A Toolkit for Statistical Data Comparison

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

A. Pfeiffer, A. Ribon
CERN

S. Donadio, S. Guatelli, B. Mascialino, M.G. Pia, P. Viarengo
INFN Genova

CHEP 2003, San Diego (CA), 24-28 March 2003

What is?

A project to develop a
statistical comparison system

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

- equivalent reference distributions
- experimental measurements
- data from reference sources
- functions deriving from theoretical calculations or fits

Main application areas:

- Physics analysis
- Simulation validation
- Detector monitoring
- Regression testing
- Reconstruction vs. expectation

Vision: the basics

- Have a vision for the project
  - Motivated by Geant4
  - First core of a statistics toolkit for HEP?

→ 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

Goodness-of-fit tests

- Pearson's χ^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!

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

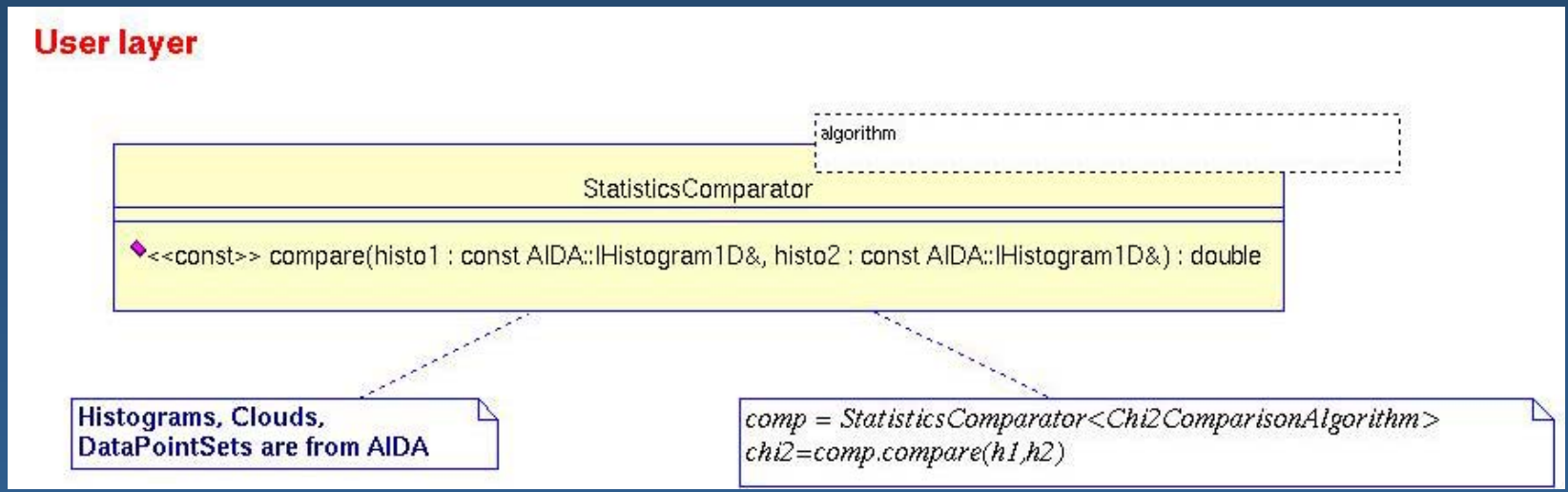
Software process guidelines

- **USDP**, specifically tailored to the project
  - practical guidance and tools from the RUP
  - both rigorous and lightweight
  - mapping onto ISO 15504
- Guidance from **ISO 15504**
  - standard!
- Incremental and iterative life cycle model
- Various software process **artifacts** available on the web
  - Vision
  - User Requirements
  - Architecture and Design model
  - Traceability matrix
  - etc.

