Verification / Validation Tools

B. Mascialino, A. Pfeiffer, M.G. Pia, A. Ribon, P. Viarengo

10th Geant4 Workshop

Bordeaux, November 8th 2005

Goodness of Fit tests

Goodness-of-Fit tests measure:

- the compatibility of a random sample with a given theoretical distribution function as its population (**one-sample problem**)
- the compatibility of two samples (**two-sample problem**)

The problem is concerned with the choice of one of these two alternative hypothesis:

$$H_0 : F_1(x) = F_2(x);$$

$$H_1 : F_1(x) \neq F_2(x), F_1(x) < F_2(x), F_1(x) > F_2(x)$$

The acceptance of the null hypothesis H_0 leads to the exact specification the distribution analysed.

HEP Statistics Project - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://www.ge.infn.it/geant4/analysis/HEPstatistics/index.html What's Related

Statistical Toolkit

- Home
- Statistical Comparison
  - User Requirements
  - Design
  - Documentation
  - Artifacts in detail
  - Download
- Talks
- Papers
- Gallery
- Useful links
  - AIDA
- Team
  - Contact us

The Statistical Toolkit is an open source **software toolkit** for **statistical data analysis.**

It adopts advanced programming techniques (OO technology, generic programming) to achieve openness to extension and evolution. The Toolkit is written in C++.

The project adopts an incremental and iterative software process. The first development cycle focuses on software tools for **Goodness-of-Fit** test.

To facilitate the application in diverse experimental environments, the statistical tools are based on the **AIDA** Abstract Interfaces for Data Analysis and do not rely on any specific implementation of analysis systems. The user may select her/his preferred analysis tool system to co-work with the Statistical System, by loading a shared library of any AIDA-compliant analysis system.

Last modified 26 March 2004 - Maria Grazia Pia

G.A.P Cirrone, S. Donadio, S. Guatelli, A. Mantero, B. Mascialino, S. Parlati, M.G. Pia, A. Pfeiffer, A. Ribon, P. Viarengo

"A Goodness-of-Fit Statistical Toolkit"

IEEE- Transactions on Nuclear Science (2004), 51 (5): 2056-2063.

StatisticsTesting-V1-01-00 release downloadable from the web:

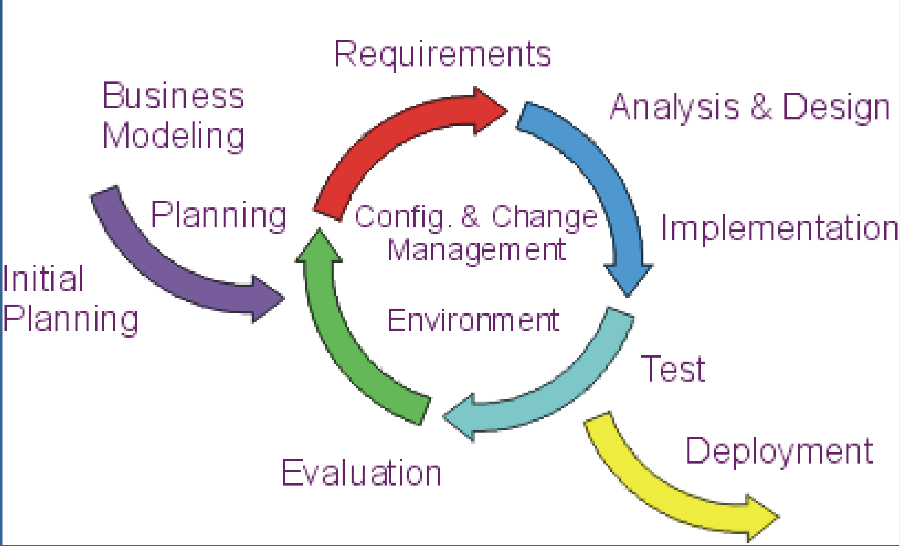
http://www.ge.infn.it/geant4/analysis/HEPstatistics/

Software process guidelines

- Unified Process, specifically tailored to the project
  - practical guidance and tools from the RUP
  - both rigorous and lightweight
  - mapping onto ISO 15504 (and CMM)

