A Toolkit for Statistical Data Analysis

S. Donadio, F. Fabozzi, L. Lista, S. Guatelli, B. Mascialino, A. Pfeiffer, M.G. Pia, A. Ribon, P. Viarengo

PHYSTAT 2003
SLAC, 8-11 September 2003

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

Maria Grazia Pia, INFN Genova
History and background
Validation of Geant4 physics models through comparison of simulation vs experimental data or reference databases

ESA Bepi Colombo mission to Mercury

test beam at Bessy

Electromagnetic models in Geant4 w.r.t. NIST reference

Fluorescence spectrum from Icelandic basalt (Mars-like rock): experimental data and simulation

Photon attenuation coefficient, Al

Maria Grazia Pia, INFN Genova
Some similar use cases

- **Regression testing**
  - Throughout the software life-cycle

- **Online DAQ**
  - Monitoring detector behaviour w.r.t. a reference

- **Simulation validation**
  - Comparison with experimental data

- **Reconstruction**
  - Comparison of reconstructed vs. expected distributions

- **Physics analysis**
  - Comparisons of experimental distributions (*ATLAS vs. CMS Higgs?*)
  - Comparison with theoretical distributions (*data vs. Standard Model*)
Based on considerations such as those given above, as well as considerable computational experience, it is generally believed that tests like the Kolmogorov or Smirnov-Cramer-Von-Mises (which is similar but more complicated to calculate) are probably the most powerful for the kinds of phenomena generally of interest to high-energy physicists. […]

The value of PROB returned by HDIFF is calculated such that it will be uniformly distributed between zero and one for compatible histograms, provided the data are not binned. […]

The value of PROB should not be expected to have exactly the correct distribution for binned data.

CDF Collaboration,
Inclusive jet cross section in p pbar collisions at sqrt(s) 1.8 TeV,
Let’s do it ourselves...

A project to develop an open-source software system for statistical analysis

Provide tools for the statistical comparison of distributions

Interest in other areas, not only Geant4
Not only GoF, but other statistical tools...

Maria Grazia Pia, INFN Genova
The vision
Vision: the basics

- Have a **vision** for the project
  - General purpose tool for statistical analysis
  - Toolkit approach (choice open to users)
  - Open source product

- **Who are the stakeholders?**
  - Who are the **users**?
  - Who are the **developers**?

- **Rigorous software process**

- **Build on a solid architecture**

- **Clearly define scope, objectives**

- **Clearly define roles**

- **Software quality**

- **Flexible, extensible, maintainable system**

Maria Grazia Pia, INFN Genova
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

**Dependencies**
- adopt a (HEP) standard (AIDA) for the user layer
- no dependence on any specific analysis tool

**Python**
- the “glue” for interactivity

The approach adopted is compatible with the recommendations of the LCG Architecture Blueprint Report
- *but the project is independent from LCG*

Maria Grazia Pia, INFN Genova
Software process guidelines

- **Adopt a process**
  - the key to software quality...

- **Significant experience in the team**
  - in Geant4 and in other projects

- **Guidance from ISO 15504**
  - standard!

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

**User requirements** elicited, analysed and formally specified

- Functional *(capability)* and not-functional *(constraint)* requirements
- User Requirements Document available from the web site
  
  http://www.ge.infn.it/geant4/analysis/HEPstatistics/

- Use case model in progress

**Requirement traceability**

- **Requirements**
- **Design**
- **Implementation**
- **Test & test results**
- **Documentation** *(coming...)*
The core Goodness-of-Fit component
In 1933 Kolmogorov published a short but landmark paper on the Italian *Giornale dell’Istituto degli Attuari*. He formally defined the empirical distribution function (EDF) and then enquired how close this would be to the true distribution $F(x)$ when this is continuous.

It must be noticed that Kolmogorov himself regarded his paper as the solution of an interesting probability problem, following the general interest of the time, rather than a paper on statistical methodology.

After Kolmogorov article, over a period of about 10 years, the foundations were laid by a number of distinguished mathematicians of methods of testing fit to a distribution based on the EDF (Smirnov, Cramer, Von Mises, Anderson, Darling, …).

The ideas in this paper have formed a platform for vast literature, both of interesting and important probability problems, and also concerning methods of using the Kolmogorov statistics for testing fit to a distribution. The literature continues with great strength today showing no sign to diminish.

Maria Grazia Pia, *INFN Genova*
Goodness-of-fit tests

- Pearson’s $\chi^2$ test
- Kolmogorov test
- Kolmogorov – Smirnov test
- Goodman approximation of KS test
- Lilliefors test
- Fisz-Cramer-von Mises test
- Cramer-von Mises test
- Anderson-Darling test
- Kuiper test
- ...

It is a difficult domain…
Implementing algorithms is easy
But comparing real-life distributions is not easy
Incremental and iterative software process
Collaboration with statistics experts
Patience, humility, time…

System open to extension and evolution
Suggestions welcome!

