Basic Example - 2

Estimer un paramètre à partir de modèles partiellement connus

Autrans, 17-21 mai 2010

Jérôme Baudot,
baudot@in2p3.fr

IPHC
Institut Pluridisciplinaire
Hubert CURIEN
STRASBOURG

D'après :

• Roger Barlow & Christine Beeston, "Fitting using Finite Monte Carlo Samples", Comp.Phys.Com. 77 (1993) 219-228

Le problème

- Distribution d'UNE variable avec PLUSIEURS sources
  - ✗ N échantillons d'une grandeur x = données réelles
  - ✗ K sources (par exemple : signal+bruit ou plusieurs saveurs,...)
  - ✗ chaque source a une distribution de x, fj(x) différente des autres
  - ✗ Les fj(x) sont inconnues analytiquement MAIS leur forme est simulée par un échantillon Monte Carlo de taille Nj
  - ✗ Quelles sont les proportions p_j de chacune de K sources dans la distribution réelle de taille N ?
    - ➔ Condition de normalisation $\sum_{j=1}^{K} p_j = 1$

Notations

- Le problème est binné :
  - On suppose qu'il y a B bins
  - Données réelles :
    - n_i où i=1..B est le nombre d'entrées dans le bin i
    - $N = \sum_{i=1}^{B} n_i$
  - Monte Carlo, source j (j=1..K) :
    - s_{ji} où i=1..B est le nombre simulé d'entrées dans le bin i
    - $N_j = \sum_{i=1}^{B} s_{ji}$
  - Prédiction d'après les modèles : f_i est le nombre prédit d'entrées dans le bin i pour un ensemble $\{p_j\}$ de proportion
    - $f_i = N \sum_{j=1}^{K} \frac{p_j \times s_{ji}}{N_j}$

Méthodes d'estimation

- Estimateur des moindres carrées
  - Les estimateurs p_j sont ceux qui minimisent :

$$s^2 = \frac{\sum_{i=1}^{B} (n_i - f_i)^2}{n_i + N^2 \sum_j \frac{s_{ji}}{N_j^2}}$$

    - où l'incertitude associée à n_i est gaussienne et vaut $\sqrt{n_i}$
    - où l'incertitude associée à f_i est gaussienne
  - Problème des faibles n_i
    - L'erreur n'est plus gaussienne → erreur sur les estimateurs délicate
- Estimateur du maximum de vraisemblance
  - Chaque f_i suit une loi de Poisson $f(n_i) = e^{-f_i} \frac{f_i^{n_i}}{n_i!}$
  - Fonction de vraisemblance binnée $\ln L(\vec{p}|\vec{s}\,\vec{n}) = \sum_{i=1}^{B} n_i \ln(f_i) - f_i$
  - Problème des fluctuations de s_{ji}
    - Pas prises en compte !

Méthode d'estimation « complète »

- Complétons la vraisemblance
  - ✗ Prise en compte d'inconnues supplémentaires :
    - → les a_{ij} = « vraies » valeurs attendues pour les s_{ji}
    - → chaque s_{ij} suit une loi de Poisson : $f(s_{ji}) = e^{-a_{ji}} \frac{a_{ji}^{s_{ji}}}{s_{ji}!}$
    - → Possible si nombre de bins suffisant et corrélations binomiales négligeables
  - ✗ Nouveau terme de vraisemblance
    - → $$\ln L(\vec{a}|\vec{s}) = \sum_{j=1}^{K} \sum_{i=1}^{B} (s_{ji} \ln(a_{ji}) - a_{ji})$$
    - → Additif avec le premier terme en f_i
  - ✗ Vraisemblance complète :
    - → $$\ln L(\vec{p}|\vec{s}\,\vec{n}) = \sum_{i=1}^{B} (n_i \ln(f_i) - f_i) + \sum_{j=1}^{K} \sum_{i=1}^{B} (s_{ji} \ln(a_{ji}) - a_{ji})$$
    - → Il y a maintenant K+KxB paramètres (p_j et a_{ji}), même si seuls les p_j nous intéressent vraiment

Pour trouver le maximum

- Dérivons par rapport aux pj et aux aji
  - $\sum_{i=1}^{B} \frac{n_i a_{ji}}{f_i} = 0 \; \forall \; j$

    $\frac{N p_j}{N_j} \left(\frac{n_i}{f_i} - 1 \right) + \frac{s_{ji}}{a_{ji}} - 1 = 0 \; \forall \; i, j$

  - Ensemble d'équations couplées…qui peuvent se simplifier en
    - K et B équations découplées
    - Solution numérique itérative possible

