A new development cycle of the Statistical Toolkit

M Batič1,2, A. M. Paganoni3, A. Pfeiffer4, M. G. Pia1, A. Ribon4
1 INFN Sezione di Genova, Genova, Italy
2 Jožef Stefan Institute, Ljubljana, Slovenia
3 MOX - Dept. of Mathematics, Politecnico di Milano, Milano, Italy
4 CERN, Geneva, Switzerland
E-mail: matej.batic@ge.infn.it

Abstract. The Statistical Toolkit is an open source system specialized in the statistical comparison of distributions. It addresses requirements common to different experimental domains, such as simulation validation (e.g. comparison of experimental and simulated distributions), regression testing in the course of the software development process, and detector performance monitoring. Various sets of statistical tests have been added to the existing collection to deal with the one sample problem (i.e. the comparison between the data distribution and a known one, including tests for normality, categorical analysis and the estimate of randomness). Improved algorithms and software design contribute to the robustness of the results. A simple user layer dealing with primitive data types facilitates the use of the toolkit both in standalone analyses and in large scale experiments.

1. Introduction
The Statistical Toolkit [1, 2] was originally conceived as a statistical data analysis toolkit for the problem of comparing data distributions. Its development follows the Unified Software Development Process [5]. According to this approach, the life-cycle of the software is iterative-incremental, every iteration representing an evolution, an improvement, an extension in comparison with the previous one. Iterations in the Statistical Toolkit development process are driven by the needs of its experimental applications; practical use cases steer the implementations of new tests.

The first development cycles of the Statistical Toolkit implemented a set of goodness-of-fit (GoF) tests for the two-sample problem, i.e. for the comparison of two distributions. These developments were motivated by experimental requirements for regression testing, validation of simulation with respect to experimental data, comparison of expected versus reconstructed distributions, and more in general for the comparison of data from different sources.

New requirements have been identified, based on the experience of using the Statistical Toolkit in several analyses for the validation of Geant4 [3, 4] physics models. These projects highlighted the need for complementary functionality, beyond the problem of assessing the compatibility of two distributions.

One of the problems faced in simulation validation (and, more in general, in the comparison of experimental distributions), consists in the identification of possible systematic effects: tests of randomness address this requirement.

Another problem encountered in the experience with the simulation validation consists of the comparison not only of individual data distributions, but also of categories (e.g. the evaluation
of differences in the behaviour of two Geant4 physics models with respect to a set of experimental test cases).

2. Overview of the current functionality of the Statistical Toolkit

Goodness-of-fit tests quantify the compatibility of the agreement between a set of sample observations and the corresponding values predicted from some model of interest, or between two (or more) sets of observations. 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 is true.

The collection of tests implemented in the current version of the Statistical Toolkit is given in table 1; extensive details can be found in [1, 2].

| GoF test          | Distribution Type | <ComparisonAlgorithm> Class                                      |
|-------------------|-------------------|------------------------------------------------------------------|
| Anderson-Darling  | Binned            | AndersonDarlingBinned                                            |
|                   |                   | AndersonDarlingBinnedApproximated                                 |
|                   | Unbinned          | AndersonDarlingUnbinned                                          |
|                   |                   | AndersonDarlingUnbinnedApproximated                               |
| Chi-squared       | Binned            | Chi2                                                             |
|                   |                   | Chi2Approximated                                                 |
|                   |                   | Chi2Integrating                                                  |
| Fisz-Cramer-von-Mises | Binned         | CramerVonMisesBinned                                             |
|                   | Unbinned          | CramerVonMisesUnbinned                                           |
|                   | Unbinned          | WeightedCramerVonMisesBuningUnbinned                             |
| Girone            | Unbinned          | Girone                                                           |
| Goodman           | Unbinned          | KolmogorovSmirnovApproximated                                    |
| Chi-squared       | Unbinned          | KolmogorovSmirnov                                                |
|                   |                   | WeightedADKolmogorovSmirnov                                      |
|                   |                   | WeightedBuningKolmogorovSmirnov                                  |
| Kuiper            | Unbinned          | Kuiper                                                           |
| Tiku              | Binned            | TikuBinned                                                       |
|                   | Unbinned          | TikuUnbinned                                                     |
| Watson            | Unbinned          | Watson                                                            |

3. Software improvements

An effort has been invested to provide an effective software development environment, which exploits more modern tools and facilitates the use of the Statistical Toolkit in a variety of computing environments.