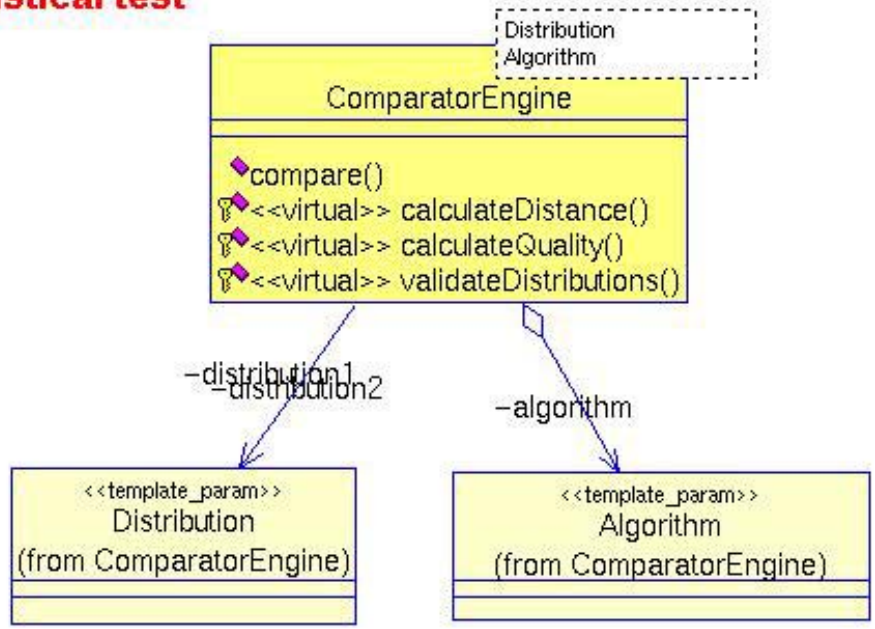
Historical introduction to EDF tests

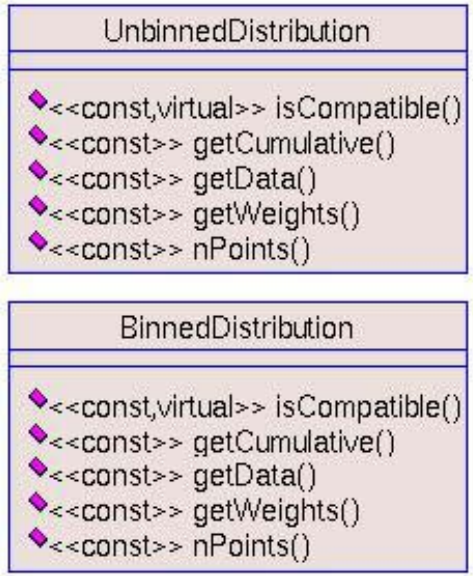
- In 1933 Kolmogorov published a short but landmark paper on the Italian *Giornale dell'Istituto degli Attuari*. He formally defined the empirical distribution function (EDF) and then enquired how close this would be to the true distribution F(x) when this is continuous.
- It must be noticed that Kolmogorov 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.