Maria Grazia Pia, INFN Genova
Statistical test

Compared Engine

- compare()
- <<virtual>> calculateDistance()
- <<virtual>> calculateQuality()
- <<virtual>> validateDistributions()

Distribution Algorithm

- distribution1
- distribution2
- algorithm

UnbinnedDistribution

- <<const, virtual>> isCompatible()
- <<const>> getCumulative()
- <<const>> getData()
- <<const>> getWeights()
- <<const>> nPoints()

BinnedDistribution

- <<const, virtual>> isCompatible()
- <<const>> getCumulative()
- <<const>> getData()
- <<const>> getWeights()
- <<const>> nPoints()

<<template_param>> Distribution (from ComparatorEngine)

<<template_param>> Algorithm (from ComparatorEngine)

Binned and unbinned distributions are different types

ComparisonAlgorithm

- setQualityChecker()
- quality()

StatisticsQualityChecker

- <<abstract>> quality()

Factory method

<<anonymous_type>> ComparisonAlgorithm<BinnedDistribution>

<<anonymous_type>> ComparisonAlgorithm<UnbinnedDistribution>
Statistical comparison: algorithms

The user should select the algorithm.

Maria Grazia Pia, INFN Genova
Simple user layer
- Shields the user from the complexity of the underlying algorithms and design
- Only deal with **AIDA objects** and choice of **comparison algorithm**

```
<int> compare(histo1 : const AIDA::IHistogram1D&, histo2 : const AIDA::IHistogram1D&) : double
```

Histograms, Clouds, DataPointSets are from AIDA

```
comp = StatisticsComparator<Chi2ComparisonAlgorithm>(
chi2 = comp.compare(h1, h2)
```
The client instantiates the two distributions and the comparison algorithm.
Pearson's $\chi^2$

- Applies to **binned** distributions
- It can be useful also in case of unbinned distributions, but the data must be grouped into classes
- Cannot be applied if the counting of the theoretical frequencies in each class is $< 5$
- When this is not the case, one could try to unify contiguous classes until the minimum theoretical frequency is reached
Kolmogorov test

The easiest among non-parametric tests
Verify the adaptation of a sample coming from a random continuous variable
Based on the computation of the maximum distance between an empirical repartition function and the theoretical repartition one
Test statistics:

\[ D = \sup | F_O(x) - F_T(x) | \]
Problem of the two samples

- mathematically similar to Kolmogorov’s

Instead of comparing an empirical distribution with a theoretical one, try to find the maximum difference between the distributions of the two samples $F_n$ and $G_m$:

$$D_{mn} = \sup |F_n(x) - G_m(x)|$$

Can be applied only to continuous random variables

Conover (1971) and Gibbons and Chakraborti (1992) tried to extend it to cases of discrete random variables
Goodman (1954) demonstrated that the Kolmogorov-Smirnov exact test statistics

\[ D_{mn} = \sup |F_n(x) - G_m(x)| \]

can be easily converted into a \( \chi^2 \):

\[ \chi^2 = 4D^2_{mn} \left[ \frac{m \times n}{(m+n)} \right] \]

This approximated test statistics follows the \( \chi^2 \) distribution with 2 degrees of freedom

Can be applied only to continuous random variables
Lilliefors test

Similar to Kolmogorov test

Based on the null hypothesis that the random continuous variable is normally distributed $N(m,\sigma^2)$, with $m$ and $\sigma^2$ unknown

Performed comparing the empirical repartition function $F(z_1,z_2,...,z_n)$ with the one of the standardized normal distribution $\Phi(z)$:

$$D^* = \sup | F_o(z) - \Phi(z)|$$
Kuiper test

- Based on a quantity that remains invariant for any shift or re-parameterisation
- Does not work well on tails

$$D^* = \max (F_O(x)-F_T(x)) + \max (F_T(x)-F_O(x))$$

- It is useful for observation on a circle, because the value of $D^*$ does not depend on the choice of the origin. Of course, $D^*$ can also be used for data on a line
Fisz-Cramer-von Mises test

- Problem of the two samples
- The test statistics contains a weight function
- Based on the test statistics:
  \[ t = \frac{n_1 \cdot n_2}{(n_1 + n_2)^2} \sum_i [F_1(x_i) - F_2(x_i)]^2 \]
- Can be performed on binned variables
- Satisfactory for symmetric and right-skewed distribution

Cramer-von Mises test

- Based on the test statistics:
  \[ \omega^2 = \text{integral} (F_O(x) - F_T(x))^2 \, dF(x) \]
- The test statistics contains a weight function
- Can be performed on unbinned variables
- Satisfactory for symmetric and right-skewed distributions
Anderson-Darling test

- Performed on the test statistics:

\[ A^2 = \text{integral} \left\{ \frac{[F_O(x) - F_T(x)]^2}{F_T(x) (1-F_T(X))} \right\} dF_T(x) \]

- Can be performed both on **binned** and **unbinned** variables

- The test statistics contains a weight function

- Seems to be suitable to any data-set \((Aksenov and Savageau - 2002)\) with any skewness (symmetric distributions, left or right skewed)

- Seems to be sensitive to **fat tail of distributions**
The study concerns monthly birth and death distributions (binned data)

\[ \chi^2 \text{ test-statistics} = 15.8 \]

\[ \text{Expected } \chi^2 = 15.8 \]

\[ \text{Exact p-value} = 0.200758 \]

\[ \text{Expected p-value} = 0.200757 \]
EXAMPLE FROM CRAMER BOOK

(MATHEMATICAL METHODS OF STATISTICS - page 447)

The study concerns the sex distribution of children born in Sweden in 1935.

