

The Goodness-of-Fit Statistical Toolkit

Wednesday, 15 February 2006 09:00 (20 minutes)

Statistical methods play a significant role throughout the life-cycle of high energy physics experiments. Only a few basic tools for statistical analysis were available in the public domain FORTRAN libraries for high energy physics. Nowadays the situation is hardly unchanged even among the libraries of the new generation. The present project in progress develops an object-oriented software toolkit for statistical data analysis. The Goodness-of-Fit (GoF) Statistical Comparison component of the toolkit provides algorithms for the comparison of data distributions in a variety of use cases typical of physics experiments. The GoF Statistical 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. A layer for user input from ROOT objects has been easily added recently, thanks to the component-based architecture. The system contains a variety of two-sample GoF tests, from chi-squared to tests based on the maximum distance between the two empirical distribution functions (Kolmogorov-Smirnov, Kuiper, Goodman), to tests based on the weighted quadratic distance between the two empirical distribution functions (Cramer-von Mises, Anderson-Darling).

Thanks to its flexible design the GoF Statistical Toolkit has been recently extended, implementing other less known GoF tests (weighted formulations of Kolmogorov-Smirnov and Cramer-von Mises tests, Watson and Tiku tests). Nowadays the GoF Statistical Toolkit represents the most complete system available for two-sample GoF hypothesis testing, not only in the domain of physics, but even in professional statistics analysis.

The toolkit is open-source and can be downloaded from the web together with user and software process documentation. It is also distributed together with the LCG Mathematical Libraries.

We present the recent improvements and extensions of the GoF Statistical Toolkit; we describe the architecture of the extended system, the new statistics methods implemented, some results of its application, and an outlook towards future developments.

Primary authors: Dr RIBON, Alberto (CERN); Dr PFEIFFER, Andreas (CERN); Dr MASCIALINO, Barbara (INFN Genova); Dr PIA, Maria Grazia (INFN GENOVA); Dr VIARENGO, Paolo (IST Genova)

Presenters: Dr RIBON, Alberto (CERN); Dr PFEIFFER, Andreas (CERN); Dr MASCIALINO, Barbara (INFN Genova); Dr PIA, Maria Grazia (INFN GENOVA); Dr VIARENGO, Paolo (IST Genova)

Session Classification: Poster

Track Classification: Software Components and Libraries