- **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**

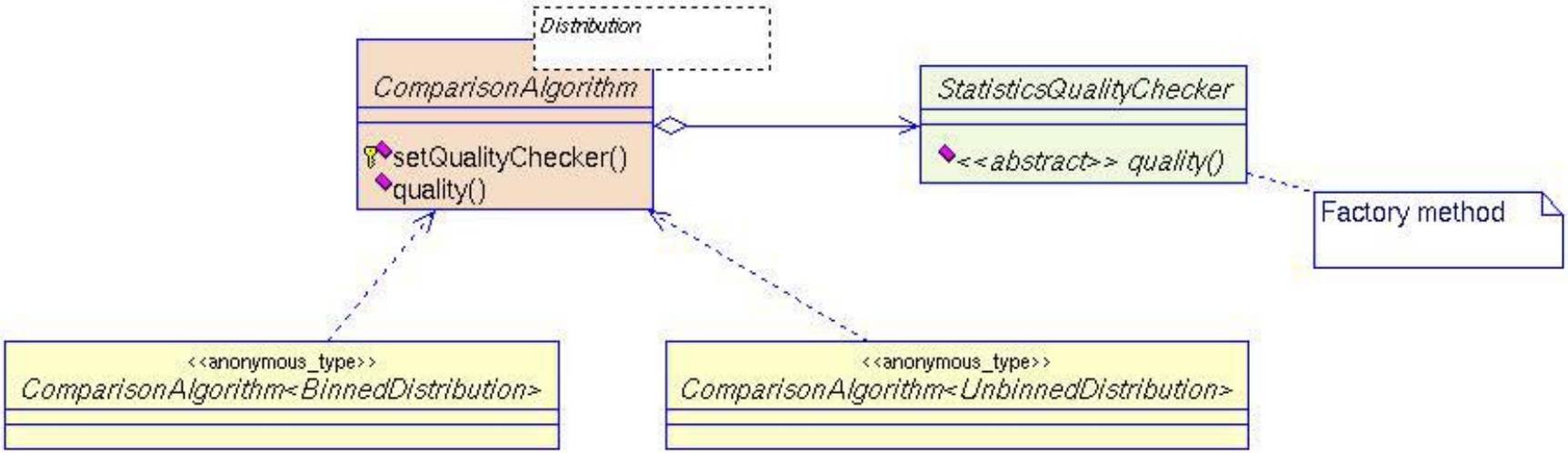


Statistical test

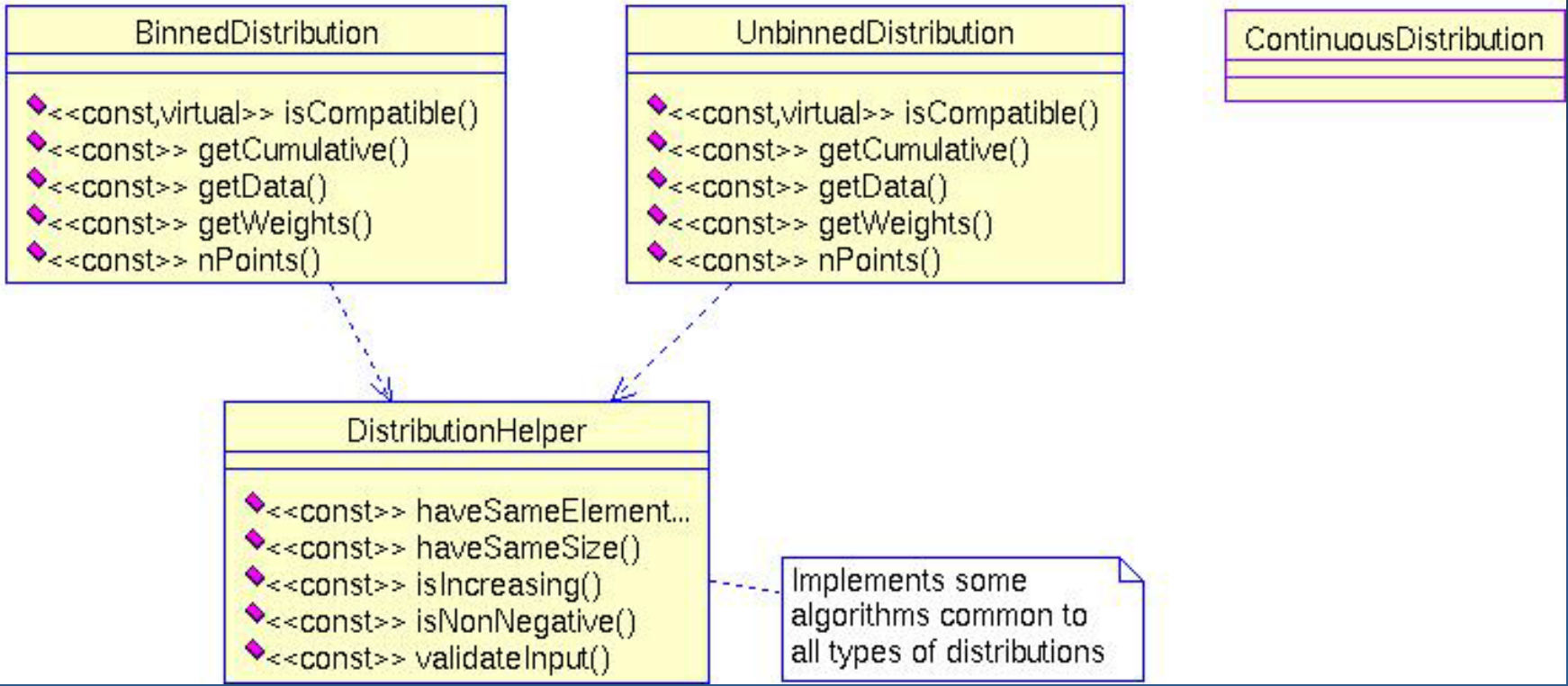




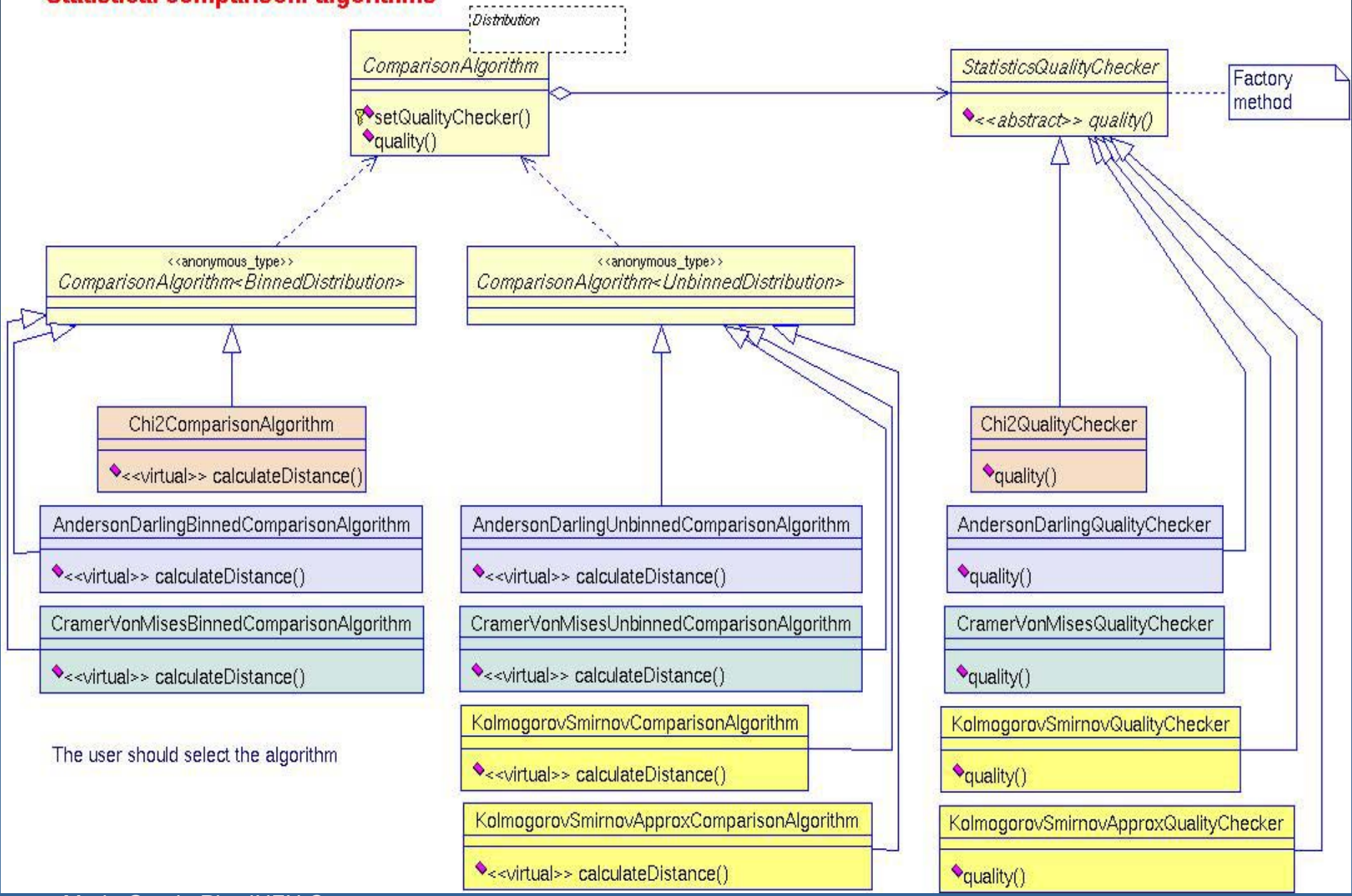
Binned and unbinned distributions are different types