- Incremental and iterative life-cycle

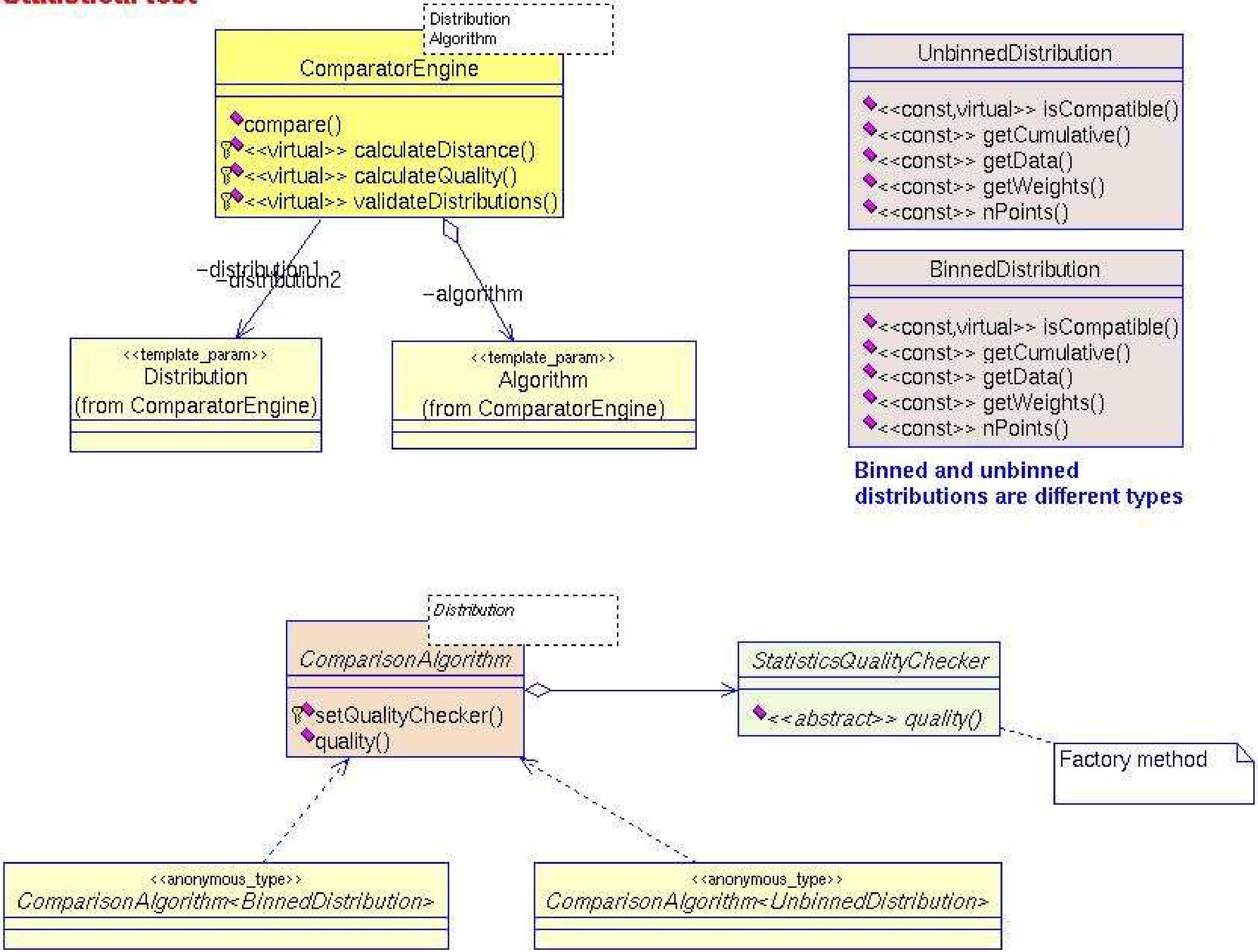
- 1st cycle: 2-sample GoF tests
  - 1-sample GoF in preparation