For the new development cycle Subversion (SVN) [6] has been selected as a tool in support of the Configuration and Change Management discipline. The Statistical Toolkit code was moved to a SVN repository.
In order to facilitate using the Statistical Toolkit on different operating systems, the build system has been moved to the Cross Platform Make (CMake) [7] system. The ctest testing tool, distributed as a part of CMake, is used for unit testing.

To be as self-consistent as possible, the number of dependencies on external software systems has been minimized. The only essential external dependency is on the GNU Scientific Library [8].

An additional user layer was implemented to facilitate the use of the Statistical Toolkit in analysis environment that are concerned neither with AIDA [9] nor with ROOT [10] analysis objects, which are supported by the two user layers available in the current version. The new user layer allows the analyst to supply input data to the Statistical Toolkit in the form of comma-separated lists of values (csv ASCII files). If no external dependencies are specified, this user layer is built by default. Otherwise, in properly set-up environments AIDA or ROOT (or both) are found by cmake and the corresponding user layer is built automatically.

The Statistical Toolkit comes with an extensive set of unitTests, which are meant to test the correct implementation of the statistical tests for each new version of the Statistical Toolkit.

4. Extension of functionality
The new development cycle extends the functionality of the Statistical Toolkit with tests for randomness, one sample goodness-of-fit tests, i.e. comparing a data to reference distributions, and tests for categorical data. Table 2 lists the new tests.

| Test                  | Input data             | Class name                                           |
|-----------------------|------------------------|------------------------------------------------------|
| Wald-Wolfowitz runs test | Sequence of signs (-1/1) | WaldWolfowitzTwoSamplesRunsTest WaldWolfowitzOneSampleRunsTest |
| Wald-Wolfowitz test of randomness | 1-dimensional distribution | WaldWolfowitzOneSampleRandomnessTest |
| Mann-Whitney U test | 1-dimensional distribution | MannWhitneyTwoSamplesTest |
| Fisher’s exact test | $2 \times 2$ matrix | FishersExact2x2Test |
| $\chi^2$ contingency test | $c \times r$ matrix | Chi2ContingencyTableTest |
| $\chi^2$ paired test | Paired values | Chi2CurvesComparisonAlgorithm |

4.1. Runs tests of randomness
Randomness tests provide complementary information to the existing goodness of fit tests (table 1): for instance, tests for randomness can highlight the presence of systematic effects in the distributions subject to comparison, which goodness of fit tests cannot detect.

A use case is illustrated in [11]: goodness of fit tests confirm the compatibility of various Geant4 proton elastic scattering models respect to reference data, nevertheless asymmetries in the distribution of differences between the results of the simulation and reference data hint to the presence of systematic effects associated with some of the Geant4 physics models.

The runs tests are statistical tests, used to test the hypothesis that the elements of the sequence are mutually independent or whether the data have some pattern. A run is defined as
a series of values of the same type (e.g. series of increasing/decreasing values, series of true/false values, etc.), the number of consequent values of the same type being the length of the run.

As an example consider tossing a coin and noting the outcome, which is either head (H) or tail (T). A run in this example is each sequence of the same type of outcome. Both too many runs (as in case of cyclic pattern HHTHT..., which has the maximum possible number of runs for given number of observations) and too few runs (where heads and tails are clustered together HH...HTT...T) exhibit evidence of a non-random relationship between the order of the experiments and the outcome.

The Wald-Wolfowitz test from [12] is the best known test that is based on the number of runs. It has been proposed as a test of whether two samples are from the same population, but as such has poor power [13] and the non parametric Mann-Whitney test is preferable.

The new version of the Statistical Toolkit (to be released) encompasses implementations of the the Wald-Wolfowitz runs test for one or two samples. When the test is used with two samples, the algorithm in [12] is used to construct one (binary) sample, and results from the test for one sample are returned.

To calculate the \( p \)-values, either the exact or an approximated formula can be used. The exact calculation of the two-tailed probability of the test statistics implemented in the Statistical Toolkit follows the description from [14], while the approximated formula takes into account that for large samples the distribution of the number of runs approaches a normal distribution.

Wilcoxon [15] published a test for comparison of two samples, based on comparison of the general size of the two samples, ranking of the (combined) samples and then comparing the average ranks of separate ranks. The developments of the test followed fast and the first to publish it were Mann and Whitney [16]. The new version of the Statistical Toolkit implements the Mann-Whitney \( U \) test and an approximated formula for the \( p \)-value calculation, again assuming that the samples are large, hence the distribution of the ranks can be described with a normal distribution.