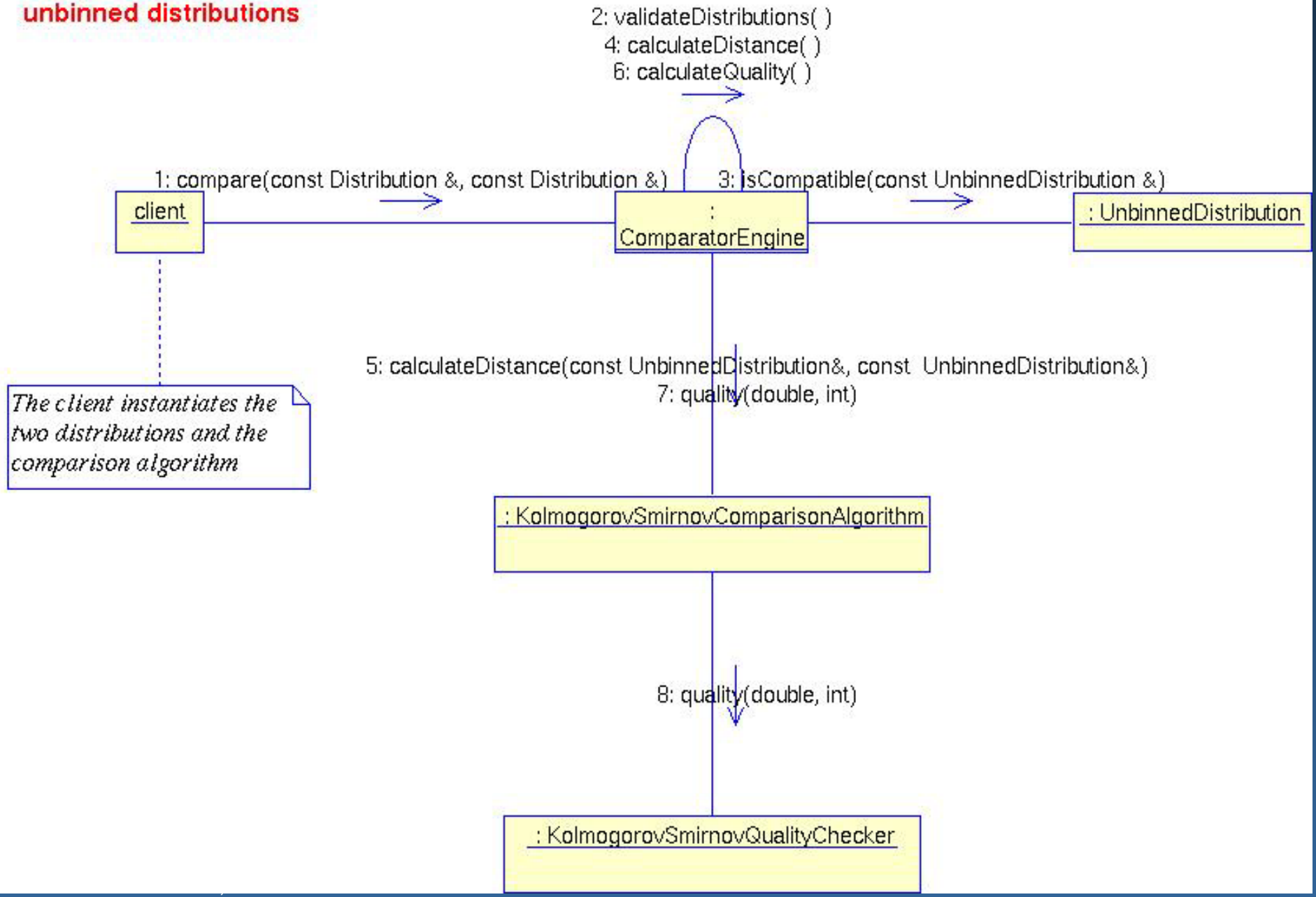
Distributions



Statistical comparison: algorithms



Compare two unbinned distributions



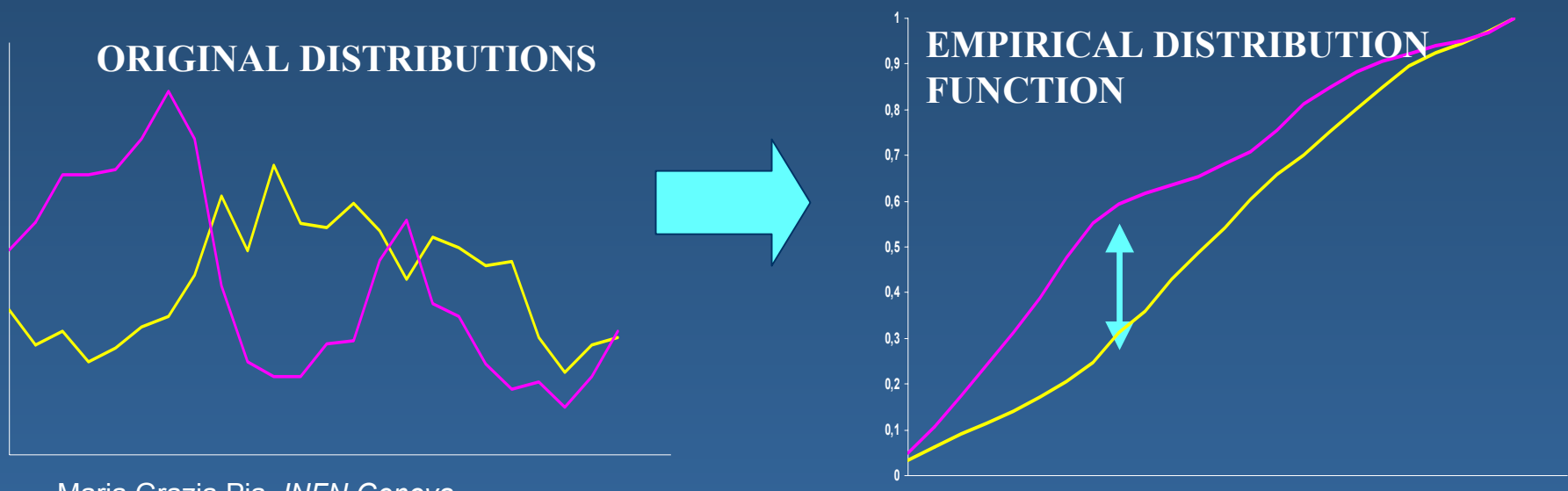
Pearson's χ^2

- 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:

$$D = \sup | F_O(x) - F_T(x)|$$



Kolmogorov-Smirnov test

- 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 approximation of K-S test