- Deux propriétés de la solution
  - Normalisation des sources conservée

    $\sum_{i=1}^{B} a_{ji} = \sum_{i=1}^{B} s_{ji} \; \forall \; j$

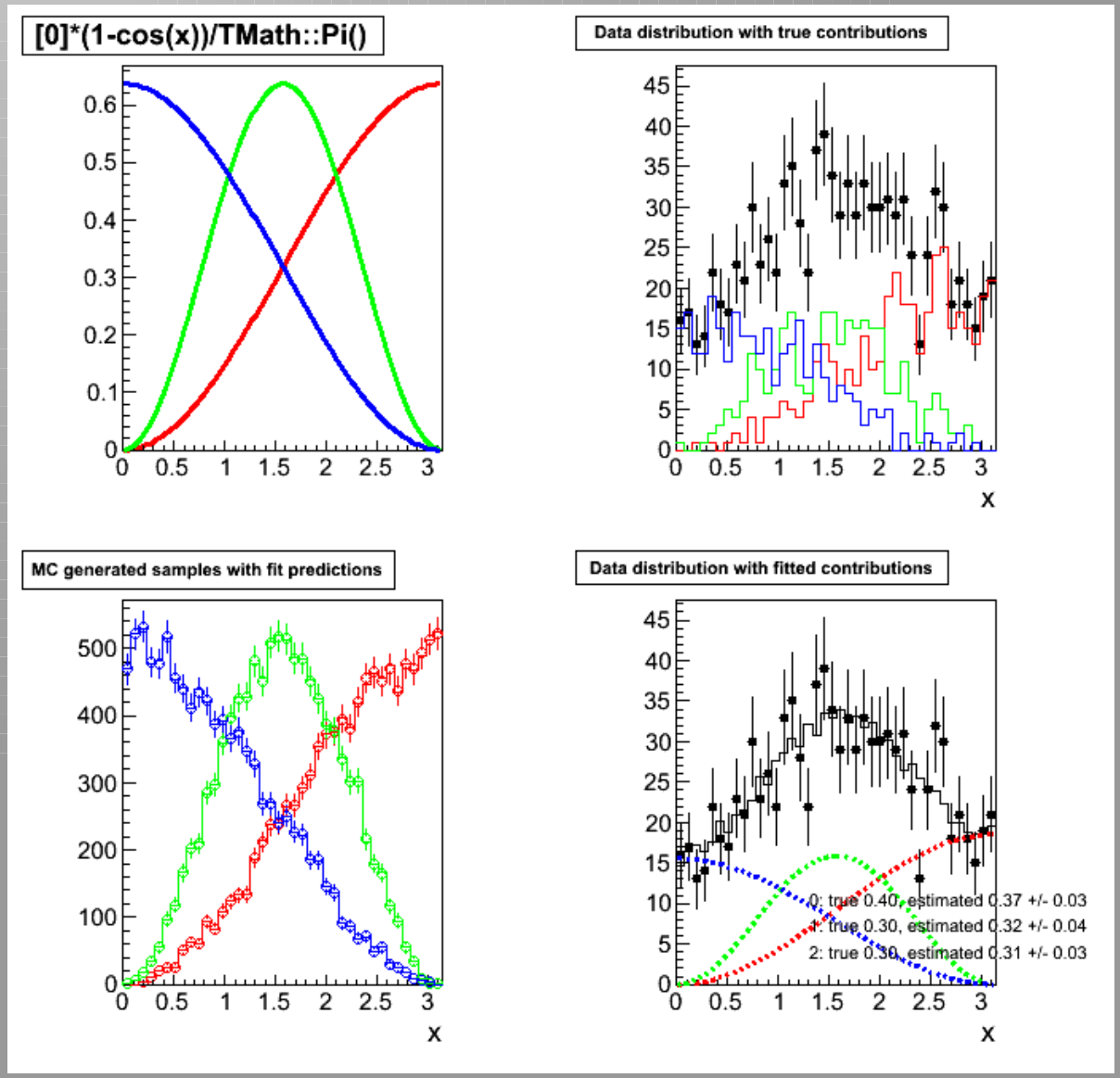
  - Normalisation des proportions assurée

    $\sum_{j=1}^{K} p_j = 1$

TFractionFitter

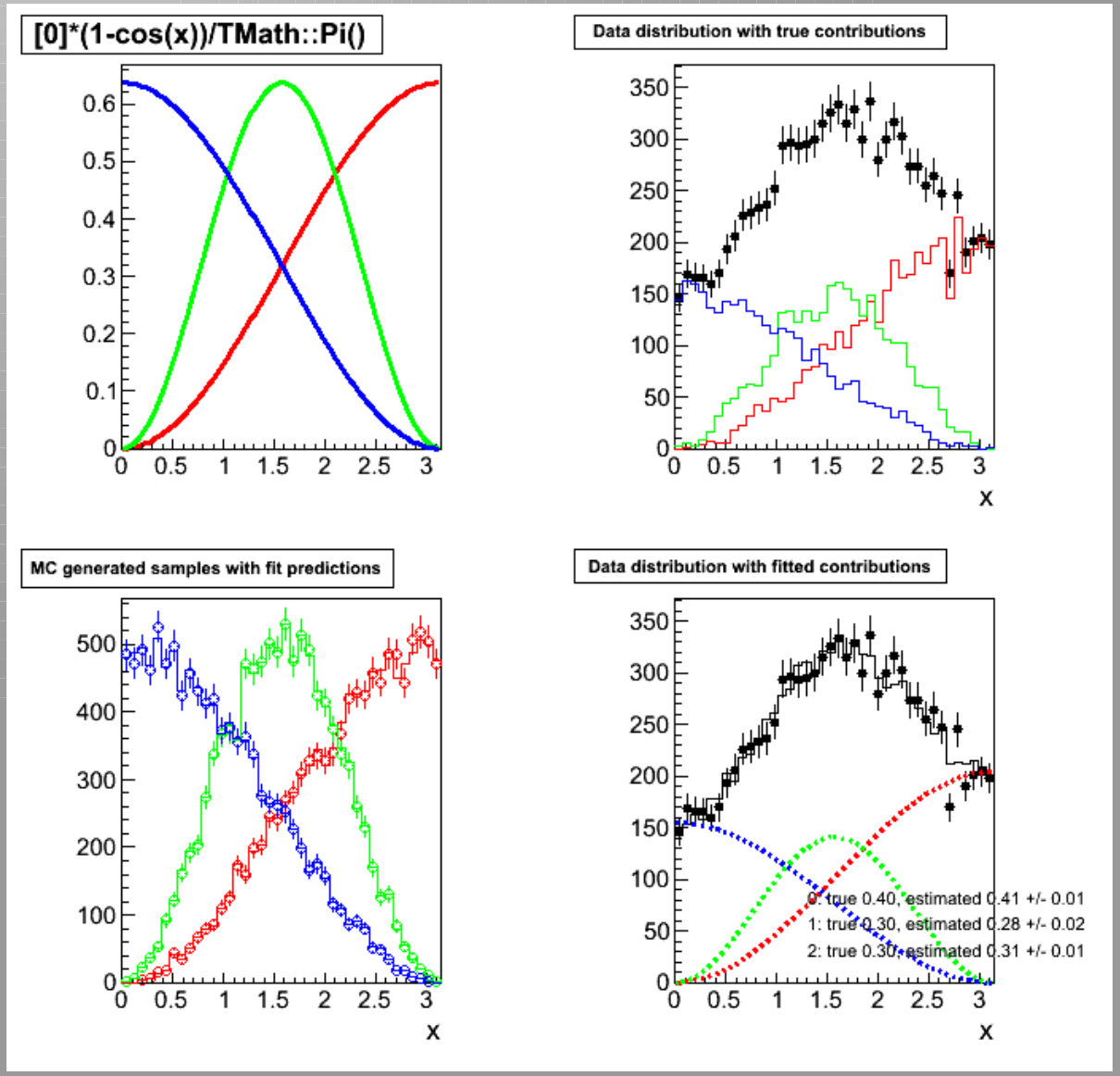
- Un exemple avec 3 contributions pour un angle [0, π]
  - Source 0: $f_0(\theta)=\frac{1-\cos(\theta)}{\pi}$
  - Source 1: $f_1(\theta)=2\frac{1-\cos^2(\theta)}{\pi}$
  - Source 2: $f_2(\theta)=\frac{1+\cos(\theta)}{\pi}$
- TfractionFitter est une classe de ROOT
  - Mise en application de l'article de Barlow & Beeston
  - Macro fractionFitterExample.C

N(MC) = N(data)



