New Developments of the Goodness-of-Fit Statistical Toolkit

Barbara Mascialino, Andreas Pfeiffer, Maria Grazia Pia, Alberto Ribon, and Paolo Viarengo

***Abstract*—The Statistical Toolkit is a project for the development of open source software tools for statistical data analysis in experimental particle and nuclear physics. The second development cycle encompassed an extension of the software functionality and new tools to facilitate its usage in experimental environments. The new developments include additional goodness-of-fit tests, new implementations of existing tests to improve their statistical precision or computational performance, a new component to extend the usability of the toolkit with other data analysis systems, and new tools for an easier configuration and build of the system in the user's computing environment. The computational performance of all the algorithms implemented has been studied.**

***Index Terms*—Data analysis, data comparison, goodness-of-fit testing, software, Statistical Toolkit.**

I. Introduction

THE comparison of data distributions with respect to other reference data or functions is a common problem in experimental physics: some typical cases are the validation of simulation results against experimental data, the evaluation of physical quantities reconstructed by the experiment's software against theoretically expected ones, or monitoring the behaviour of a particle detector with respect to its nominal operation reference. Moreover, the regression testing of an experiment's software usually involves some comparisons of data distributions to monitor the software stability or to verify its evolution.

A recent project, named the Statistical Toolkit [1], undertook the development of an open source software system for the comparison of data distributions, especially addressing, but not limited to, applications in particle and nuclear physics. This project is characterized by an iterative and incremental software process, according to established best practices in software development [2]; the first development cycle is documented in [1]. This paper describes the new developments and improvements, which have been released [3] for public usage in version 2. The new features available respond to experimental user requirements. The first two development cycles concern the comparison of two samples of one-dimensional distributions.

Several goodness-of-fit tests have been added to the already extensive collection available in the first released version; some of them introduce new weighted formulations of established tests, for the first time available in a software tool for data analysis. Other tests have been significantly improved, either in their mathematical algorithms or in their computational performance. New developments in the architectural user layer have extended the usability of the Statistical Toolkit, addressing in particular the requirements for data analysis in high energy physics experiments. A significant redesign of the supporting software tools package facilitates the configuration of the system in the user's own computing environment.

The paper also reports a comparative analysis of the computing performance of all the algorithms implemented: these results provide useful guidance to experimental users to select the test appropriate to their requirements among those available in the toolkit. A comparative study of the statistical performance of the various goodness-of-fit tests is the object of current research activity [4], [5]; it will be documented in a dedicated paper.

II. Overview of the Goodness-of-Fit Statistical Toolkit

The Statistical Toolkit is a software system for statistical data analysis; it is especially targeted to common applications in experimental nuclear and particle science. It exploits the object oriented technology and generic programming techniques; it is implemented in C++. Its life-cycle is based on the iterative-incremental model of the Unified Process [2]; the process framework adopted emphasizes the role of the software architecture and the relevance of use cases in the software development.

The Statistical Toolkit adopts a component-based architecture, which facilitates its usage in association with other data analysis software systems widely used in particle physics experiments; its sound object oriented design makes it open to extension and evolution.

A. Statistical Background

Goodness-of-fit testing provides the mathematical foundation for a rigorous, quantitative evaluation of the compatibility of two data samples, or of a data sample against a reference function. A detailed overview of goodness-of-fit testing is reported in [1]; the brief summary below is meant to introduce the basic mathematical concepts involved in the software developments described in the following sections.

Let X and Y be two real-valued random variables, and let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ be, respectively, two samples of independent and identically distributed observations with empirical distribution functions F_n and G_m. For every x the empirical distribution function $F_n(x)$ denotes the observed fraction

Manuscript received July 12, 2006; revised September 14, 2006.

B. Mascialino and M. G. Pia are with INFN Sezione di Genova, 16146 Genova, Italy (e-mail: Barbara.Mascialino@ge.infn.it; Maria-Grazia.Pia@ge.infn.it).

A. Pfeiffer and A. Ribon are with CERN, CH 1122 Geneva, Switzerland (e-mail: Andreas.Pfeiffer@cern.ch; Alberto.Ribon@cern.ch).

P. Viarengo is with IST National Institute for Cancer Research, 16132 Genova, Italy (e-mail: paolo.viarengo@istge.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNS.2006.885386

0018-9499/$20.00 © 2006 IEEE

of values smaller or equal to x; it ranges between 0 and 1. An empirical distribution function statistic evaluates the deviation between F_n and G_m; the value of the corresponding test statistic is computed as a mathematical function of the differences between the two empirical distribution functions.

Two-sample inferences test the null hypothesis

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

without specifying the common distribution function between the two samples.

Under the hypothesis that the two samples were drawn from the same population distribution, the corresponding empirical distribution functions are expected to be close to each other; if the two-sample empirical distribution functions differ significantly, it is likely that the samples derive from different populations.

The result of a goodness-of-fit test is expressed through a p-value, which represents the probability that the test statistic has a value at least as extreme as that observed, assuming the null hypothesis H_0 is true. The exact distributions associated with empirical distribution function statistics are usually complex and scarcely documented in the literature; for this reason the Statistical Toolkit provides the asymptotic distributions for the calculation of the p-value associated to the observed test statistic of the various algorithms, whenever they have been documented. In a limited number of cases, documented in the following sections, the toolkit provides the critical values, that is the values corresponding to a given significance, instead than the p-value.

Several formulations of goodness-of-fit tests have been devised in statistical science. They apply to binned or unbinned distributions, or are especially suitable to compare distributions with specific characteristics, such as cyclic observations or so-called fat tails. The Statistical Toolkit refers to authoritative sources in statistical literature for a sound theoretical foundation of the algorithms implemented in the software and distributed to the scientific community. It aims to provide an exhaustive collection of algorithms, among which the user can choose the most appropriate one to her or his experimental problem.