4.2. Tests for categorical data
Categorical data analysis involves testing the significance of the association (contingency) between the groups. In practice the number of categories is usually small (below 20), although in principle the tests for categorical data could be used for any number of groups.

The difference between the observed and the expected data, considering the given marginal and the assumptions of the model of independence, can be calculated using the \( \chi^2 \) test (already available in the Statistical Toolkit); however, the \( \chi^2 \) test gives only an estimate of the true probability value. The estimate might be inaccurate in case the marginal is very uneven or if there is a small value (less than five) in one of the cells of the contingency table.

Fisher’s exact test for contingency tables [17, 18] is most widely known exact test for categorical data analysis. It is calculated by generating all tables that are more extreme than the table given by the user. To get the two-tailed \( p \)-value, the \( p \)-values of the tables that have \( p \)-values of the same size or smaller than the data table probability are added up to form the cumulative \( p \)-value, including the \( p \)-value of the data table itself. This method becomes computationally intensive already for moderately sized tables, since the number of table probabilities to be enumerated can easily reach billions.

Fisher’s exact test for \( 2 \times 2 \) contingency tables is available in the new development version of the Statistical Toolkit.

An algorithm to calculate the \( \chi^2 \) test with Yates continuity correction has also been implemented as part of the new development cycle.

The \( \chi^2 \) tests can be applied to general \((c \times r)\) contingency tables, while due to computational reasons Fisher’s exact test is only implemented for \( 2 \times 2 \) tables.
5. Conclusions
The new development cycle of the Statistical Toolkit comes with a more versatile build system and provides the user significant extensions in testing capabilities.

The new tests extend the Statistical Toolkit capabilities with tests for randomness and tests for categorical data analysis. The new user layer component makes it possible to use the Toolkit with many spreadsheet applications that allow exporting data directly to comma separated list of values.

New tests, together with the new user layer, make the Statistical Toolkit a powerful data analysis tool for experimental physics problems concerned with data comparisons.

The new developments described in this paper are included in a prototype version of the Statistical Toolkit currently under evaluation. They are planned to be part of the next public release of the toolkit, along with additional new features.

Acknowledgment
This paper is dedicated to the memory of the late Paolo Viarengo, who contributed to the previous development cycles of the Statistical Toolkit.

References
[1] Cirrone G A P, Donadio S, Guatelli S, Mantero A, Mascalino B, Parlati S, Pia M G, Pfeiffer A, Ribon A and Viarengo P 2004 IEEE Trans. Nucl. Sci. 51 2056
[2] Mascalino B, Pfeiffer A, Pia M G, Ribon A and Viarengo P 2006 IEEE Transactions on Nuclear Science 53 3834
[3] Agostinelli S et al 2003 Nucl. Instrum. Meth. A 506 250
[4] Allison J et al 2006 IEEE Trans. Nucl. Sci. 53 270
[5] Jacobson I, Booch G and Rumbaugh J 1999 The unified software development process (Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.) ISBN 0-201-57169-2
[6] Pilato C M et al 2008 Version Control with Subversion O’Reilly Media
[7] Hoffman W and Martin K 2012 Cmake, the cross-platform, open-source build system available online URL http://www.cmake.org
[8] 2012 GSL - GNU Scientific Library available online URL http://www.gnu.org/software/gsl/
[9] Barrand G, Binko P, Duszelmann M, Johnson A and Pfeiffer A 2001 Abstract interfaces for data analysis tools. oai:cds.cern.ch:519004 Tech. Rep. CERN-IT-2001-013 CERN Geneva.
[10] Brun R and Rademakers F 1997 Nucl. Instr. and Meth. A 389 81 – 86 proceedings AHEP ’96 Workshop
[11] Pia M, Begalli M, Lechner A, Quintieri L and Saracco P 2010 IEEE Trans. Nucl. Sci. 57 2805 –2830 ISSN 0018-9499
[12] Wald A and Wollowitz J 1940 Ann. Math. Statist., Ann Arbor, 11 147–162
[13] Zar J H 2007 Biostatistical Analysis (5th Edition) (Upper Saddle River, NJ, USA: Prentice-Hall, Inc.) ISBN 0131008463
[14] Swed F S and Eisenhart C 1943 Ann. Math. Stat. 14 66–87
[15] Wilcoxon F 1945 Biometrics Bulletin 1 80–83
[16] Mann H B and Whitney D R 1947 The Annals of Mathematical Statistics 18 50–60
[17] Fisher R A 1922 J. Royal Stat. Soc. 85 87–94
[18] Fisher R A 1935 J. Royal Stat. Soc. 98 39–82