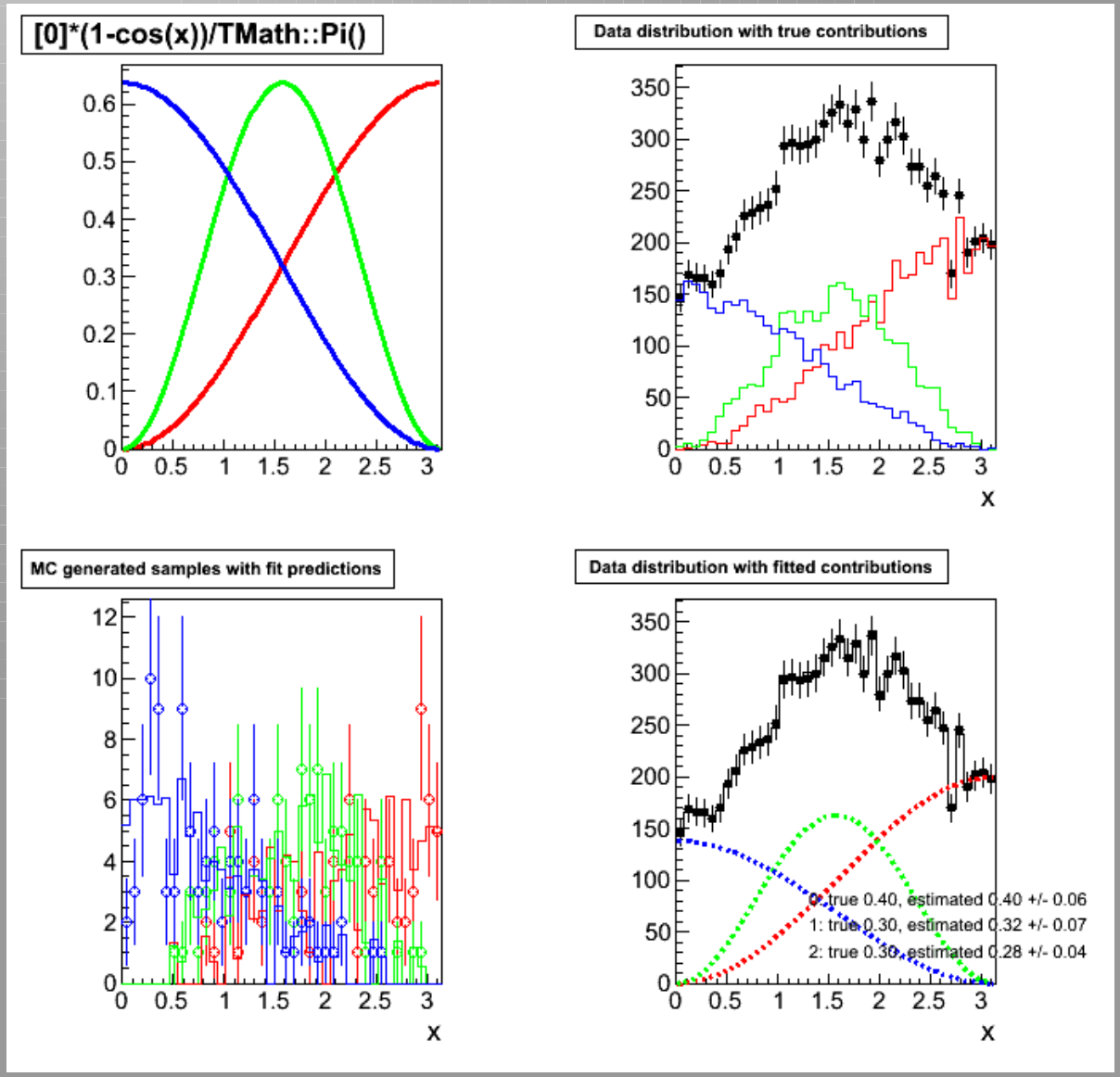
	true	estim	incert
p0	0.4	0.37	0.03
p1	0.3	0.32	0.04
p2	0.3	0.31	0.04

N(MC) = 10xN(data)



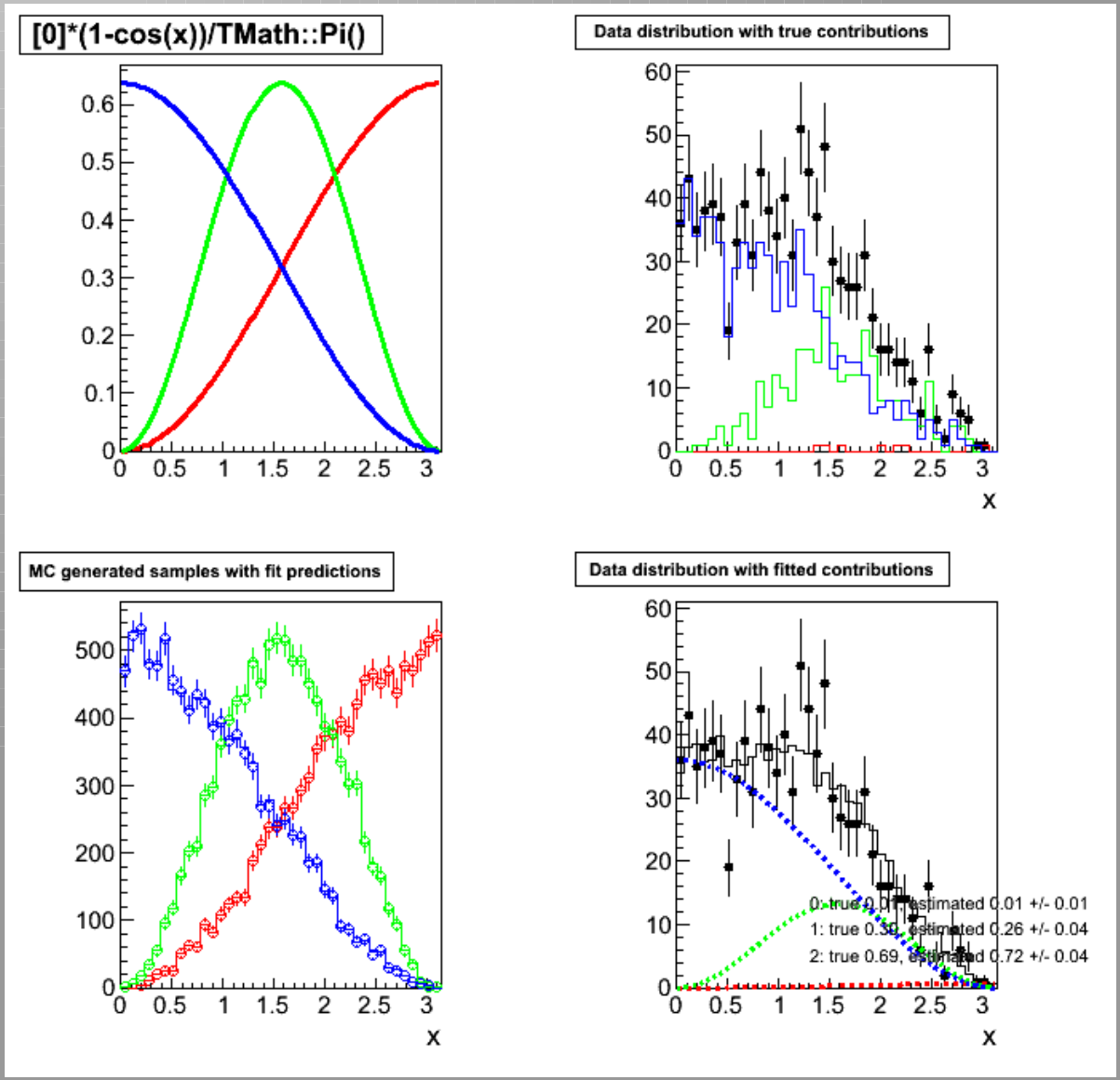
	true	estim	incert
p0	0.4	0.41	0.01
p1	0.3	0.28	0.02
p2	0.3	0.31	0.01