A complete and systematic evaluation of the power of the existing goodness-of-fit tests, that is of their capability to accept or reject the null hypothesis H_0 correctly, has not been documented in statistical literature yet; only some sparse data about a few specific tests and application cases are available.

B. Software Features

The first released version of the Statistical Toolkit [1] defined its architecture and implemented an initial set of goodness-of-fit tests, chosen among the most commonly used in the data analysis of particle physics experiments.

The component-based architecture of the Statistical Toolkit is articulated through a Layer architectural pattern [6]. The architecture distinguishes a core statistical layer, which is responsible for mathematical computations, and a user layer, which is responsible for any user analysis actions and for the interface to external data analysis systems. The architectural pattern adopted, which clearly decouples the functionality of the mathematical component from its usage, allows for an independent, unlimited extension of the statistical functionality of the Toolkit; it also facilitates the usability of the software in different data analysis environments.

The main types in the core statistical component are the *Distribution*, the *ComparatorEngine* and the *ComparisonAlgorithm*. The *ComparatorEngine* class is responsible for driving the comparison; it is parameterized over the data *Distribution* type (binned or unbinned) and the goodness-of-fit *Algorithm*. Concrete goodness-of-fit tests correspond to specializations of the *ComparisonAlgorithm* parameterized class, respectively bound to binned or unbinned *Distributions*. A Strategy pattern [7] handles the evaluation of the quality of the fit. The main features of the statistical test design are shown in Fig. 1.

The goodness-of-fit tests implemented are listed in Table I, where names in italic highlight the new tests released in version 2 of the Toolkit. The set of tests available in the first release covered various mathematical approaches, encompassing the χ^2 test, tests based on the empirical distribution functions maximum distance and tests based on the empirical distribution functions quadratic distance. It included the two-sample goodness-of-fit tests most widely used in experimental physics (χ^2, Kolmogorov-Smirnov), and a few other tests applied in more sophisticated analyses of particle physics experiments, like the Anderson-Darling test and the Cramér-von Mises test.

Some tests were implemented with preliminary algorithms in the first development cycle; the focus of the development process was on the mathematical correctness of the algorithms, rather than on their performance optimization. Thanks to the sound software design, implementation improvements could be addressed in further development cycles without interfering with the rest of the software, according to the iterative-incremental software process adopted.

The first release of the Statistical Toolkit encompassed one User Layer component, based on the AIDA [8] Abstract Interfaces for Data Analysis. This User Layer component interfaced the Statistical Toolkit to all the AIDA-compliant analysis systems: JAS [9], Open Scientist [10], PAIDA [11] and PI [12]. The component-based architecture of the system allows extending the interface to other analysis systems, without affecting the core statistical component.

C. Extension of the Functionality

The second development cycle, which is the object of this paper, addressed a well defined set of requirements in various domains:

- an extension of the set of goodness-of-fit tests covered: an ample set of software tools would enable the choice of the most appropriate test for any specific data analysis; the availability of software tools for an extensive number of tests also allows for the first time a comprehensive, quantitative evaluation of their applicability and their respective power in various experimental conditions;
- more precise calculations of some test algorithms available in the first release, such as the computation of p-values instead of critical values;

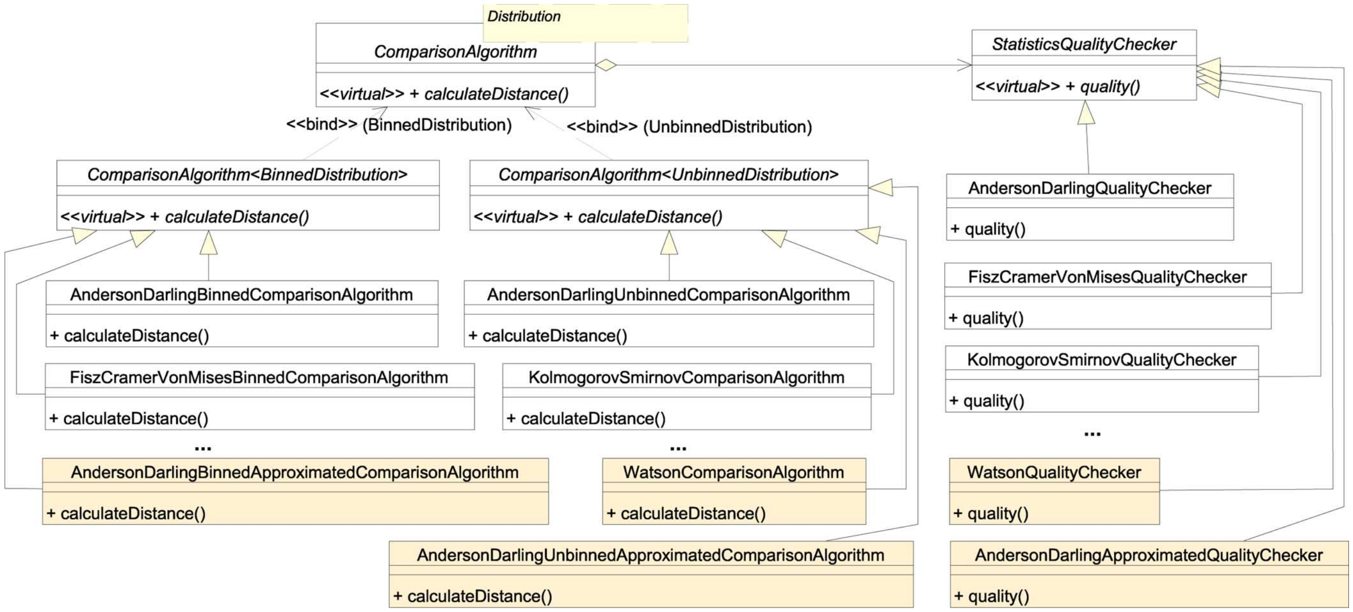
Distribution
ComparisonAlgorithm
<<virtual>> + calculateDistance()
StatisticsQualityChecker
<<virtual>> + quality()
<<bind>> (BinnedDistribution)
<<bind>> (UnbinnedDistribution)
ComparisonAlgorithm<BinnedDistribution>
<<virtual>> + calculateDistance()
ComparisonAlgorithm<UnbinnedDistribution>
<<virtual>> + calculateDistance()
AndersonDarlingQualityChecker
+ quality()
AndersonDarlingBinnedComparisonAlgorithm
+ calculateDistance()
AndersonDarlingUnbinnedComparisonAlgorithm
+ calculateDistance()
FiszCramerVonMisesQualityChecker
+ quality()
FiszCramerVonMisesBinnedComparisonAlgorithm
+ calculateDistance()
KolmogorovSmirnovComparisonAlgorithm
+ calculateDistance()
KolmogorovSmirnovQualityChecker
+ quality()
...
...
...
AndersonDarlingBinnedApproximatedComparisonAlgorithm
+ calculateDistance()
WatsonComparisonAlgorithm
+ calculateDistance()
WatsonQualityChecker
+ quality()
AndersonDarlingUnbinnedApproximatedComparisonAlgorithm
+ calculateDistance()
AndersonDarlingApproximatedQualityChecker
+ quality()

