# Statistics Toolkit Project

Maria Grazia Pia, INFN Genova

AIDA Workshop

*CERN, 2 July 2003*



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

# History and background

- Activity started as a side project motivated by Geant4 testing
  - comparison of distributions for regression testing and physics validation
- No open source OO system with reliable tools for statistical data comparison on the market
  - $\chi^2$ and Kolmogorov-Smirnov (for binned histograms!) have been the only tools publicly available in HEP for many years...
- Let's write it ourselves to satisfy our own requirements
  - ...and let's offer it as a service to the community
- Projects, experiments and experts contacted
  - interest and collaboration both from statistics experts (F. James and L. Lyons) and physicists in experiments (L. Lista - BaBar, G. Punzi - CDF...)
  - project open to collaboration, feedback from the experiments...

# Vision: the basics

- Have a vision for the project
  - An internal tool for Geant4 physics & ST?
  - A generic system for data comparison?
  - A toolkit for statistical data analysis?

→ Clearly define scope, objectives

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

→ Clearly define roles

- Rigorous software process

→ Software quality

- Build on a solid architecture

→ Flexible, extensible, maintainable system

# 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 approach
  - to facilitate use in diverse frameworks
- AIDA
  - adopt a (HEP) standard
  - no dependence on any specific analysis tool
- Python
- The approach adopted is compatible with the recommendations of the LCG Architecture Blueprint RTAG

# Software process guidelines

- Significant experience in the team
  - in Geant4 and in other projects
- Guidance from ISO 15504
  - standard!
- USDP, specifically tailored to the project
  - practical guidance and tools from the RUP
  - both rigorous and lightweight
  - mapping onto ISO 15504

# Basic strategy

- The 1st cycle of the project provides tools for statistical testing of Geant4
  - needed for physics comparisons and regression testing
  - multiple comparison algorithms
- Generality *(for application also in other areas)* should be pursued
  - facilitated by a component-based architecture
- The statistical tools should be **used** *(in Geant4 and in other frameworks)*
  - tool to be used in testing/analysis frameworks
  - not a framework itself
- Re-use existing tools whenever possible
  - no attempt to re-invent the wheel
  - but critical, scientific evaluation of candidate tools

# GoF component

A project to develop a
**statistical comparison system**,
to be used in Geant4 testing

Main application areas in Geant4:

- physics validation
- regression testing
- system testing

Provide tools for the **statistical comparison** of distributions

- equivalent reference distributions *(for instance, regression testing)*
- experimental measurements
- data from reference sources
- functions deriving from theoretical calculations or from fits

Interest in other areas, not only Geant4

# Goodness-of-fit tests

- Pearson's $\chi^2$ test
- Kolmogorov test
- Kolmogorov – Smirnov test
- Lilliefors 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!

# Current status

- First β-release March 2003
  - mainly to get early feedback and first set-up the release process
  - GoF component
- First release May 2003
  - 1st (ample) set of GoF algorithms + user layer
  - meant to be used by Geant4
  - still limited documentation and examples (they are on the way...)
- Luca's PDF/likelihood component in progress
- News on GoF component from Barbara/Stefania

# Today's meeting

- Objectives for the next months
  - agreement on strategies to adopt
  - design
  - development
  - support (documentation, examples etc.), assistance to users to get started
  - definition and documentation of the software process
  - promotion of the product
- Milestones, tasks, responsibilities
  - proposal: GDPM as project management model
- Preparation for next conferences (PHYSTAT, IEEE-NSS)
  - work and speakers
- Publications and publication policy
  - documentation of the product
  - essential to our younger collaborators (and for our funding agencies...)
  - model: collective code ownership? or independent systems?