N(MC) = N(data)/10



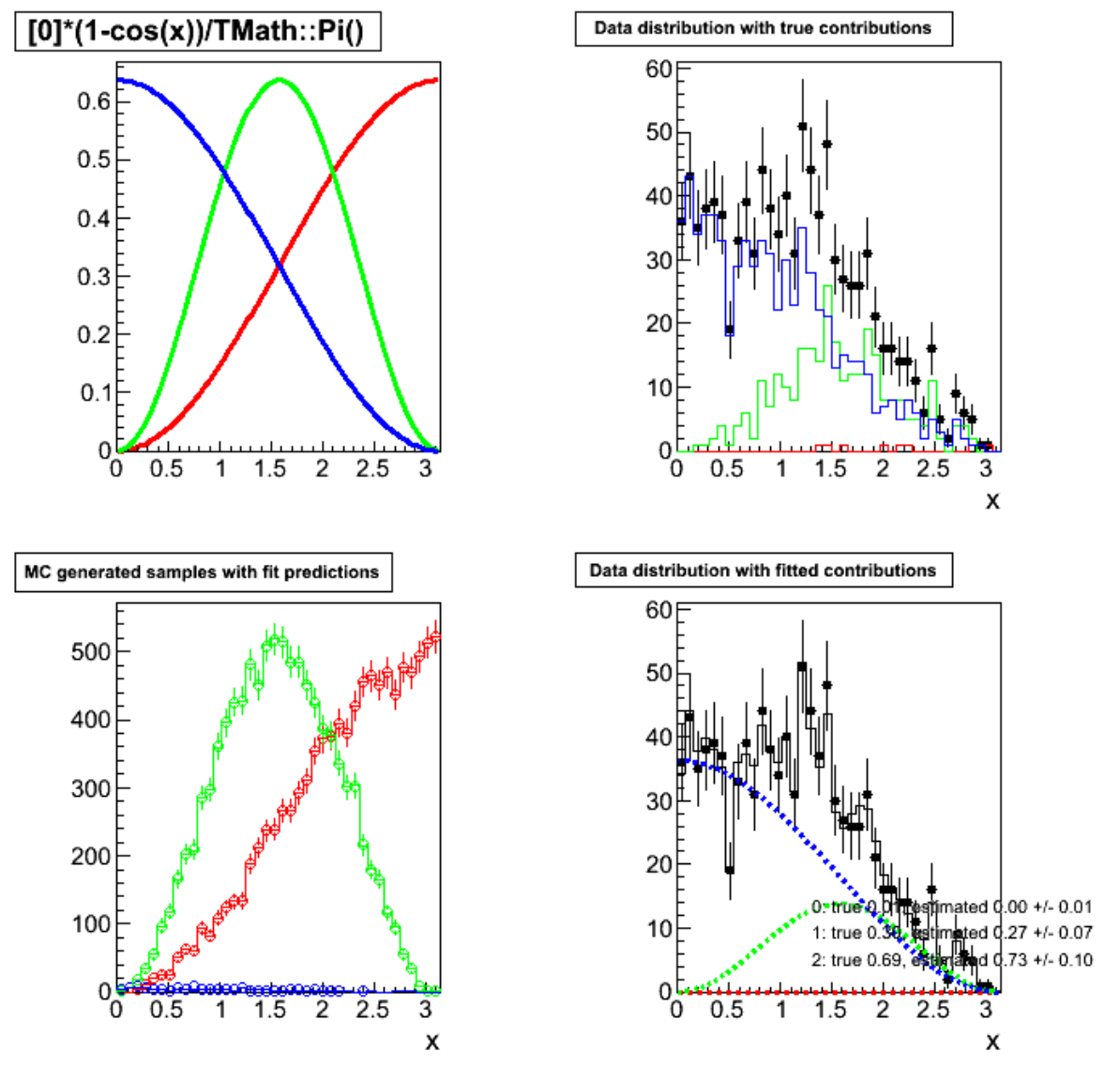
	true	estim	incert
p0	0.4	0.40	0.06
p1	0.3	0.32	0.07
p2	0.3	0.28	0.04