Fig. 1. Detail of the Goodness-of-Fit Statistical Toolkit design: the white classes represent components present in the first released version, the colored ones represent classes of the new goodness-of-fit tests added in the second development cycle.

TABLE I
LIST OF ALL THE GOODNESS-OF-FIT TESTS IMPLEMENTED IN THE STATISTICAL TOOLKIT [1]. THE NAMES IN ITALIC HIGHLIGHT THE NEW TESTS AVAILABLE IN THE 2.0 RELEASE

Goodness-of-fit tests for binned distributions	Anderson-Darling *Anderson-Darling, approximated algorithm* *Chi-squared (new implementation)* Chi-squared (old implementation) Fisz-Cramér-von Mises Tiku
Goodness-of-fit tests for unbinned distributions	Anderson-Darling *Anderson-Darling, approximated algorithm* Fisz-Cramér-von Mises *Generalised Girone* Goodman Kolmogorov-Smirnov Kuiper Tiku *Watson* *Weighted Kolmogorov-Smirnov (AD)* *Weighted Kolmogorov-Smirnov (Buning)* *Weighted Cramér-von Mises*

- a faster implementation of the χ^2 test suitable for application in online data analysis;
- the possibility of using the Statistical Toolkit in association with ROOT [13], that is a data analysis system widely used in high energy physics experiments.

The component-based architecture of the Statistical Toolkit and its sound object oriented design facilitated its evolution to satisfy the new requirements. The new features and improved functionality were included in the system as extensions of the existing design, or just as new implementations of already existing classes, without needing any modification of interfaces in the core statistical component or of user's code. The process of extending the toolkit functionality for the inclusion of a new goodness-of-fit test is illustrated in the UML (Unified Modelling Language) [14] class diagram of Fig. 1: adding a new test implies the creation of a concrete class implementing the *ComparisonAlgorithm* interface for the appropriate distribution type, and the implementation of the associated *StatisticsQualityChecker*. Since the main conceptual entities of the problem domain are handled through abstract interfaces, the extension of the system is completely transparent to the core computational component.

Similarly, the addition of a new User Layer component enabled the usability with ROOT in a transparent way.

III. New Goodness-of-Fit Tests

The second version of the Statistical Toolkit includes various new goodness-of-fit tests: the Watson test, a generalised formulation of the Girone test, and variants of the Kolmogorov-Smirnov and Cramér-von Mises tests.

A. *Generalised Girone Test*

The second release of the Statistical Toolkit provides the first implementation of a generalized version of the Girone test [15]–[20] in an open source software tool.

The Girone test [21]–[23] presents some interesting characteristics: when the variable under study is asymmetric, it appears more powerful than the Kolmogorov-Smirnov and Cramér-von Mises tests [15]. However, it is affected by an intrinsic limitation: it is applicable only to the comparison of samples of equal sizes $(n = m)$, whereas in experimental practice one often encounters data samples of different size. Various studies aimed at generalizing the original Girone test are documented in statistics literature, and their validity has been proved to be asymptotically equivalent [24].

A version of general applicability, resulting from a critical analysis of the mathematical literature on this issue, has been implemented in the Statistical Toolkit as *GironeGeneralisedComparisonAlgorithm*. This test can be applied to unbinned distributions; the test statistic implemented is:

$$GG(F_n, G_m) = \frac{nm}{(n+m)}\left[\sum_{1}^{n}|F_n - G_m| + \sum_{1}^{m}|F_n - G_m|\right]. \quad (2)$$

B. Watson Test

An implementation of the Watson test [25]–[27] is also provided in this release of the Statistical Toolkit. This test was originally proposed [25]–[27] for application to unbinned cyclic observations, however, it is valid even for samples not exhibiting a periodicity [28]. This test is quite powerful [27] and provides equal sensitivity to the tails as to the median of the two empirical distribution functions.

The Watson test statistic involves the integral of the squared deviations between the empirical distribution functions of the two-samples; in the Statistical Toolkit the following formula [26] has been implemented, which is more suitable for numerical calculations, as it does not require any computationally intensive integration, while it has been demonstrated to be equivalent to the original one:

$$U^2 = \frac{nm}{(n+m)^2} \times \left[\sum_{1}^{n+m}(F_n - G_m)^2 - \frac{1}{n+m}\sum_{1}^{n+m}(F_n - G_m)^2\right]. \quad (3)$$

C. Weighted Kolmogorov-Smirnov and Cramér-Von Mises Tests

Two modified versions of the Kolmogorov-Smirnov test, have been introduced in the second release of the Statistical Toolkit.