**χ² test-statistics = 123.203**

**Expected χ² = 123.203**

**Exact p-value=0**

**Expected p-value=0**
EXAMPLE FROM PICCOLO BOOK (*STATISTICS* - page 711)

The study concerns monthly birth and death distributions (unbinned data)

\[ \chi^2 \text{ test-statistics} = 3.9 \]

\[ \text{Expected } \chi^2 = 3.9 \]

Exact p-value = 0.140974

Expected p-value = 0.140991
EXAMPLE FROM LANDENNA BOOK
(NONPARAMETRIC TESTS BASED ON FREQUENCIES - page 287)

We consider body lengths of two independent groups of anopheles

- Distribution 1
- Distribution 2

χ² test-statistics = 1.5
Expected χ² = 1.5

Exact p-value = 0.472367
Expected p-value = 0.472367
EXAMPLE FROM http://www.physics.csbsju.edu/stats/KS-test.html

The study concerns how long a bee stays near a particular tree (Redwell/Whitney)

D test-statistics = 0.2204
Expected D = 0.2204

Exact p-value = 0.0354675
Expected p-value = 0.035

Maria Grazia Pia, INFN Genova
EXAMPLE FROM LANDENNA BOOK
(NONPARAMETRIC STATISTICAL METHODS - page 318-325)

We consider one clinical parameter of two independent groups of patients

\[
\text{D test-statistics} = 0.65 \\
\text{Expected D} = 0.65
\]

\[
\text{Exact p-value} = 2 \times 10^{-19} \\
\text{Expected p-value} = 8 \times 10^{-19}
\]
...and more

- No time to illustrate all the algorithms and details...

- more at http://www.ge.infn.it/geant4/analysis/HEPstatistic

- The code can be downloaded from the web site
  - instructions for installation and usage

- Further work in progress
  - regular releases with updates, extensions and improvements
  - comprehensive user documentation in progress
  - feedback would be appreciated
Application results

ESA Bepi Colombo mission to Mercury

Electromagnetic models in Geant4 w.r.t. NIST reference

Fluorescence spectrum from Icelandic basalt (Mars-like rock): experimental data and simulation

Anderson-Darling

\[ A_c \text{ (95\%) } = 0.752 \]

Maria Grazia Pia, INFN Genova

\[ \chi^2_{N-L} = 13.1 - \nu = 20 \quad p = 0.87 \]

\[ \chi^2_{N-S} = 23.2 - \nu = 15 \quad p = 0.08 \]
A toolkit for modeling multi-parametric fit problems

F. Fabozzi, L. Lista
INFN Napoli

Initially developed while rewriting a fortran fitter for BaBar analysis

– Simultaneous estimate of:
  - $B(B^+ \rightarrow J/\psi\pi^+) / B(B^+ \rightarrow J/\psi K^+)$
  - direct CP asymmetry

– More control on the code was needed to justify a bias appeared in the original fitter
Requirements

- Provide Tools for modeling parametric fit problems
- Unbinned Maximum Likelihood (UML\[^{[*]}\]) fit of:
  - PDF parameters
  - Yields of different sub-samples
  - Both, mixed
- $\chi^2$ fits
- Toy Monte Carlo to study the fit properties
  - Fitted parameter distributions
    - Pulls, Bias, Confidence level of fit results

[^{*}]: not Unified Modeling Language … 😊 …

New components included in the Statistical Toolkit

Architecture open to extension and evolution
Conclusions
The reason why we are here...

The project is of general interest to the physics community

- This is the reason why we present it here...
  - to establish a scientific discussion on a topic of common interest
  - to see if there are any interested collaborators
  - to see if there are any interested users

- We would all benefit of a collaborative approach to common problems
  - share expertise, ideas, tools, resources…

Maria Grazia Pia, INFN Genova
A project to develop an open source, general purpose software toolkit for statistical data analysis is in progress
  - to provide a product of common interest to user communities

Rigorous software process
  - to contribute to the quality of the product

Component-based architecture, OO methods + generic programming
  - to ensure openness to evolution, maintainability, ease of use

GoF component

Component for modeling multi-parametric fit problems

First implementation and results available
  - toolkit in use for Geant4 physics validation

Open to scientific collaboration

B. Mascialino et al.,
A Toolkit for statistical data analysis

L. Pandola et al.,
Precision validation of Geant4 electromagnetic physics

L. Lista et al.,
A Generic Toolkit for Multivariate Fitting Designed with Template Metaprogramming