N(MC) = 10xN(data), low p_0



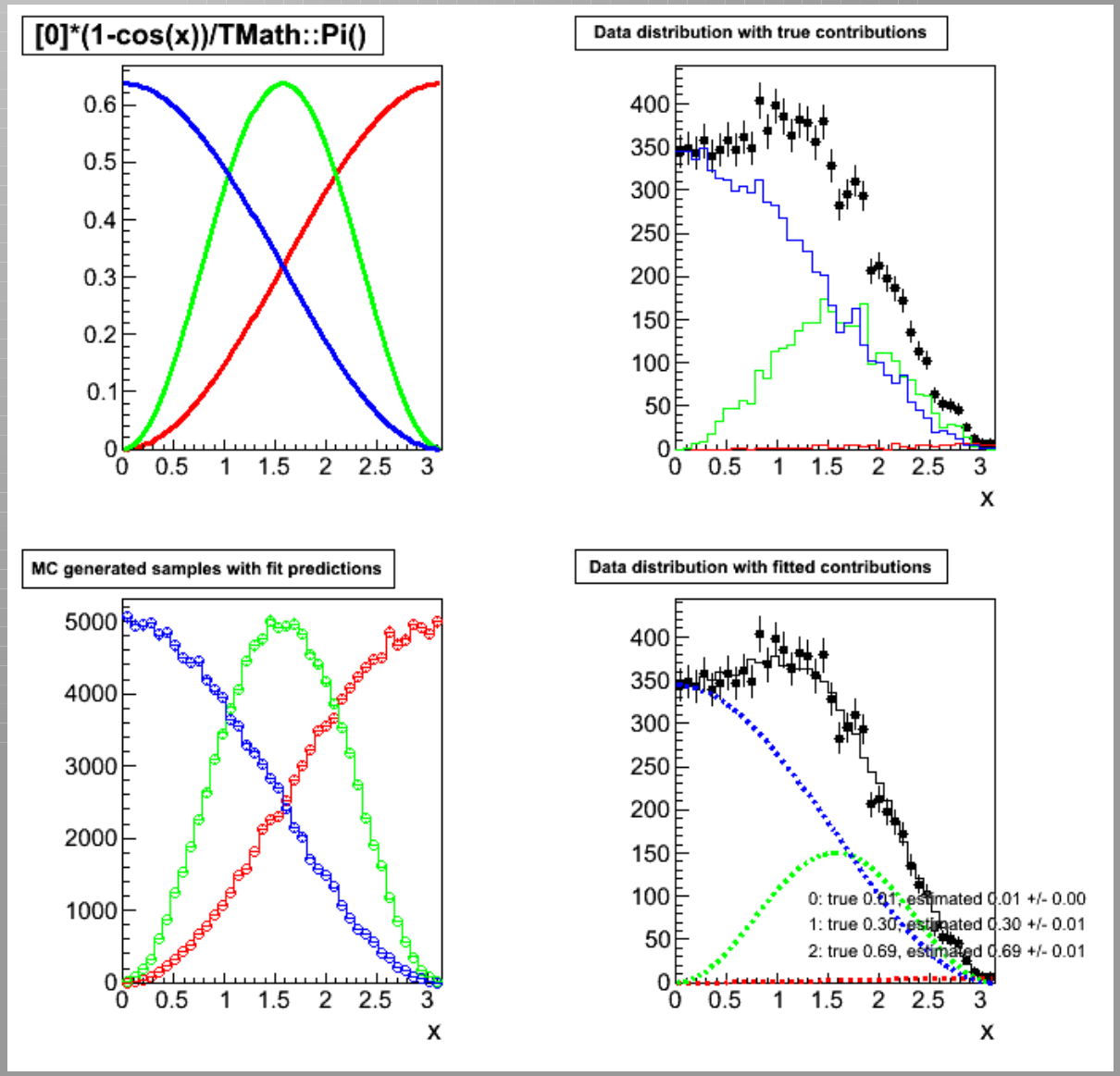
	true	estim	incert
p0	0.01	0.01	0.01
p1	0.3	0.26	0.04
p2	0.69	0.72	0.04