- 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 χ^2:

$$\chi^2 = 4D^2_{mn} [m*n / (m+n)]$$

- This approximated test statistics follows the χ^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 σ^2 unknown
- Performed comparing the empirical repartition function $F(z_1,z_2,...,z_n)$ with the one of the standardized normal distribution $\Phi(z)$:

$$D^* = \sup | F_O(z) - \Phi(z)|$$

Fisz-Cramer-von Mises test

- Problem of the two samples
- The test statistics contains a weight function
- Based on the test statistics:

$$t = n_1 * n_2 / (n_1+n_2)^2 \ \Sigma_i [F_1(x_i) - F_2(x_i)]^2$$

- 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:

$$A^2 = \text{integral } \{ [F_O(x) - F_T(x)]^2 / [F_T(x) (1-F_T(X))] \} \, dF_T(x)$$

- 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**

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

Power of the tests

In terms of power:

The power of a test is the probability of rejecting correctly the null hypothesis

χ^2 < Kolmogorov-Smirnov < Tests containing a weight function

- χ^2 loses information in a test for continuous distribution by grouping the data into cells
  - *Kac, Kiefer and Wolfowitz (1955) showed that D requires $n^{4/5}$ observations compared to n observations for χ^2 to attain the same power*
- Cramer-von Mises and Anderson-Darling statistics are expected to be superior to D, since they make a comparison of the two distributions all along the range of x, rather than looking for a marked difference at one point

Status

- First β release last week, can be downloaded from http://www.ge.infn.it/geant4/HEPstatistics
  - χ^2
  - Kolmogorov-Smirnov *(exact and Goodman's)*
  - Anderson-Darling *(discrete)*
  - Fisz-Cramer-von Mises *(discrete)*
- Used in the test process of Geant4 electromagnetic physics
  - Talk on Monday 24th, Precision Validation of Geant4 Electromagnetic Physics
- and in Geant4 regression testing
- Various experimental groups interested
- Open to suggestions, collaboration – feedback welcome!
- Work will continue…

Preliminary results

- Results from unit tests
  - test cases from well known reference sources (books, NIST web site)
  - comparison of results: software vs reference
- Only a few examples of results shown in the next slides
  - extensive set of unit and system tests to validate the code

Unit tests: χ^2 (1)

EXAMPLE FROM PICCOLO BOOK (*STATISTICS* - page 711)

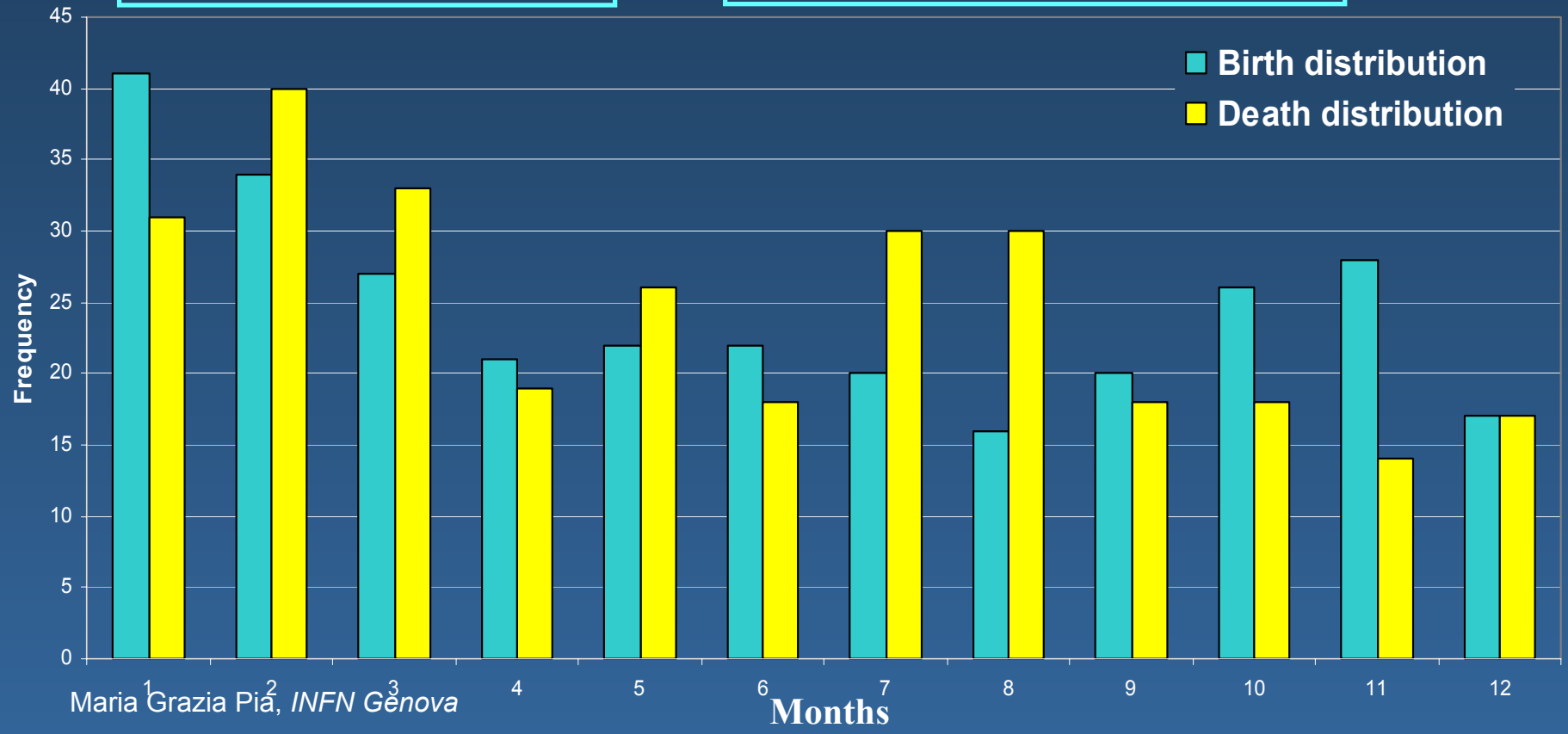
The study concerns monthly birth and death distributions