The Kolmogorov-Smirnov [29], [30] test statistic is defined as the maximum unsigned deviation between the two empirical distribution functions derived from the data samples to be compared. Modified versions of the Kolmogorov-Smirnov test have been proposed [31], [32]; they introduce appropriate non-negative weight functions $W_i(F_n, G_m)$ to attribute different weights to different parts of the distributions:

$$WKS_i(F_n, G_m) = \max|F_n - G_m| \cdot W_i(F_n, G_m). \quad (4)$$

$W_i(F_n, G_m)$ can have many formulations; two of them, the Anderson- Darling and the Buning weighting functions, have been implemented in the Statistical Toolkit. These functions attribute larger weight to different parts of the distributions; therefore, the corresponding weighted Kolmogorov-Smirnov tests are suitable to look for specific types of deviations.

The Anderson and Darling [33], [34] weighting function W_{AD} is symmetric; it attributes larger weight to the lower (close to 0) and upper (close to 1) part of the empirical distribution functions:

$$W_{AD}(F_n, G_m) = \frac{1}{\sqrt{\frac{(nF_n + mG_m)}{n+m} \cdot \left(1 - \frac{(nF_n + mG_m)}{n+m}\right)}}. \quad (5)$$

The Buning [32] weighting function W_B is asymmetric and emphasises the lower part of the empirical distribution functions:

$$W_B(F_n, G_m) = \frac{1}{\sqrt{\frac{(nF_n + mG_m)}{n+m} \cdot \left(2 - \frac{(nF_n + mG_m)}{n+m}\right)}}. \quad (6)$$

A new weighted Cramér-von Mises test $WCvM_B(F_n, G_m)$, which adopts the Buning [32] weighting function W_B, has been added to the second released version of the Statistical Toolkit.

The Cramér-von Mises test statistic [35], [36] measures the sum of the integrated squared discrepancy between F_n and G_m. For historical reasons, when the Cramér-von Mises test statistic is weighted with W_{AD}, the test is named the Anderson-Darling test [33], [34]. Both the original Cramér-von Mises and the Anderson-Darling tests were already present in the first version of the Statistical Toolkit (the former named Fisz-Cramér-von Mises test in its generalized version for the two-sample problem [37]). The test statistic of the new weighted Cramér-von Mises test has the following formulation:

$$WCvM_B(F_n, G_m) = \frac{nm}{(n+m)^2} \times \sum_{1}^{n+m}[F_n - G_m]^2 \cdot W_B(F_n, G_m). \quad (7)$$

The asymptotic distributions associated to the weighted test statistic are not documented in statistical literature; only some critical values corresponding to the significance level of 0.05 were obtained by Monte Carlo simulation for a few selected sample sizes [31], [32]. This deficiency would have limited the applicability of the weighted tests, in spite of the fact that these tests are more powerful than other goodness-of-fit tests for some data sample comparisons [32]. An original approach was developed to overcome this limitation in the Statistical Toolkit: it consists of an approximated method to calculate the critical values of the test statistic corresponding to the significance level of 0.05 for any sample size. For each test, a regression model has been devised on an empirical basis and evaluated on the critical values available in literature; these models are capable of predicting the expected critical values ($\hat{\mathrm{cv}}$) as a function of the sample sizes.

The regression models associated to the two weighted Kolmogorov-Smirnov tests and to the weighted Cramér-von Mises test are listed in Table II. The table also reports the models' coefficient of determination R^2: R^2 ranges from 0 (when no agreement is found between the model and the data points) to 1 (in the case of perfect agreement); it represents the proportion of variation in the dependent variable $\hat{\mathrm{cv}}$ explained by the model. The high R^2 values in Table II confirm the excellent predictive capabilities of the model developed for the Statistical Toolkit implementation of the weighted goodness-of-fit tests.

The implementation in the Statistical Toolkit makes these weighted tests available for the first time in an open-source software tool; it enables researchers to exploit their peculiar features in physics data analyses, by selecting the version of the well known Kolmogorov-Smirnov test, or of the powerful

TABLE II
REGRESSION MODELS TO EVALUATE THE EXPECTED CRITICAL VALUES ($\hat{cv}$) CORRESPONDING TO THE SIGNIFICANCE LEVEL 0.05 AS A FUNCTION OF SAMPLE SIZES FOR THE WEIGHTED KOLMOGOROV-SMIRNOV AND CRAMÉR-VON MISES TESTS. THE DETERMINATION COEFFICIENTS R^2 HIGHLIGHT THE EXCELLENT PREDICTIVE QUALITY OF THE MODELS

Test	Expected critical values	Model quality
$\mathrm{WKS_{AD}}$	$\hat{cv} = (2.45 \pm 0.06) + (0.11 \pm 0.01) \cdot \ln(\sqrt{\frac{nm}{n+m}})$	$R^2 = 0.90$
$\mathrm{WKS_{B}}$	$\hat{cv} = (1.16 \pm 0.04) + (0.17 \pm 0.02) \cdot \ln(\sqrt{\frac{nm}{n+m}})$	$R^2 = 0.98$
$\mathrm{WCvM_{B}}$	$\hat{cv} = (0.69 \pm 0.01) + (0.03 \pm 0.01) \cdot \ln(\sqrt{\frac{nm}{n+m}})$	$R^2 = 0.99$

Cramér-von Mises test most appropriate to the data distributions to be compared.