N(MC) = N(data)/10, low p_0



	true	estim	incert
p0	0.01	0.00	0.01
p1	0.3	0.27	0.07
p2	0.69	0.73	0.10

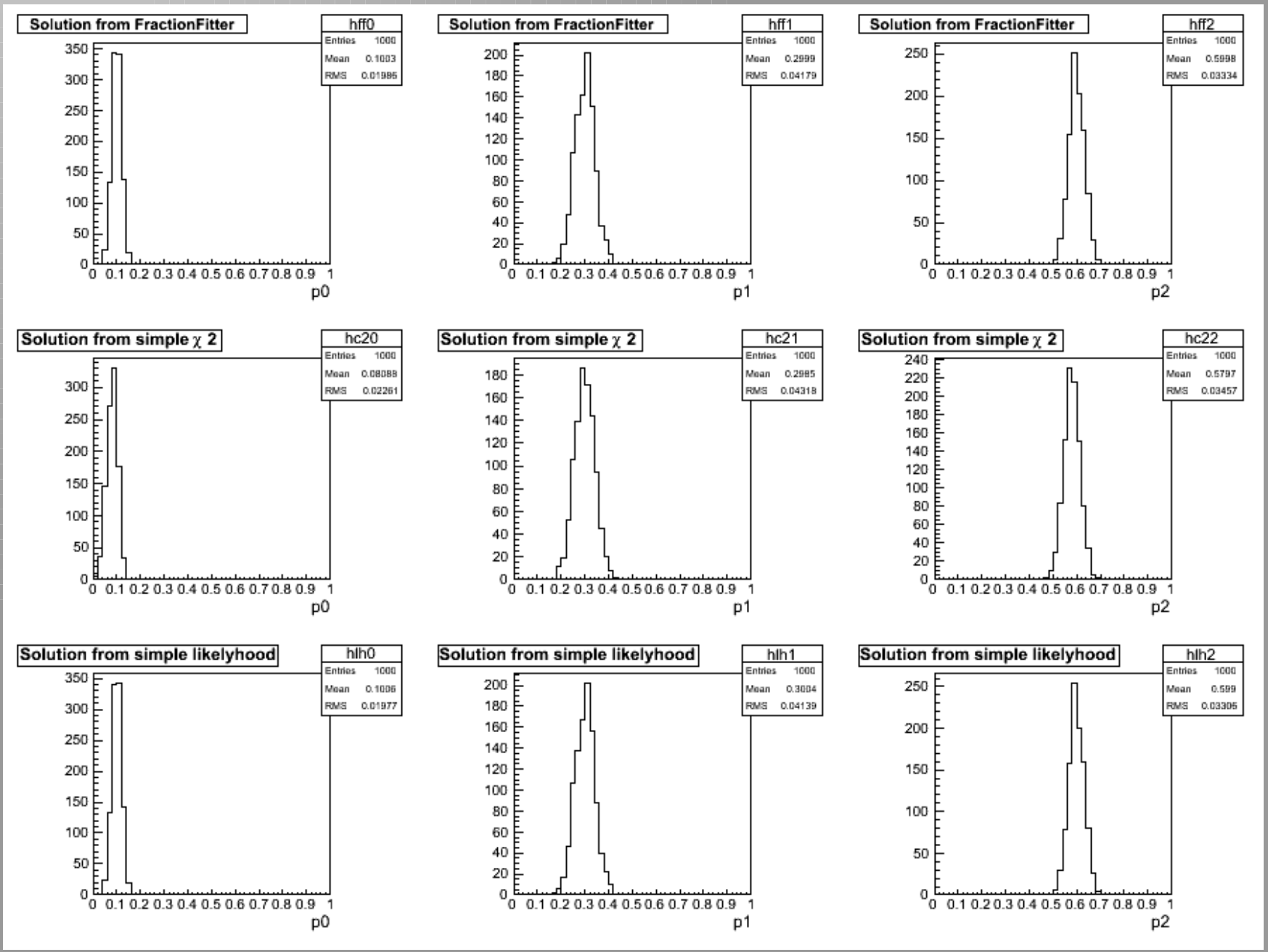
N(MC) = 10xN(data), low p_0



N(data) augmenté x10

	true	estim	incert
p0	0.01	0.01	0.00
p1	0.3	0.30	0.01
p2	0.69	0.69	0.01

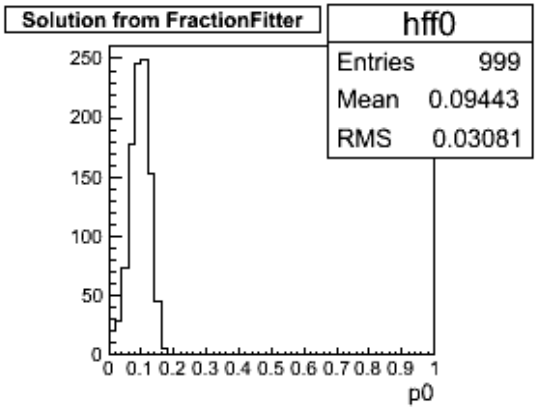
Comparons les incertitudes

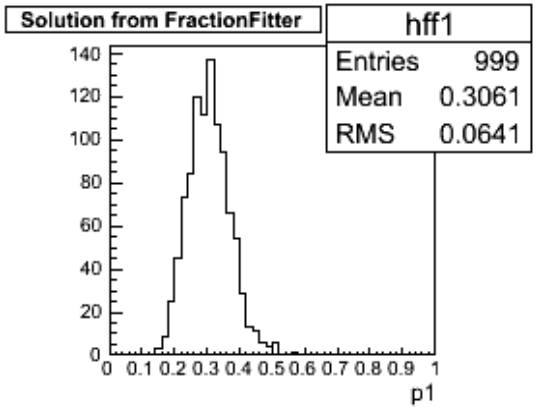


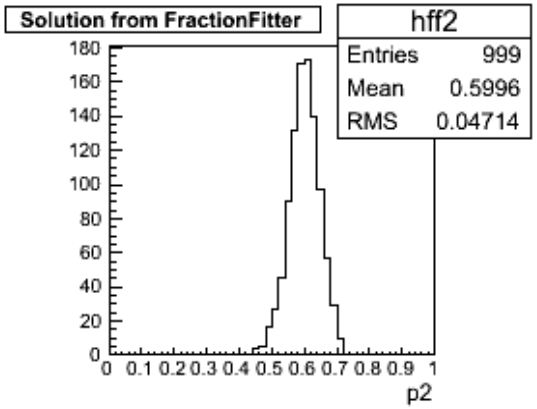
N(MC)
= N(data)*10

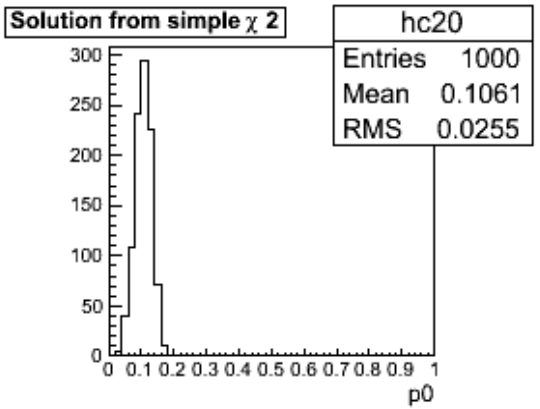
Macro fractionFitterComparison.C

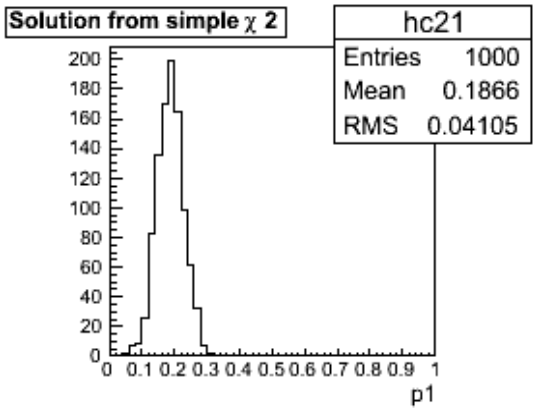