χ^2 test-statistics = 15.8

Expected χ^2 = 15.8

Exact p-value=0.200758

Expected p-value=0.200757



Unit tests: χ^2 (2)

EXAMPLE FROM CRAMER BOOK

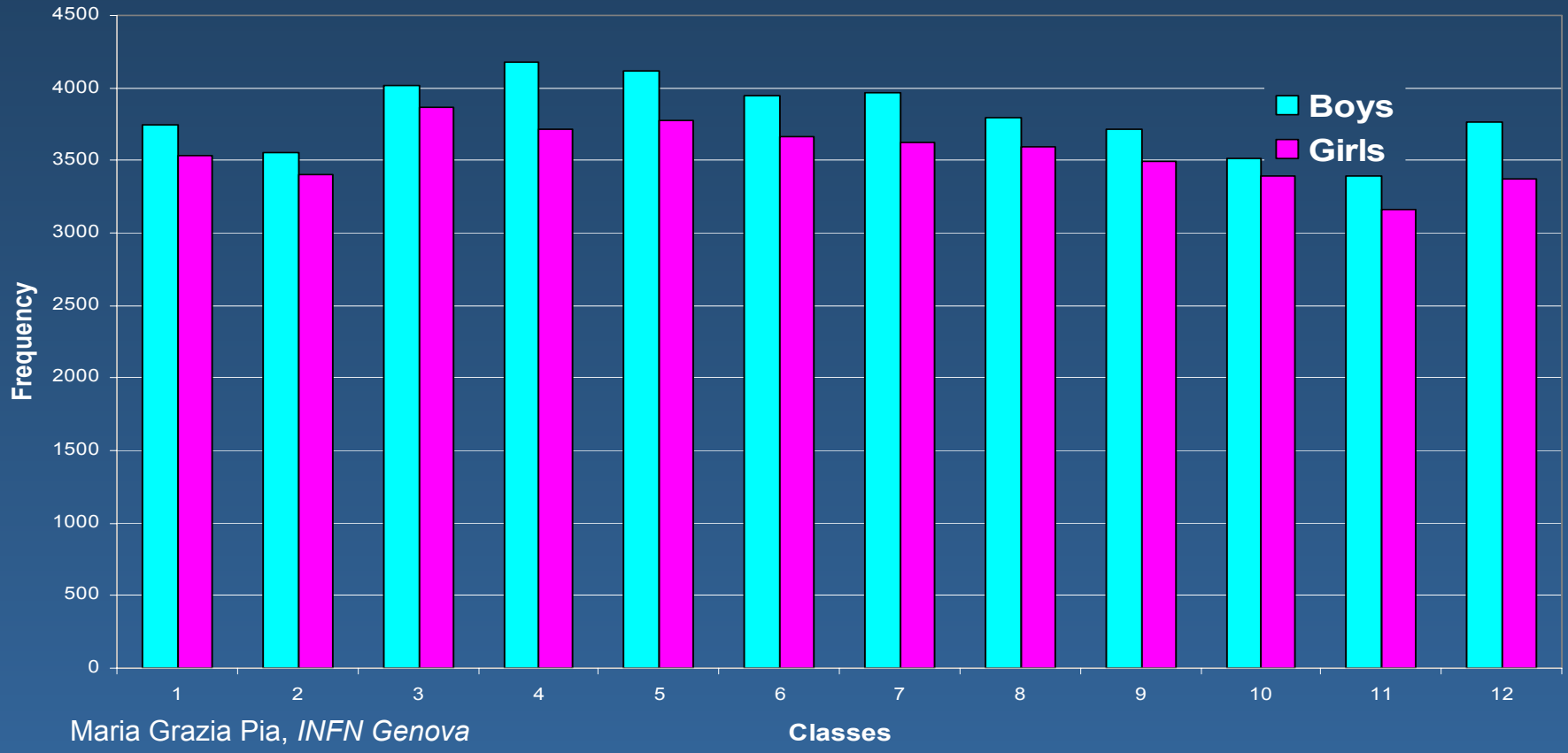
(*MATHEMATICAL METHODS OF STATISTICS* - page 447)