D. Approximation of the Anderson-Darling Test

This release of the Statistical Toolkit also includes a new approximated version of the Anderson-Darling test, in addition to the standard one. The implementation of the approximated algorithm was motivated by the interest to evaluate potential benefits in terms of computing performance, that could be valuable in some use cases. The Anderson-Darling test is known as one of the most powerful goodness-of-fit tests [38]; the investigation of the performance of an approximation is relevant to users concerned with the speed of execution, as well as the power, of goodness-of-fit tests.

The Anderson-Darling [39], [40] test statistic is based on a doubly weighted sum of the integrated squared differences between the two empirical distribution functions weighted by the weighting function proposed by Anderson and Darling and defined in (5). This test has a highly skewed and complex limit distribution. The calculation of the test statistic is left unchanged in the approximated variant, while an empirical approximation [41] is used for the calculation of the p-value of the test. If A^2 represents Anderson-Darling test statistic, the alternative formulation of its limiting distribution implemented in the Statistical Toolkit is:

$$P(A^2 \leq z) = 1 - \frac{1}{1 + e^{1.784 + 0.9936z + \frac{0.3287}{z} - \frac{1}{\sqrt{z}}\left(2.018 + \frac{0.2029}{z}\right)}}. \quad (8)$$

This formula is adequate for both the upper and lower tails of the asymptotic distribution; it was demonstrated to be more accurate than other theoretical approximations proposed [41].

The computational performance of the Anderson-Darling test and of its approximated variant is documented in Table III and discussed in Section VII.

IV. IMPROVED ALGORITHMS

The Statistical Toolkit offers new implementations of some goodness-of-fit tests already released in its first version. The improvements concern either the precision of the statistical calculations or the computational speed of the algorithms.

A. New Implementations of the Anderson-Darling and Fisz-Cramér-Von Mises Tests

The first release of the Statistical Toolkit provided the critical values corresponding to the significance level 0.05 for two of the most powerful goodness-of-fit tests: the Anderson-Darling and the Fisz-Cramér-von Mises. The critical values have been replaced by the statistic asymptotic distributions in the second release; this evolution addressed the requirement for a p-value evaluation based on a more accurate and rigorous method.

In the Statistical Toolkit the Anderson-Darling p-value calculation is based on the one-sample limit distribution [40]:

$$P(A^2 \leq z) = \frac{\sqrt{2\pi}}{z} \sum_{j=0}^{\infty} \binom{-\frac{1}{2}}{j} (4j+1) e^{-\frac{(4j+1)^2 \pi^2}{8z}} \times \int_0^{\infty} e^{\frac{z}{8(w^2+1)} - \frac{(4j+1)^2 \pi^2 w^2}{8z}} dw. \quad (9)$$

In fact, the Glivenko-Cantelli theorem [42] allows using the asymptotic distributions of the one-sample problem in the two-sample case too, given the null hypothesis $\mathrm{H_0}$ as in (1).

Similarly, under the null hypothesis $\mathrm{H_0}$, the two-sample Fisz-Cramér-von Mises asymptotic distribution has the same limiting distribution as the one-sample test statistic [43]:

$$P(T \leq z) = \frac{1}{\pi\sqrt{z}} \sum_{j=0}^{\infty} (-1)^j \binom{-\frac{1}{2}}{j} \sqrt{4j+1} \cdot e^{-\frac{(4j+1)^2}{16z}} K_{1/4}\left[\frac{(4j+1)^2}{16z}\right], \quad (10)$$

where $K_{1/4}$ is the modified Bessel function of the second kind.

B. New Implementation of the χ^2 Test

Among the many goodness-of-fit tests devised in statistical science, the χ^2 test is the most commonly used in experimental particle physics. Because of the wide usage of this test, a software tool for statistical analysis should address a variety of use cases, corresponding to different applications in the experimental practice: some, like physics analysis, may be more concerned with the precision of the calculation, while other ones, like online detector monitoring, may be more concerned with the speed of the computation. The implementation of χ^2 test in the first released version of the Statistical Toolkit satisfied the requirement of precision of calculation; a new implementation in the released version 2 also addresses the computational speed.

A new version of the χ^2 test is based on the fact that the χ^2 cumulative distribution function is related to the incomplete gamma function [44]–[47]. The algorithm implementation exploits the $\mathrm{gsl\_cdf\_chisq\_}P$ function of the GNU Scientific Library [48] for generating random variables and computing their probability, using the incomplete gamma function. The class corresponding to this new implementation has been named *Chi2ComparisonAlgorithm*; the old version is still distributed in the Statistical Toolkit as *Chi2IntegratingComparisonAlgorithm* to enable comparative evaluations.

The computational performance of the two χ^2 versions is documented in Table III. The new formulation represents a significant time gain with respect to the first released version; it is suitable to online data analysis, where computational speed is a significant requirement. The p-values resulting from either version of the algorithm differ less than 10^{-6}.

A similar performance improvement, based on the same tools of the GNU Scientific Library, has been implemented in the Goodman [49] and Tiku [50] tests, which respectively represent

TABLE III
AVERAGE CPU TIME IN MS FOR THE EXECUTION OF THE GOODNESS-OF-FIT TESTS

Algorithms for Binned Distributions	**CPU time** (ms)
Anderson-Darling	0.69 ± 0.01
Anderson-Darling, approximated	0.60 ± 0.01
χ^2, new implementation	0.55 ± 0.01
χ^2, old implementation	812 ± 8
Fisz-Cramér-von Mises	0.44 ± 0.01
Tiku	0.69 ± 0.01

Algorithms for Unbinned Distributions	**CPU time** (ms)
Anderson-Darling	16.9 ± 0.2
Anderson-Darling, approximated	16.1 ± 0.2
Fisz-Cramér-von Mises	16.3 ± 0.2
Generalised Girone	15.9 ± 0.2
Goodman	11.9 ± 0.1
Kolmogorov-Smirnov	8.9 ± 0.1
Kuiper	12.1 ± 0.1
Tiku	16.7 ± 0.2
Watson	14.2 ± 0.1
Weighted Kolmogorov-Smirnov (AD)	14.0 ± 0.1
Weighted Kolmogorov-Smirnov (Buning)	14.0 ± 0.1
Weighted Cramér-von Mises	14.0 ± 0.1

the approximation of the Kolmogorov-Smirnov and Cramér-von Mises tests to a χ^2 statistic.