Comparons les incertitudes

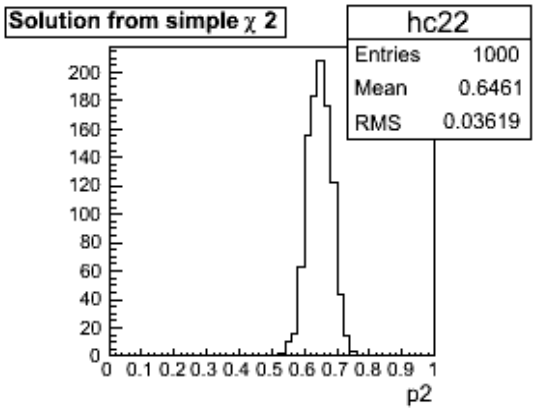


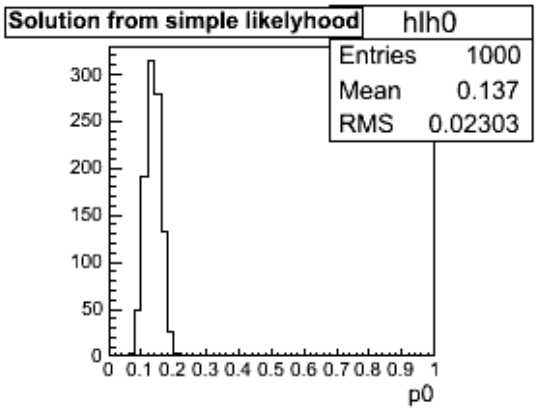


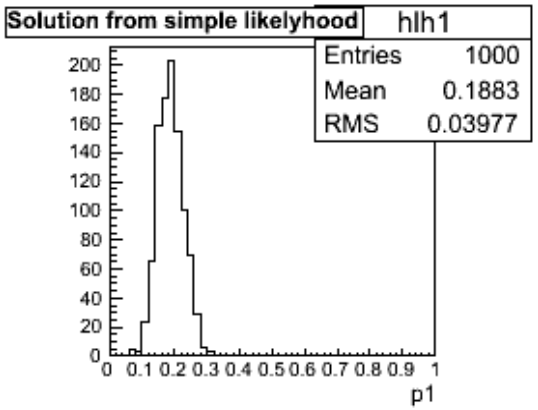


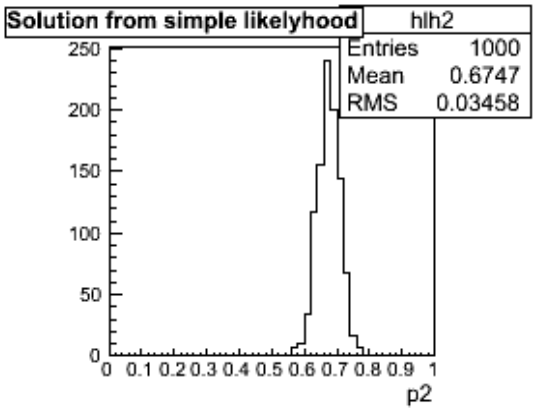












N(MC) = N(data)/100

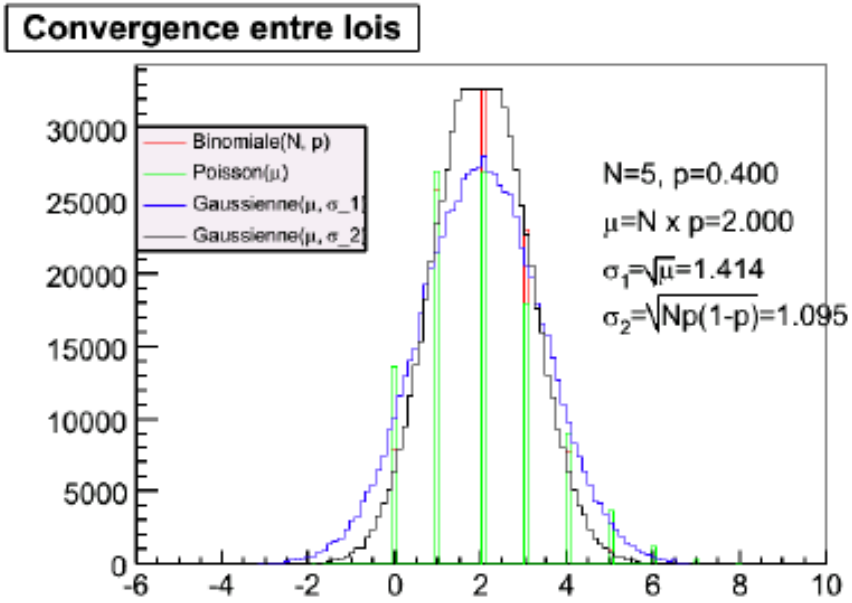
Conclusion

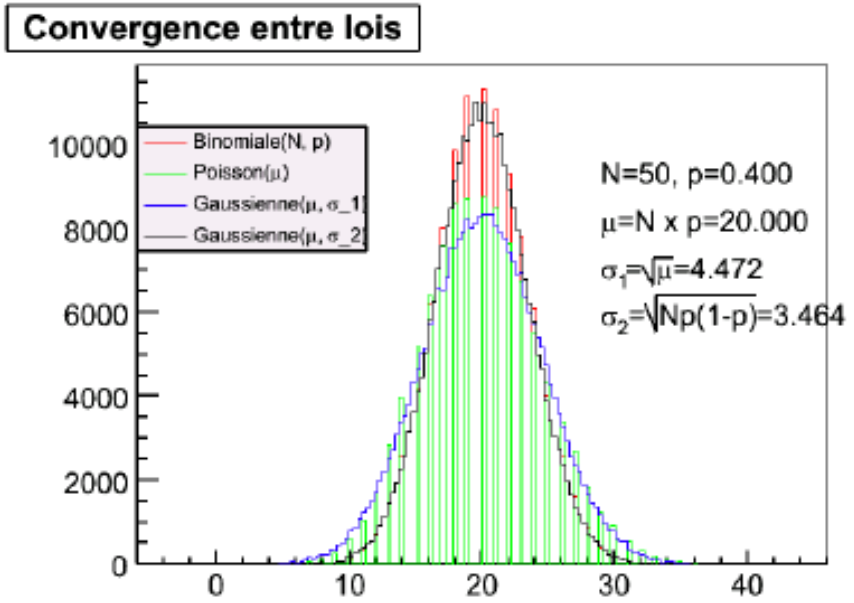
- Du point de vue « estimation »
  - ✗ Importance de la définition des fonctions à minimiser
    - ➔ Risque de biais !
- Du point de vue « pratique »
  - ✗ Vérification des incertitudes par Toy Monte Carlo