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

χ^2 test-statistics = 123.203
Expected χ^2 = 123.203

Exact p-value=0

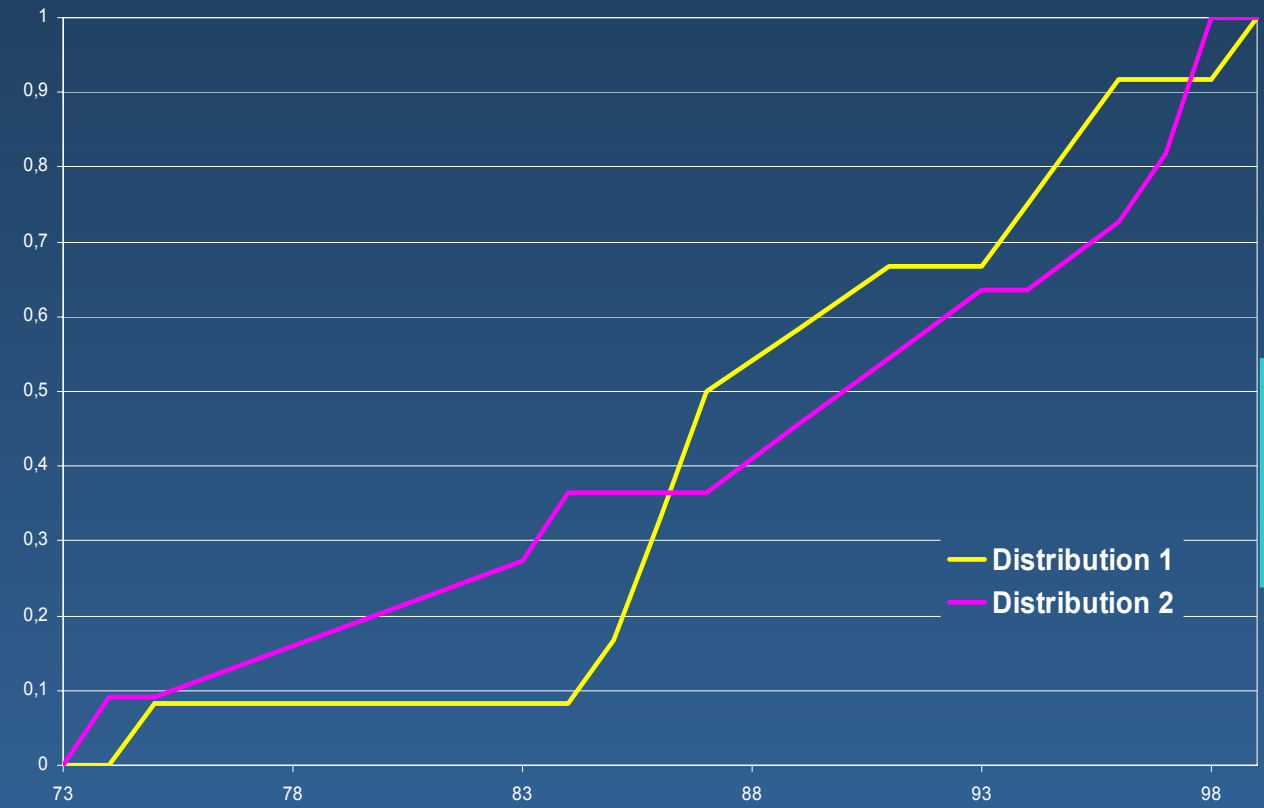
Expected p-value=0



Unit tests: K-S Goodman

EXAMPLE FROM LANDENNA BOOK
(*NONPARAMETRIC TESTS BASED ON FREQUENCIES* - page 287)

We consider body lengths of two independent groups of anopheles



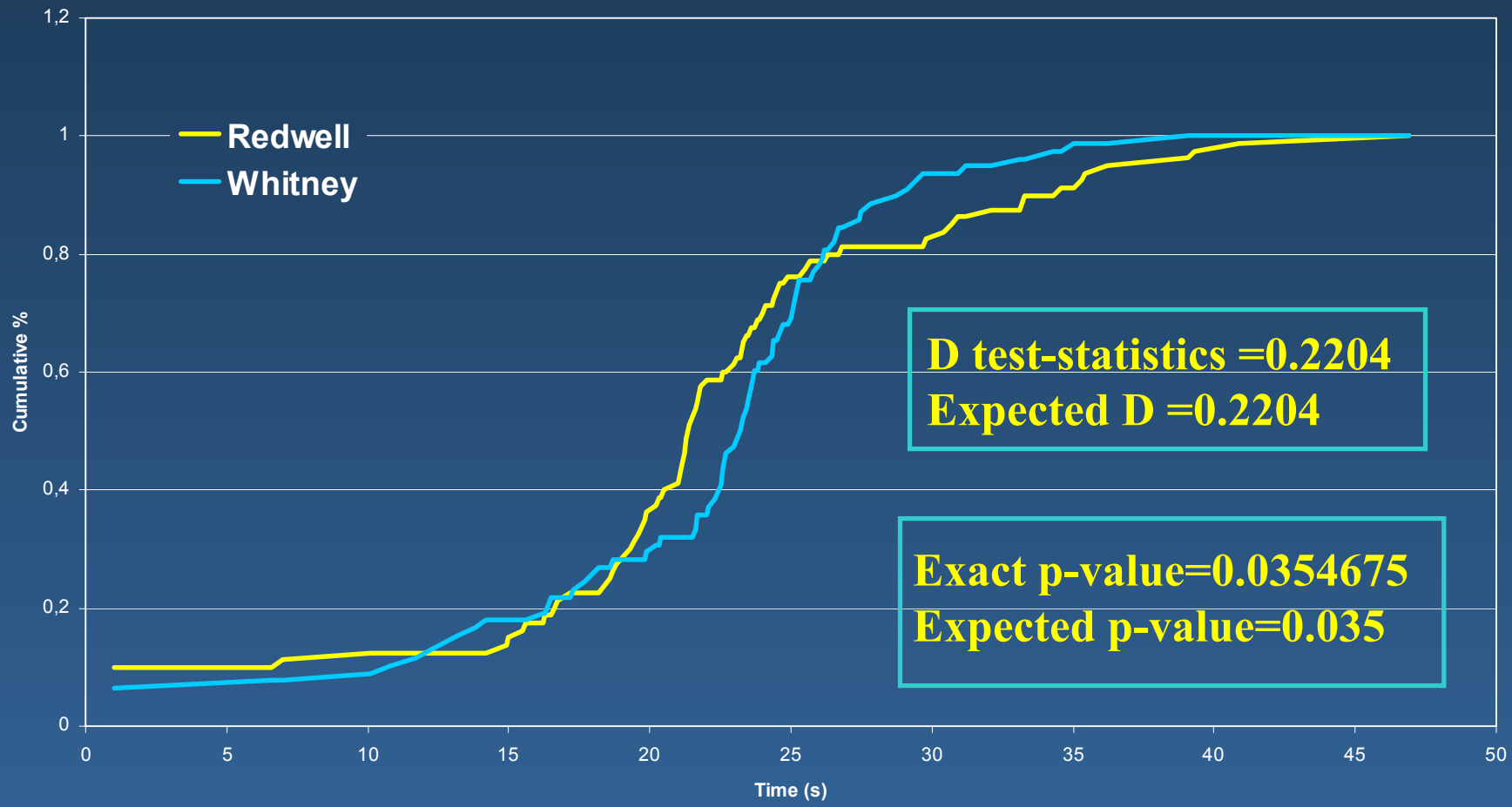
χ^2 test-statistics = 1.5
Expected χ^2 = 1.5