V. EXTENSION OF THE USER LAYER

The user requirement of enabling the usage of the Statistical Toolkit to compare ROOT analysis objects was motivated by the wide popularity of this analysis system in high energy physics experiments. In fact, the rich functionality of the Statistical Toolkit effectively complements the limited tools for the comparison of histograms currently available in ROOT.

The User Layer bridges the user's representation of analysis objects to be compared to the binned and unbinned *Distribution* classes handled by the core Statistical Layer. An implementation of a new User Layer component is easily performed by creating a single class called *StatisticsComparator* and implementing one public method for the comparison (*compare*). This is typically done in the header file of the class itself, thus reducing the overhead to rebuild the library for adding new user layer components.

The usability with ROOT was easily satisfied by extending the User Layer with a new component implementing the functionality required to handle user supplied ROOT histograms (i.e., binned analysis objects); three protected helper methods are employed, allowing an easy extension for the case that future versions of ROOT may provide unbinned data classes too. The helper methods convert ROOT one-dimensional histograms into the *BinnedDistributions* type employed by the toolkit, create an instance of the templated *Algorithm*, then forward the actual evaluation to the *Algorithm*.

Because of the clear separation of the User and Statistical Layers in the architecture of the Statistical Toolkit, and of different components in the User Layer itself, the extension for the operability with ROOT did not affect any of the existing software. As described in Section VI, the user may decide to configure her or his system with either the ROOT or the AIDA User Layer components, or with both.

VI. CONFIGURATION AND BUILD TOOLS

The new version of the Statistical Toolkit represents a substantial improvement concerning the installation and configuration in the user's software environment. Standard procedures provide a more flexible and user friendly way to configure and build it.

The only external dependency of the system is on the GNU Scientific Library (GSL) [48]. The initial analysis objects supplied by the user as input to the comparison also depend on external analysis systems; however, this dependency is limited to the User Layer, while the core Statistical Layer does not depend on any external analysis tools.

In the version 2 release the configuration and build of the Statistical Toolkit software is fully based on GNU Autotools [51]; these tools are the de facto standard for portably building and installing C and C++ applications across different UNIX flavours and systems.

The Statistical Toolkit provides two scripts (*configure.in* and *Makefile.am*) for the fully automated configuration and build of the system in the user environment. The configuration scripts accept the options to configure the system with either the AIDA or the ROOT user layer, or with both; the user can also specify the location of the external systems (GSL, ROOT, AIDA-compliant analysis tools) in her or his computing environment.

The build procedure provides the option to build and run all the unit tests of the Statistical Toolkit; therefore, the user has the possibility to verify the correctness of her or his installation of the software.

The reference supported platform for the Statistical Toolkit is the open source Scientific Linux CERN (SLC Version 3 at the time of the second release of the toolkit). This platform is widely used in particle and nuclear physics experiments; it is not a site-specific product: all CERN site customizations are optional and need not be activated for users in other environments than CERN. The adoption of the standard GNU Autotools and of coding guidelines compliant with the C++ standard greatly facilitates porting the system to other platforms.

VII. PERFORMANCE OF THE ALGORITHMS

The computational speed of the various goodness-of-fit algorithms available in the Statistical Toolkit is one of the criteria to select a test appropriate to a given application. A complementary selection criterion is the power of the test, that is its capability to identify compatible distributions correctly, while minimizing the chance of spurious compatibility results; this topic will be extensively treated in two papers currently in preparation.

A thorough investigation of the computational performance of all the available goodness-of-fit tests has been carried out in connection with the release of version 2 of the Statistical Toolkit. This study exploited the software tools regularly used in the system testing process of the toolkit.

The system testing of the Statistical Toolkit encompasses a series of tests, that exercise the goodness-of-fit algorithms in realistic use cases. It consists of the Monte Carlo generation of a large number of pseudo-experiments: unbinned and binned data samples of variable size are drawn from a set of parent distributions (flat, gaussian, right and left tailed exponential), and compared to one another through each of the goodness-of-fit algorithms. These tests were equipped with a calculation of the CPU time spent in the computation of the comparison algorithm.

The tests were executed over 10000 pseudo-experiments on a PC with a Pentium ™ IV (3 GHz) processor and 512 MB of RAM; the size of the samples subject to comparison was 500; for the binned tests the number of bins was 20. The results are listed in Table III. The absolute values of the execution times depend on the characteristics of the distributions subject to the tests, not only on the intrinsic features of the goodness-of-fit algorithms; therefore, Table III should be considered as a relative indication of the computational speed of the various algorithms, rather than as an absolute reference.

The computational performance of the various tests depends on the the distributions type (binned or unbinned): binned data analysis is in general faster, since, as a consequence of the grouped nature of data, the empirical distribution functions exhibit simpler computational features.

The CPU times of the tests for binned distributions are all quite similar, with the exception of the old implementation of the χ^2 test. The new implementation of this test shows a significant improvement, which makes it suitable for usage in online data analysis. It is worthwhile noting that the Fisz-Cramér-von Mises test is the fastest among those available for binned distributions, with a performance 20% better than the more popular χ^2 test. Such a consideration may be relevant to application cases where the computational performance is a critical issue. The calculation of the χ^2 probability requires the evaluation of the cumulative distribution function to determine the probability p corresponding to a given test statistic value χ^2_{obs}; therefore, it is not surprising that the execution of the software for the p-value calculation is computationally intensive.