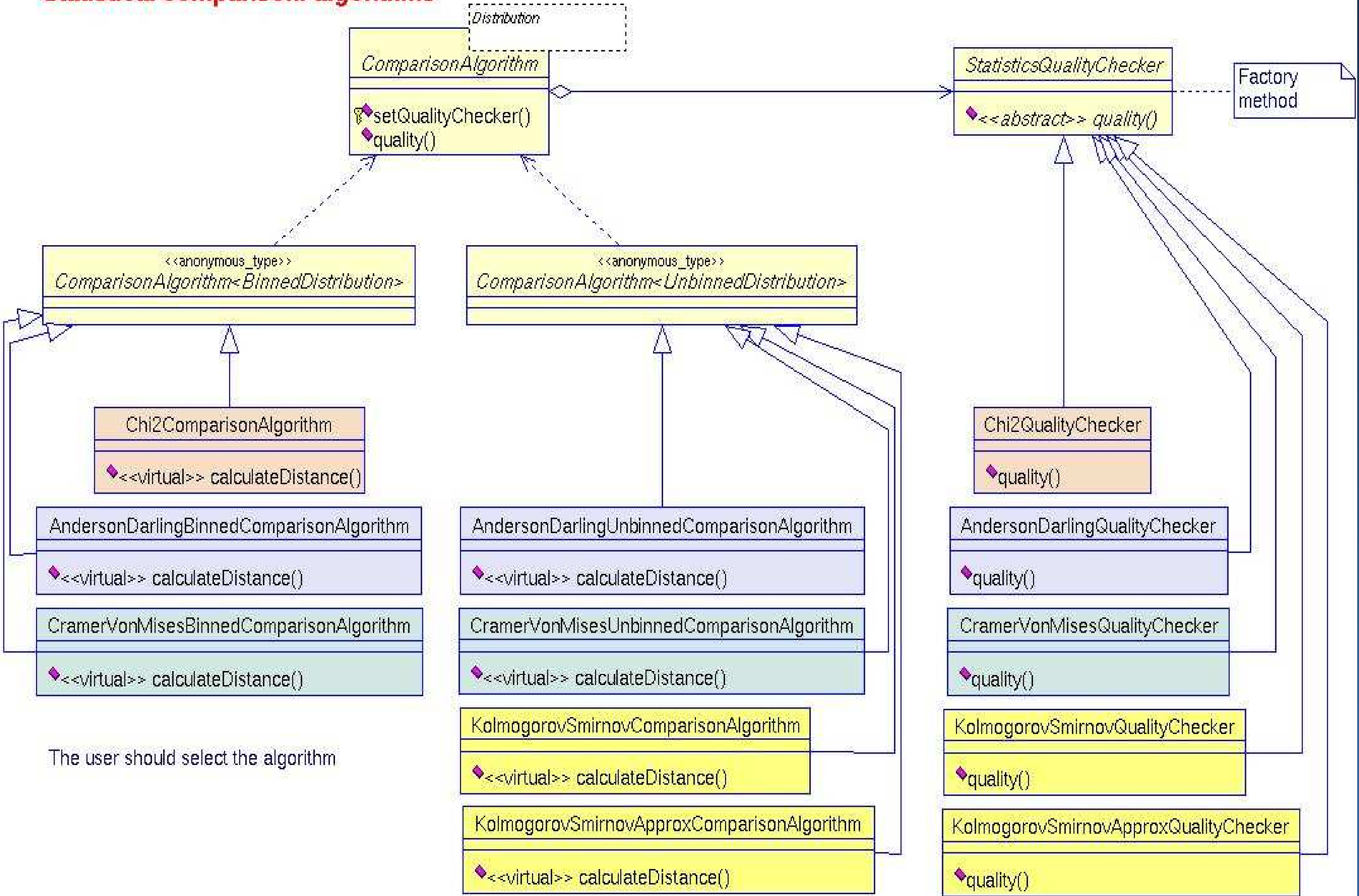
Architectural guidelines

- The project adopts a solid **architectural** approach
  - to offer the *functionality* and the *quality* needed by the users
  - to be *maintainable* over a large time scale
  - to be *extensible*, to accommodate future evolutions of the requirements
- Component-based architecture
  - to facilitate re-use and integration in diverse frameworks
  - layer architecture pattern
  - core component for statistical computation
  - independent components for interface to user analysis environments
- Dependencies
  - no dependence on any specific analysis tool
  - can be used by any analysis tools, or together with any analysis tools
  - offer a (HEP) standard (AIDA) for the user layer