Exact p-value=0.472367
Expected p-value=0.472367

Unit tests: Kolmogorov-Smirnov

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)

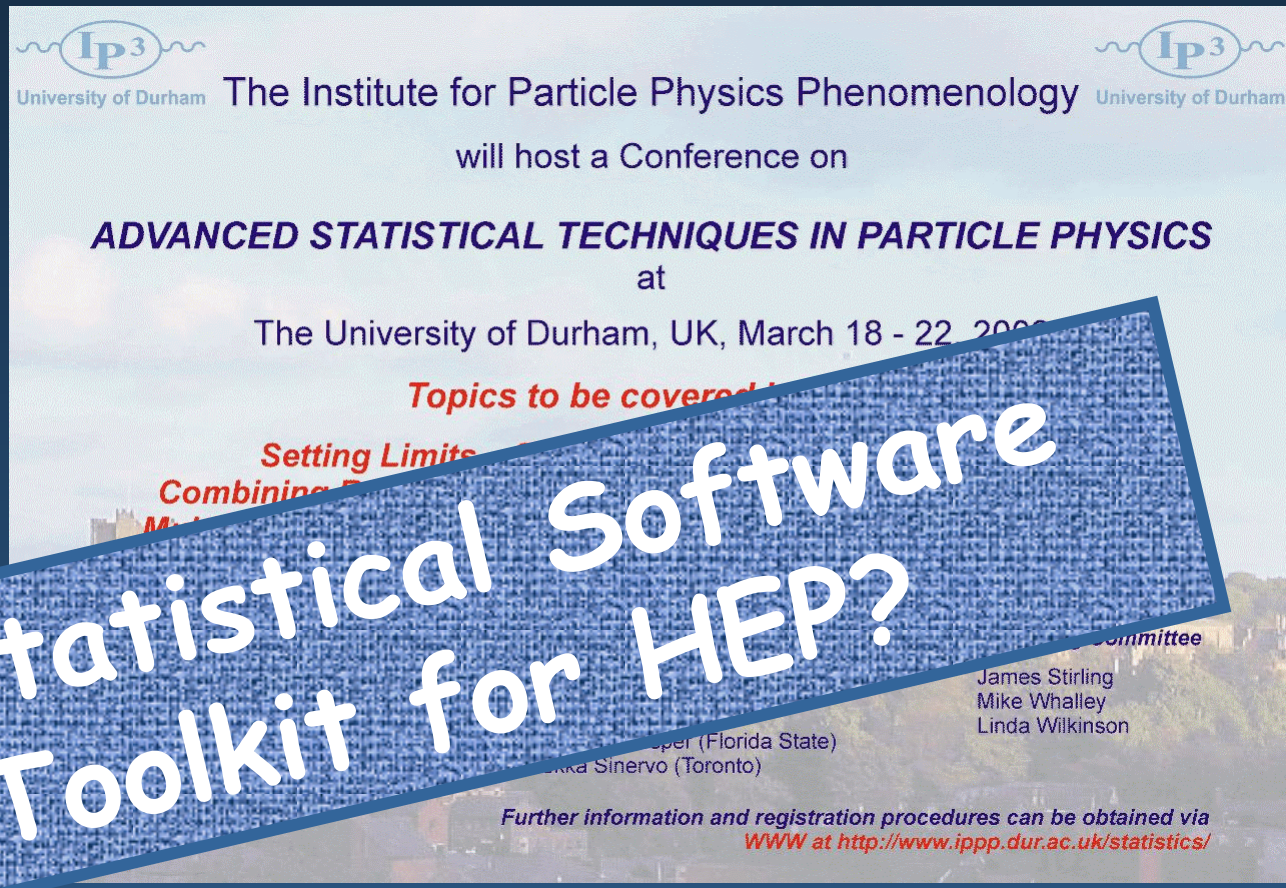


A lot of interest

Significant knowledge and expertise in the HEP community

Comm
softwa
exper

A Statistical Software Toolkit for HEP?



CERN WORKSHOP Confidence Limits

Workshop on 'Confidence Limits'
17-18 January, 2000
CERN Council Chamber

WORKSHOP Confidence Limits

Workshop on Confidence Limits
27-28 March, 2000
Fermilab 1-West Conference Room

Maria Grazia