Among unbinned distributions, the Kolmogorov-Smirnov test exhibits the fastest execution time, while most of the other tests show a comparable computational performance.

The performance gain of the approximated Anderson-Darling test is modest with respect to its full formulation: it is limited to approximately 15% for binned distributions and about 5% for unbinned ones. Other approximated versions of established tests exhibit a worse computational performance than the original formulation: this is the case, for instance, of the Goodman and Tiku tests, which are approximated versions of the Kolmogorov-Smirnov and Cramér-von Mises tests respectively. This conclusion is apparently surprising with respect to the naive expectation that an approximated algorithm would be faster. For historical reasons approximated tests have been devised to facilitate the manual calculation of the test statistic in some practical application cases, for instance introducing the reference to commonly available tabulated values of the χ^2; nevertheless, their comparative computational performance has not been previously documented quantitatively in a software environment. The present study shows that approximated tests play a limited role in a modern computational environment. The availability of a wide set of goodness-of-fit tests in the same open-source software environment makes a thorough investigation of their properties achievable for the first time: this holds not only for their comparative computational performance, but also for a rigorous study of their statistical power.

VIII. Conclusion

The second development cycle of the Statistical Toolkit has resulted in a significant extension and improvement with respect to the first publicly released version.

New goodness-of-fit tests have been implemented, including a few available for the first time in a publicly released software product. The extensive collection of algorithms implemented makes the Statistical Toolkit the most complete software system for the comparison of two data distributions, not only among data analysis systems for physics research, but even in the professional domain of statistics software tools. To the authors' knowledge, the new version of the Statistical Toolkit implements all the two-sample goodness-of-fit tests based on the empirical distribution function statistics known in statistical science as well as the χ^2 test.

New implementations of existing tests have improved their computational performance or their precision with respect to the first version. Original developments have allowed to extend the applicability of some tests, like the weighted ones, with respect to the formulations available in statistical literature.

A new User Layer component has been added to facilitate the usage of the Statistical Toolkit in high energy physics experiments.

The sound architecture of the Statistical Toolkit enabled the implementation of new functionality without affecting the existing core software or the user's one. This characteristics, together with its rigorous iterative-incremental software process, makes the Statistical Toolkit open to further extension and evolution. New development cycles are planned to address complementary problems in the domain of data comparisons: the one-sample and the k-sample problem, the comparison of multi-dimensional distributions, and the treatment of uncertainties. Besides these major planned extensions, the requirements of more precise p-value calculations or of possibly new tests will continue to be addressed too.

A comprehensive study is in progress to analyse quantitatively the power of all the goodness-of-fit tests, with the purpose to identify the most appropriate ones for different experimental applications. The results of this study will be documented in future publications currently in preparation.

Acknowledgment

The authors are grateful to L. Moneta for fruitful discussions on the performance improvement of the χ^2 test, and for the support to distribute the Statistical Toolkit in association to the CERN Mathematical Libraries. The authors thank P. Arce for useful comments and suggestions, and users in the ATLAS and CMS Collaborations for valuable feedback.

References

[1] G. A. P. Cirrone, S. Donadio, S. Guatelli, A. Mantero, B. Mascialino, and S. Parlati *et al.*, "A goodness-of-fit statistical toolkit," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 5, pp. 2056–2063, Oct. 2004.

[2] I. Jacobson, G. Booch, and J. Rumbaugh, *The Unified Software Development Process*. Reading, MA: Addison-Wesley, 1999.

[3] [Online]. Available: http://www.ge.infn.it/statisticaltoolkit/

[4] S. Guatelli, B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon, and P. Viarengo, "Application of statistical methods for the comparison of data distributions," in *Proc. Conf. Rec. 2004 IEEE Nuclear Science Symp.*, vol. N40-5.

[5] S. Guatelli, B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon, and P. Viarengo, "An update on the goodness-of-fit statistical toolkit," in *Proc. PHYSTAT'05, Statistical Problems in Particle Physics, Astrophysics and Cosmology,*, Oxford, U.K., 2005.

[6] F. Buschmann, E. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, , Chichester, Ed., *A system of patterns*. New York: Wiley, 1996.

[7] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns*. Reading, MA: Addison-Wesley, 1995, ch. 5.

[8] G. Barrand, P. Binko, M. Donszelmann, A. Johnson, and A. Pfeiffer, "Abstract interfaces for data analysis—Component architecture for data analysis tools," in *Proc. Computing in High Energy and Nuclear Physics*, Beijing, China, 2001, pp. 215–218.

[9] T. Johnson, "Java analysis studio," in *Proc. Computing in High Energy and Nuclear Physics*, Padova, Italy, 2000, pp. 741–745.

[10] G. Barrand, "Open scientist. Status of the project," in *Proc. Computing in High Energy and Nuclear Physics*, Interlaken, Switzerland, 2004.

[11] [Online]. Available: http://paida.sourceforge.net/

[12] A. Pfeiffer, L. Moneta, V. Innocente, H. C. Lee, and W. L. Ueng, "The LCG PI project: using interfaces for physics data analysis," *IEEE Trans. Nucl. Sci.*, vol. 52, no. 6, pp. 2823–2826, Dec. 2005.

[13] R. Brun and F. Rademakers, "ROOT, an object oriented data analysis framework," *Nucl. Instrum. Methods Phys. Res. A*, vol. A389, no. 1–2, pp. 81–86, 1997.

[14] G. Booch, J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide*. Reading, MA: Addison-Wesley, 1999.