Statistical test



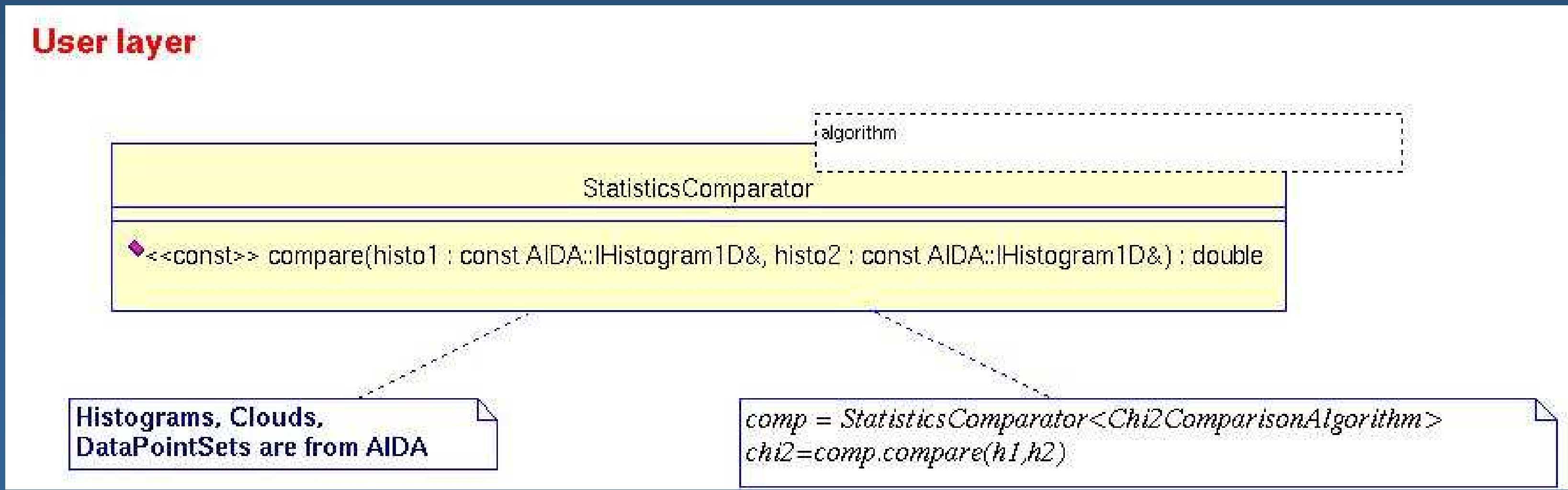
Statistical comparison: algorithms



The user should select the algorithm

User Layer

- **Simple user layer**
- Shields the user from the complexity of the underlying algorithms and design
- Only deal with the user's **analysis objects** and choice of **comparison algorithm**



GoF algorithms (current release)

- Algorithms for binned distributions
  - **Anderson-Darling** test
  - **Chi-squared** test
  - **Fisz-Cramer-von Mises** test
  - **Tiku** test *(Cramer-von Mises test in chi-squared approximation)*
- Algorithms for unbinned distributions
  - **Anderson-Darling** test
  - **Cramer-von Mises** test
  - **Goodman** test *(Kolmogorov-Smirnov test in chi-squared approximation)*
  - **Kolmogorov-Smirnov** test
  - **Kuiper** test
  - **Tiku** test *(Cramer-von Mises test in chi-squared approximation)*

Recent extensions: algorithms

- Fisz-Cramer-von Mises test and Anderson-Darling test
  - exact asymptotic distribution (earlier: critical values)
- New tests:

  - weighted Kolmogorov-Smirnov, weighted Cramer-von Mises
    - various weighting functions available in literature
    - Watson test (can be applied in case of cyclic observations, like Kuiper test)
    - Girone test

- New features: approximated p-value calculation

- chi-squared test
- Anderson-Darling test

- It is the most complete software for the comparison of two distributions, even among commercial/professional statistics tools
  - goal: provide all edf 2-sample GoF algorithms existing in statistics literature

- Publication in preparation to describe the new algorithms

Recent extensions: user layer

- First release: user layer for AIDA analysis object

- July 2005: added user layer for ROOT histograms
  - in response to user requirements

- Other user layer implementations
  - easy to add
  - sound architecture decouples the mathematical component and the user's representation of analysis objects
  - different requirements from various user communities: satisfy them without introducing dependencies on any analysis tools

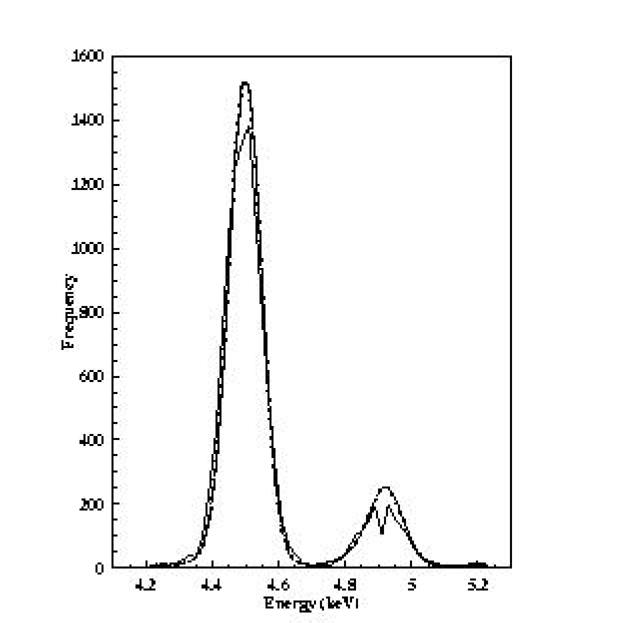
Usage

- Geant4 physics validation
  - quantitative evaluation of Geant4 physics models with respect to established reference data (K. Amako et al., Comparison of Geant4 electromagnetic physics models against the NIST reference data - IEEE Trans. Nucl. Sci. 52- 4 (2005) 910-918)
  - Bragg peak validation in collaboration with the CATANA team – LNS INFN
  - Radioactive decay in collaboration with Luciano Pandola - LNGS INFN
- Space applications
  - Radioprotection applications for manned space missions – REMSIM – Susanna Guatelli INFN Genova
  - ESA test beam validation at Bessy
- Medical applications
  - IMRT – in collaboration with the IST team Genova
- LCG Simulation Validation project – see Alberto Ribon talk!
- CMS
  - validation of "new" histograms w.r.t. "reference" ones in OSCAR Validation Suite