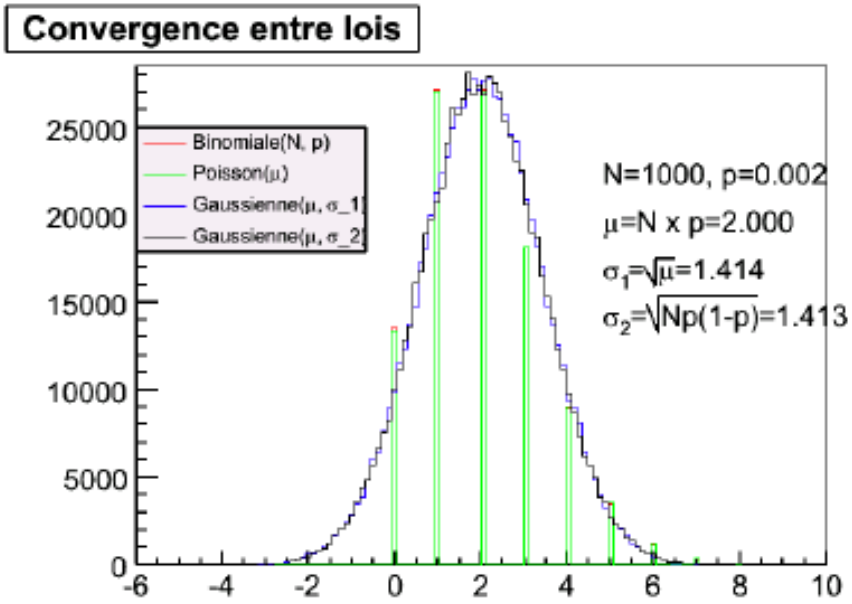
School Of Statistics

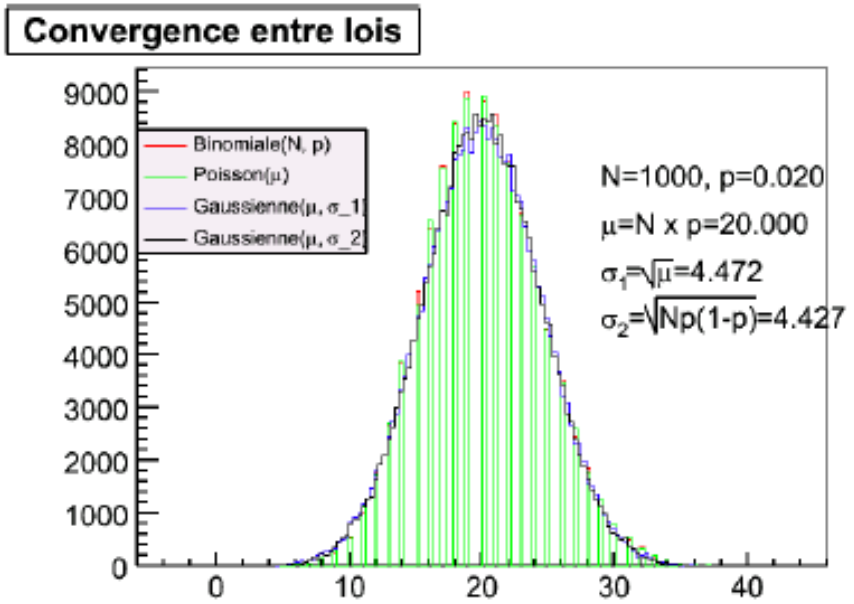
BACKUP

Convergence des lois

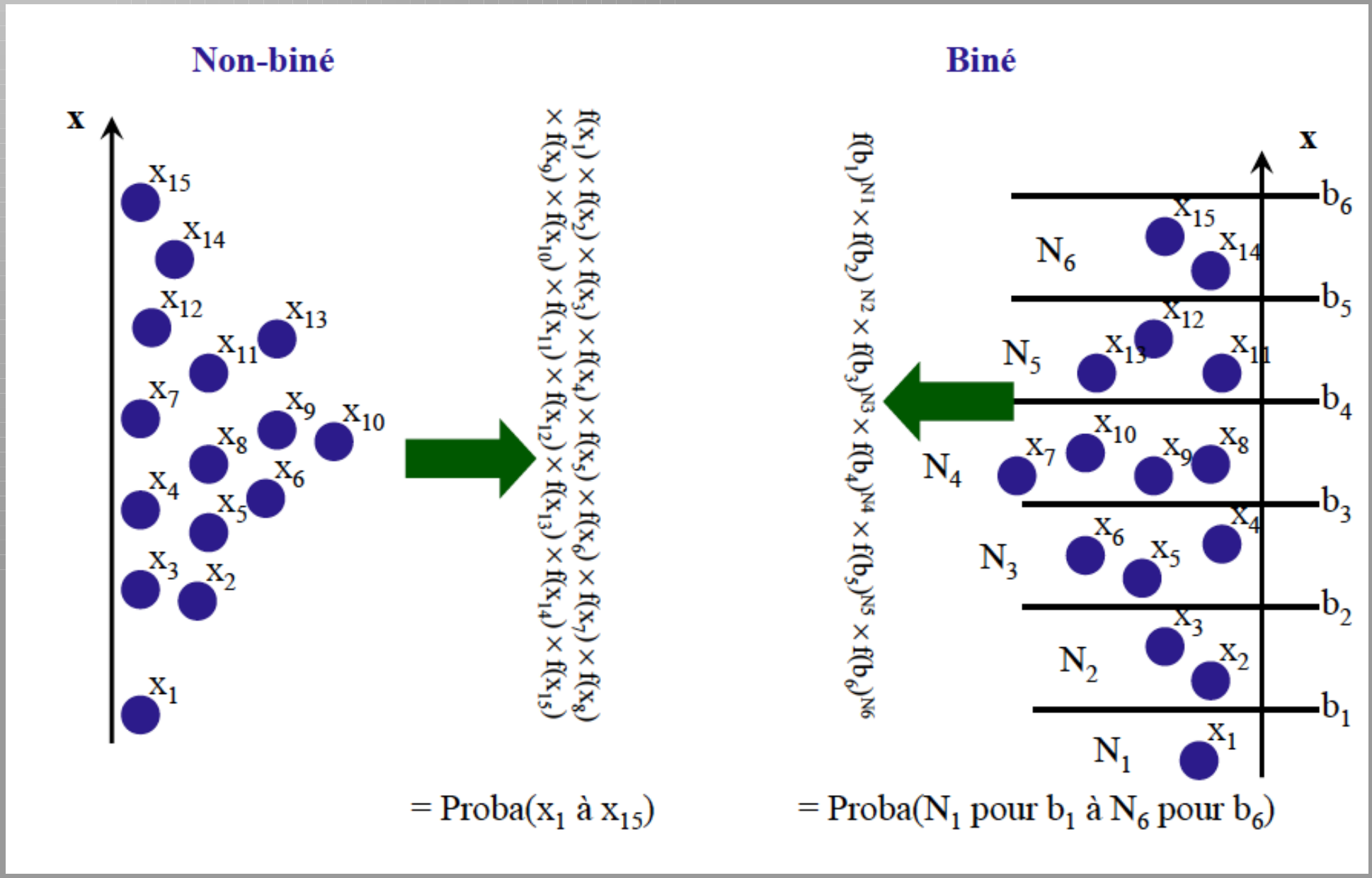








Vraisemblance binnée



Loi de distribution d'un histo

- Quelle est la loi de distribution des entrées par bin ?
  - Loi multinomiale = généralisation binomiale pour B alternatives
    - P la probabilité d'observer sur N tirages, chacune des B alternatives N_i fois, sachant que la probabilité de l'alternative i est p_i
  - Moments de la loi multinomiale
    - Moyenne de l'alternative i : Np_i
    - Variance de l'alternative i : $Np_i(1-p_i)$
    - Covariance des alternatives i et j : $-Np_ip_j \neq 0$!!
- Un histogramme = loi multinomiale
  - Un histogramme = B bins
  - Le bin i (i=1..B) contient N_i entrées avec Ni = N
  - Chaque bin est une alternative de la multinomiale où les pi dépendent la loi de distribution de la variable représentée
  - Si B est suffisamment grand les $p_i \ll 1$ et p_ip_j est négligeable devant p_i
  - Chaque bin est une var. de Poisson : moyenne = variance = Np_i