[15] C. G. Borroni, "Some notes about nonparametric tests for the equality of two populations," *Test*, vol. 10, pp. 147–159, 2001.

[16] M. Goria, "Some generalizations of Girone's test," *Annali dell'Istituto di Statistica - Università di Bari*, vol. 36, pp. 53–72, 1971–72.

[17] D. M. Cifarelli and E. Regazzini, "On the asymptotic distribution of a statistic arising in testing the homogeneity of two samples," *Giorn. Eco.*, vol. 2, pp. 233–249, 1975.

[18] L. A. Shepp, "On the integral of the absolute value of the pinned Wiener process," *Anls. Prob.*, vol. 10, no. 1, pp. 234–239, 1982.

[19] F. Schmid and M. Trede, "A distribution free test for the two sample problem for general alternatives," *Cmp. St. D. A.*, vol. 20, pp. 409–419, 1995.

[20] L. A. Shepp, "Acknowledgement of priority," *Anls. Prob.*, vol. 19, p. 1397, 1991.

[21] G. Girone, "Su un indice di omogeneità di due distribuzioni del tipo dell'indice semplice di dissomiglianza," *Atti Riunione Scient. S.I.S.*, vol. 24, pp. 53–58, 1964.

[22] G. Girone, "Sulla media e sulla varianza di un indice di dissomiglianza calcolato su ranghi," *Annali dell'Istituto di Statistica—Universitá di Bari*, vol. 36, pp. 40–51, 1971–72.

[23] G. Landenna and D. Marasini, *Metodi statistici non parametrici*. Bologna: Il Mulino, 1990, ch. 7.

[24] C. G. Borroni, "A comparison of some nonparametric test in presence of data from skewed populations," *Atti Riunione Scient. S.I.S*, 2002.

[25] G. S. Watson, "Goodness-of-fit tests on a circle," *Biometrka*, vol. 48, pp. 109–114, 1961.

[26] J. H. Zar, *Biostatistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1984, ch. 2.

[27] E. Batschelet, *Circular statistics in biology*. New York: Academic, 1981.

[28] M. A. Stephens, "EDF statistics for goodness-of-fit and some comparisons," *JASA*, vol. 69, pp. 730–737, 1974.

[29] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Gior. Ist. Ital. Attuari*, vol. 4, pp. 83–91, 1933.

[30] N. V. Smirnov, "On the estimation of the discrepancy between empirical curves of distributions for two independent samples," in *Bull. Math. Univ. Moscou*, 1939.

[31] P. L. Canner, "A simulation study of one- and two-sample Kolmogorov-Smirnov statistics with a particular weight function," *JASA*, vol. 70, no. 349, pp. 209–211, 1975.

[32] H. Buning, "Kolmogorov-Smirnov and Cramér-von Mises type two-sample tests with various weight functions," *Comm. St. B*, vol. 30, no. 4, pp. 847–865, 2001.

[33] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain goodness of fit criteria based on stochastic processes," *Anls. Ma. St.*, vol. 23, pp. 193–212, 1952.

[34] T. W. Anderson and D. A. Darling, "A test of goodness of fit," *JASA*, vol. 49, pp. 765–769, 1954.

[35] H. Cramér, "On the composition of elementary errors. Second paper: statistical applications," *Skand. Aktuarietidskr.*, vol. 11, pp. 13,141–74,180, 1928.

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

[37] M. Fisz, "On a result by M. Rosenblatt concerning the von Mises-Smirnov test," *Anls. Ma. St.*, vol. 31, pp. 427–429, 1960.

[38] M. A. Stephens, "Tests based on EDF statistics," in *Goodness-of-Fit Techniques*, New York, 1986, pp. 97–185, Marcel Dekker.

[39] A. N. Pettitt, "A two-sample Anderson-Darling rank statistic," *Biometrka*, vol. 63, pp. 161–168, 1976.

[40] J. Scholz and M. A. Stephens, "K-sample anderson-darling tests," *JASA*, vol. 82, pp. 918–924, 1987.

[41] C. D. Sinclair and B. D. Spurr, "Approximations to the distribution function of the Anderson-Darling test statistics," *JASA*, vol. 83, no. 404, pp. 1190–1191, 1988.

[42] G. Cicchitelli, *Probabilità e Statistica*. Rimini: Maggioli, 2001, p. 357.

[43] M. Rosenblatt, "Limit theorems associated with variants of the von Mises statistic," *Anls. Ma. St.*, vol. 23, pp. 617–623, 1952.

[44] G. P. Battacharjee, "Algorithm AS 32: The incomplete gamma integral," *Appl. Stat.*, vol. 19, pp. 285–287, 1970.

[45] W. J. Kennedy and J. E. Gentle, *Statistical Computing*. New York: Marcel Dekker, 1980, ch. 5.

[46] E. W. Weisstein, Chi-Squared Distribution Wolfram MathWorld [Online]. Available: http://mathworld.wolfram.com/Chi-SquaredDistribution.html, 1999

[47] M. Abramowitz and I. Stegun, "Handbook of mathematical functions," *National Bureau of Standards Applied Mathematics Series 55*, 1964.

[48] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual Revised Second Edition*. Bristol: Network Theory Limited, 2005.

[49] L. A. Goodman, "Kolmogorov-Smirnov tests for psychological research," *Psychol. Bull.*, vol. 51, pp. 160–168, 1954.

[50] M. L. Tiku, "Chi-square approximations for the distributions of goodness-of-fit statistics U_N^2 and W_N^2," *Biometrika*, vol. 52, pp. 630–633, 1965.

[51] G. V. Vaughan, B. Elliston, T. Tromey, and I. L. Taylor, *GNU Autoconf, Automake, and Libtool*. Berkeley, CA: New Riders, 2000.