Power of the tests

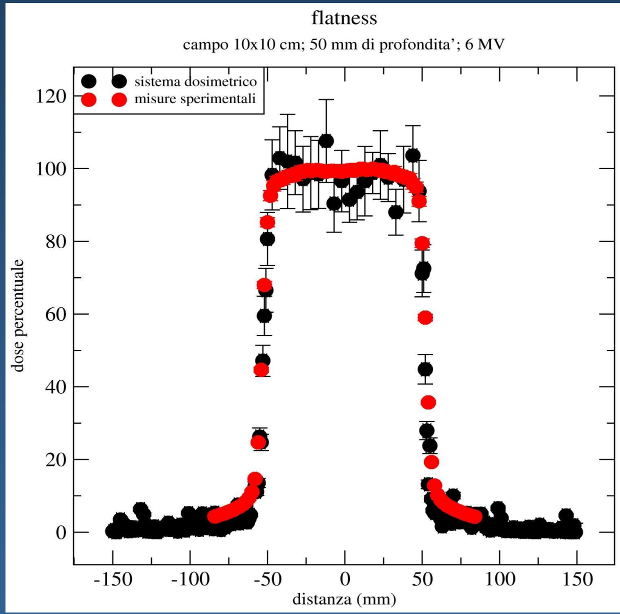
EXAMPLE 1: binned data

X-ray fluorescence spectrum



EXAMPLE 2: unbinned data

Dosimetric distribution from a medical LINAC

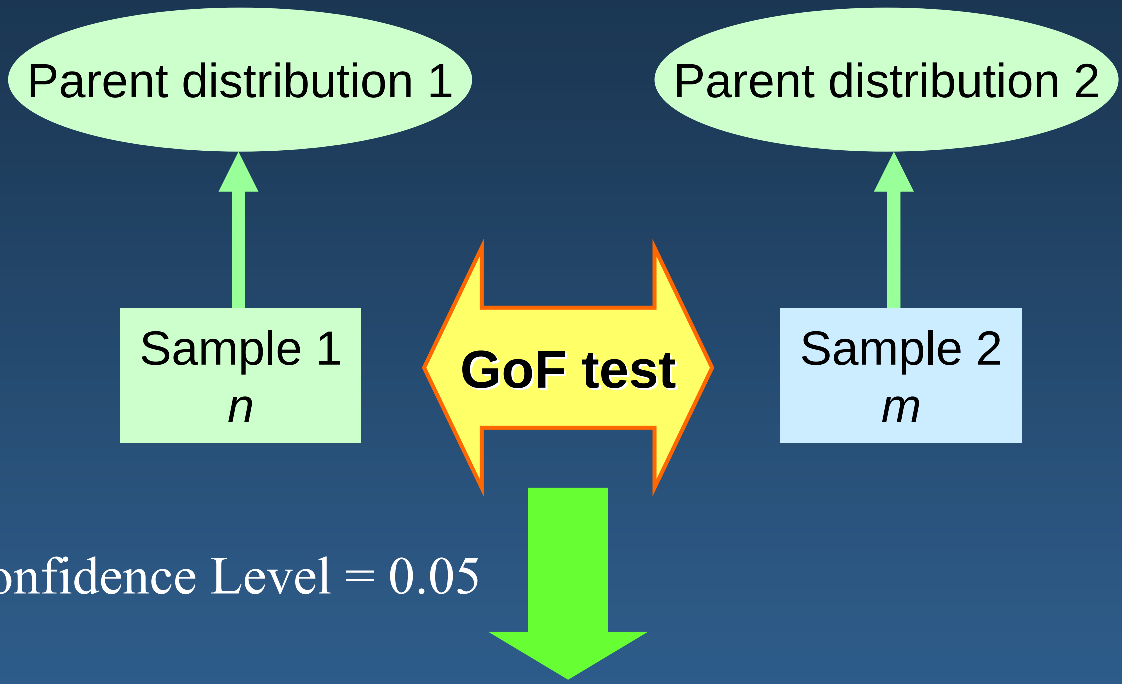


Which is the most suitable goodness-of-fit test?

Systematic study of GoF tests

- No comprehensive study of the relative power of GoF tests exists in literature
  - **novel research in statistics** (not only in physics data analysis!)
- Systematic study of all existing GoF tests in progress
  - made possible by the extensive collection of tests in the Statistical Toolkit
- Provide **guidance to the users** based on sound quantitative arguments
- Preliminary results available
- Publication in preparation

Method for the evaluation of power



Pseudoexperiment:
a random drawing of two samples from two parent distributions

N=1000 Monte Carlo replicas

Confidence Level = 0.05

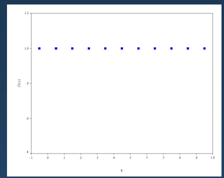
$$\textbf{Power} = \frac{\#\ \text{pseudoexperiments with p-value} < (1\text{-CL})}{\#\ \text{pseudoexperiments}}$$

For each test, the p-value computed by the GoF Toolkit derives from the analytical calculation of the asymptotic distribution, often depending on the samples sizes

