pp. 13-74, pp. 141-180, 1928.

[17] R. von Mises, "Wahrscheinliehkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik", *Leipzig: F.Duticke*, 1931.

[18] T.W. Anderson, D.A. Darling, "A test of Goodness of Fit", *Journal of the American Statistical Association*, vol.49, pp. 765-769, 1954.

[19] T.W. 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", 2002, in press.

[20] N.V. Smirnov, "Sur les écarts de la courbe de distribution empirique (Russian/French summary)", *Matematiceskĭi Sbornik N.S.*, vol.6, pp. 3-26, 1939.

[21] N.V. Smirnov, "Table for estimating the Goodness-of-Fit of empirical distributions", *Annals of Matemathical Statistics*, vol.19, pp. 279-281, 1948.

[22] D.A. Darling, "The Kolmogorov-Smirnov, Cramer-von Mises tests", *Annals of Mathematical Statistics*, vol. 28, pp. 823-838, 1957.

[23] L.A. Goodman, "Kolmogorov-Smirnov tests for psychological research", *Psychological Bullettin*, vol. 51, pp. 160-168, 1954.

[24] T.W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion", *Annals of Mathematical Statistics*, vol. 33, pp. 1148-1159, 1962.

[25] E.J. Burr, "Distribution of two-sample Cramer-von Mises criterion for small equal sample", *Annals of Mathematical Statistics*, vol. 33, pp. 95-101, 1963.

[26] T.W. Anderson, D.A. Darling, "Asymptotic theory of certain Goodness of fit criteria based on stochastic processes", *Annals of Mathematical Statistics*, vol. 23, pp. 193-212, 1952.

[27] M. Fisz, “On a result by M.Rosenblatt concerning the von Mises-Smirnov test”, *Annals of Mathematical Statistics*, vol. 31, pp. 427-429, 1960.

[28] A. Heckert (2003, April 4). NIST, Available: http://www.itl.nist.gov/div898/software/dataplot/refman1/

[29] I. Jacobson, G. Booch, J. Rumbaugh, “The Unified Software Development Process”, 1^{st} ed., Ed. Addison-Wesley, 1999.

[30] ISO/IEC Joint Technical Committee 1 (JTC1), “ISO/IEC DTR 15504-5 Part 5: an assessment model and indicator guidance”, Ed. Jean Martin Simon.

[31] E. Gamma, R. Helm, R. Johnson, J. Vlissides, “Design Patterns”, 1^{st} ed., Ed. Addison Wesley Professional Computing Series, 1994.

[32] G. Barrand, P. Binko, M. Donszelmann, A. Johnson, A. Pfeiffer, “Abstract Interfaces for Data Analysis - Component Architecture for Data Analysis Tools”, proceesings of CHEP, Beijing, September 2001, Science Press, pp. 215-218.

[33] D. Piccolo, “Statistica”, 1^{st} ed., Ed. Il Mulino, Bologna Italy, 1998, pp. 711-712.

[34] G. Landenna, D. Marasini, “Metodi statistici non parametrici”, 1^{st} ed., Ed. Il Mulino, Bologna Italy, 1990, pp. 318-337.

[35] F.C. Delse, B.W. Feather, “The effect of augmented sensory feedback on the control of salivation”, *PsychoPhysiology*, vol. 5, 1968, pp. 15-21.

[36] http://www.ge.infn.it/geant4/analysis/HEPstatistics/index.html

[37] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, et al., “GEANT4a simulation toolkit”, *NIM - Section A*, vol. 506 (3), 2003, pp. 250-303.

[38] G.A.P. Cirrone, G. Cuttone, S. Donadio, V. Grischine, S. Guatelli, P. Gumplinger, et al., “Precision Validation of Geant4 Electromagnetic Physics”, Proceedings of IEEE Nuclear Science Symposium, Portland, Oregon, October 2003.

[39] A. Mantero, B. Bavdaz, A. Owens, T. Peacock, M.G. Pia, “Simulation of X-ray Fluorescence and Application to Planetary Astrophysics”, Proceedings of IEEE Nuclear Science Symposium, Portland, Oregon, October 2003.

[40] G.A.P. Cirrone, G. Cuttone, S. Guatelli, S. Lo Nigro, B. Mascialino, M.G. Pia, et al., “Implementation of a New Monte Carlo Simulation Tool for the Development

of a Proton Therapy Beam Line and Verification of the Related Dose Distributions”, Proceedings of IEEE Nuclear Science Symposium, Portland, Oregon, October 2003.

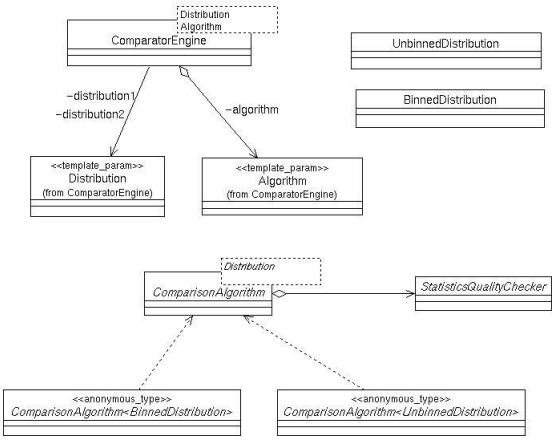

Figure 1: Design of the core component of the statistical Toolkit: the *ComparatorEngine* is responsible of the whole statistical comparison process.