Parent distributions

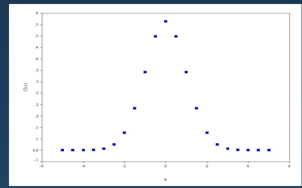
Uniform

$$f_1(x) = 1$$



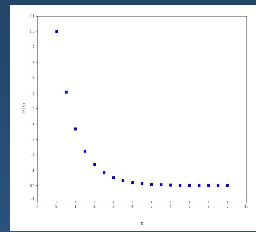
Gaussian

$$f_2(x) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{x^2}{2})}$$



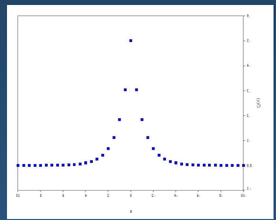
Exponential

$$f_5(x) = e^{-x}$$



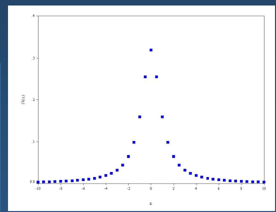
Double exponential

$$f_3(x) = \frac{1}{2} e^{-|x|}$$



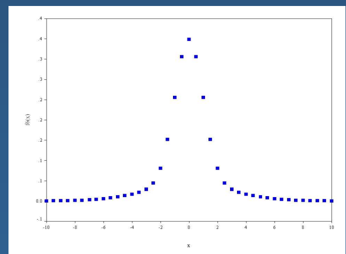
Cauchy

$$f_4(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$



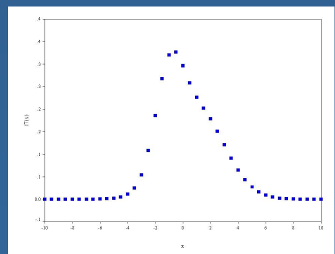
Contaminated Normal Distribution 1

$$f_6(x) = 0.9\mathrm{N}(0,1) + 0.1\mathrm{N}(0,9)$$



Contaminated Normal Distribution 2

$$f_7(x) = 0.5\mathrm{N}(1,4) + 0.5\mathrm{N}(-1,1)$$



Also Breit-Wigner, other distributions being considered

Characterization of distributions

Skewness

$$S = \frac{x_{0.975} - x_{0.5}}{x_{0.5} - x_{0.025}}$$

Tailweight

$$T = \frac{x_{0.975} - x_{0.025}}{x_{0.875} - x_{0.125}}$$

Parent distribution	S	T
$f_1(x)$ Uniform	1	1.267
$f_2(x)$ Gaussian	1	1.704
$f_3(x)$ Double exponential	1	2.161
$f_4(x)$ Cauchy	1	5.263
$f_5(x)$ Exponential	4.486	1.883
$f_6(x)$ Contamined normal 1	1	1.991
$f_7(x)$ Contamined normal 2	1.769	1.693

Comparative evaluation of tests

Preliminary

Tailweight

Skewness

	Short (T<1.5)	Medium (1.5 < T < 2)	Long (T>2)
S~1	KS	KS – CVM	CVM - AD
S>1.5	KS - AD	AD	CVM - AD

Preliminary results

- No clear winner for all the considered distributions in general
  - the performance of a test depends on its intrinsic features as well as on the features of the distributions to be compared
- Practical recommendations
  1) first classify the type of the distributions in terms of skewness and tailweight
  2) choose the most appropriate test given the type of distributions
- Systematic study of the power in progress
  - for both binned and unbinned distributions
- Topic still subject to research activity in the domain of statistics
- Publication in preparation

Outlook

- 1-sample GoF tests (comparison w.r.t. a function)
- Comparison of two/multi-dimensional distributions
- Systematic study of the power of GoF tests
- Goal to provide **an extensive set** of algorithms so far published in statistics literature, with a critical evaluation of their relative strengths and applicability
- Treatment of errors, filtering
- New release coming soon
- New papers in preparation

Conclusions

- A novel, complete software software toolkit for statistical analysis is being developed
  - rich set of algorithms
  - rigorous architectural design
  - rigorous software process
- A systematic study of the power of GoF tests is in progress
  - unexplored area of research
- Application in various domains
  - Geant4, HEP, space science, medicine…