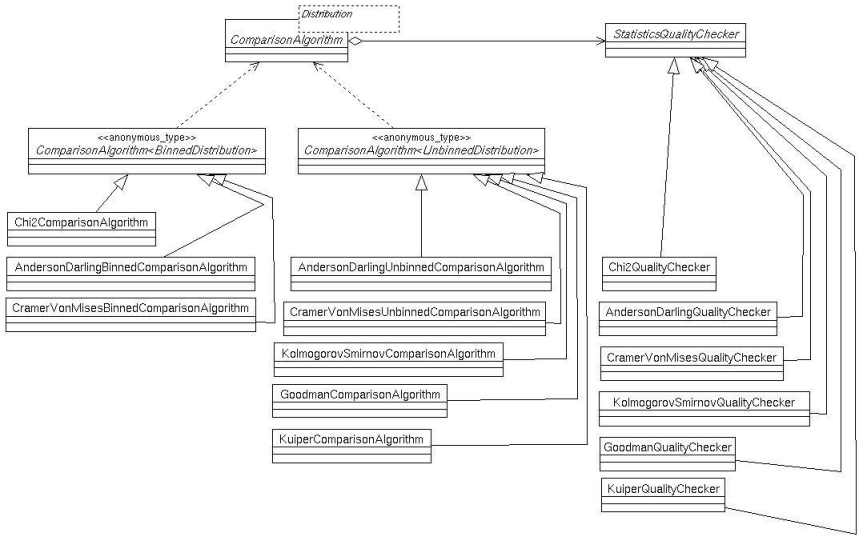

Figure 2: Detail of the statistical Toolkit design: algorithms implemented for binned (Chi-squared, Fisz-Cramer-von Mises and Anderson-Darling tests) and unbinned (Kolmogorov-Smirnov, Goodman-Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling and Kuiper tests) distributions.

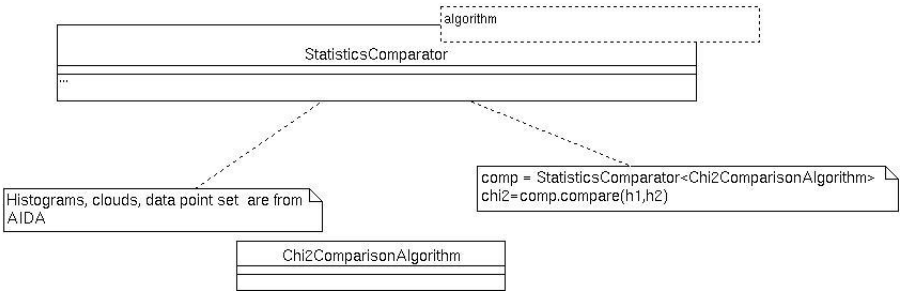

Figure 3: User layer of the **GoF** Toolkit: dealing with AIDA objects, the user is completely shielded from the complexity of both design and statistics. He has only to extract the algorithm he wants to use and to run the comparison. In the specific case shown, the user is running the Chi-squared algorithm.

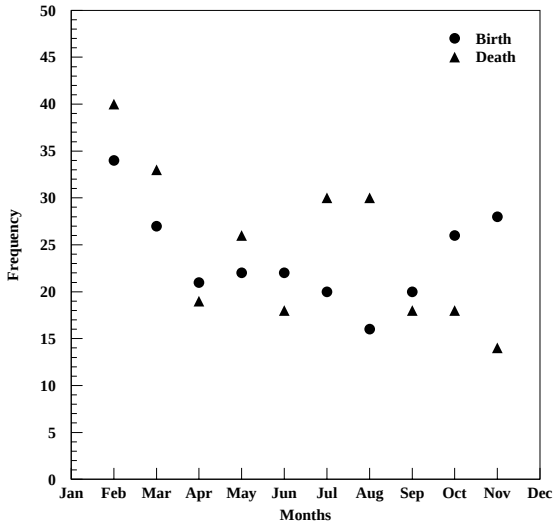

Figure 4: Example from Piccolo's book [33]: the author wants to compare birth against death monthly distributions in a sample of 294 people. Statistical comparisons lead to the same result computed by Piccolo.

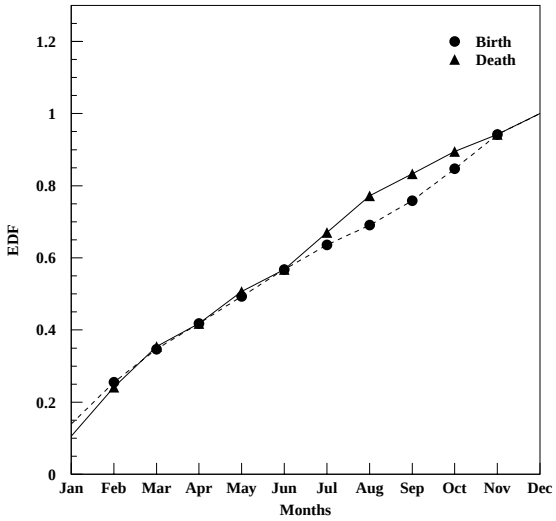

Figure 5: Example from Landenna's book [34]: the author wants to compare the results of a physiology experiment performed by Delse and Feather [35] onto two groups made up by 10 people each. Statistical comparisons lead to the same results computed by Landenna.