School Of Statistics

Basic Example - 1

Comment estimer une incertitude ?
Le cas de l'efficacité

Autrans, 17-21 mai 2010

D'après :

• Marc Patterno, “Calculating Efficiencies and Their Uncertainties”, FERMILAB-TM-2286-CD

• Thomas Ullrich & Zangbu Xu, “Treatment of Errors in Efficiency Calculations”, arXiv:physics/0701199v1

Institut Pluridisciplinaire Hubert CURIEN
STRASBOURG

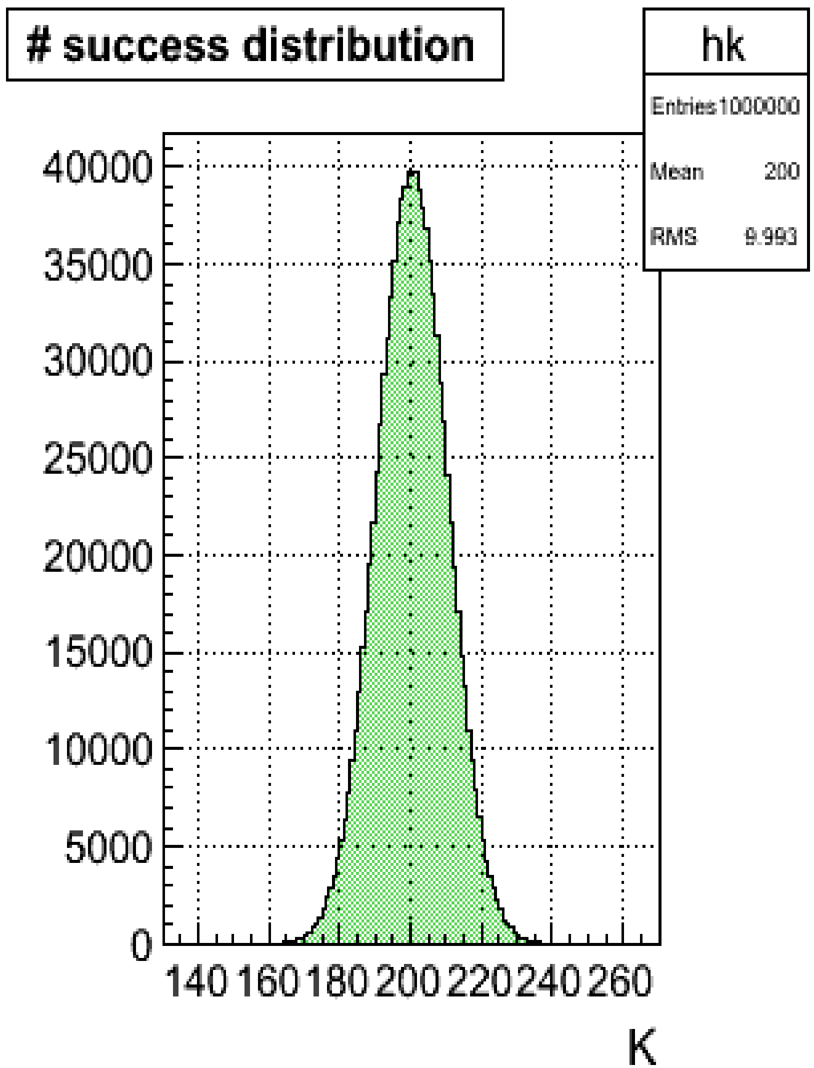
Le problème

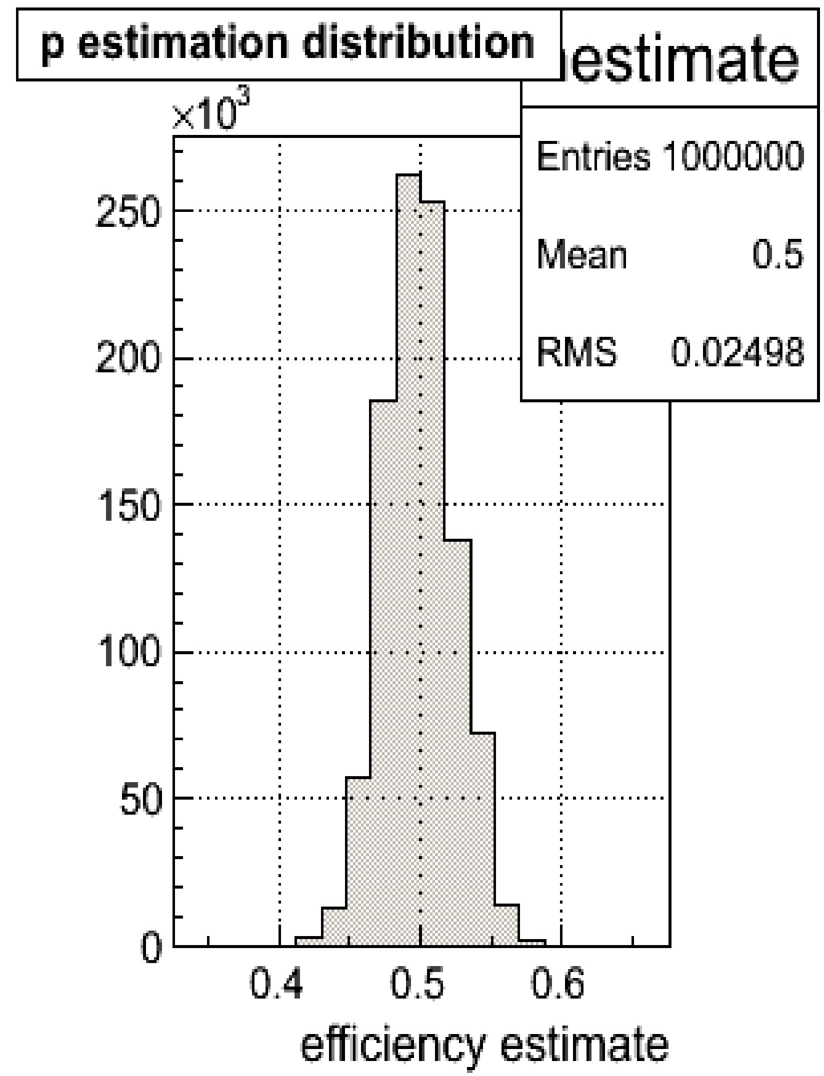
- Processus de sélection
  - ✗ Efficacité d'un détecteur, d'un critère de sélection, ...
  - ✗ p est la probabilité de sélectionner un événement
- Pour UNE expérience
  - ✗ Échantillon initial = N événements
  - ✗ Échantillon final = K événements sélectionnés
  - ✗ Quel est un estimateur de p ?
    - ➔ Un estimateur intuitif de p est K/N
  - ✗ Quelle est l'incertitude sur cette estimation ?
    - ➔ Selon la statistique de Poisson : $\sqrt{K} / N$

Approche empirique

- Let's go for a Toy MonteCarlo
  - ✗ Simulation de M expériences à p connu
  - ✗ Pour chaque expérience :
    - ➔ N tirages de la loi de Bernouilli (0=failed ou 1=success)
    - ➔ La sommation des 1 donne K
  - ✗ La distribution de K/N sur les M expériences fourni :
    - ➔ La moyenne de l'estimateur de p
    - ➔ L'écart-type de l'estimateur de p = incertitude recherchée ?
  - ✗ En pratique pour Bernouilli :
    - ➔ Tirer un nombre u uniforme [0,1]
    - ➔ Si u<p : success
    - ➔ Si u>p : failed

Toy MonteCarlo p=0.5





N=400

Succes

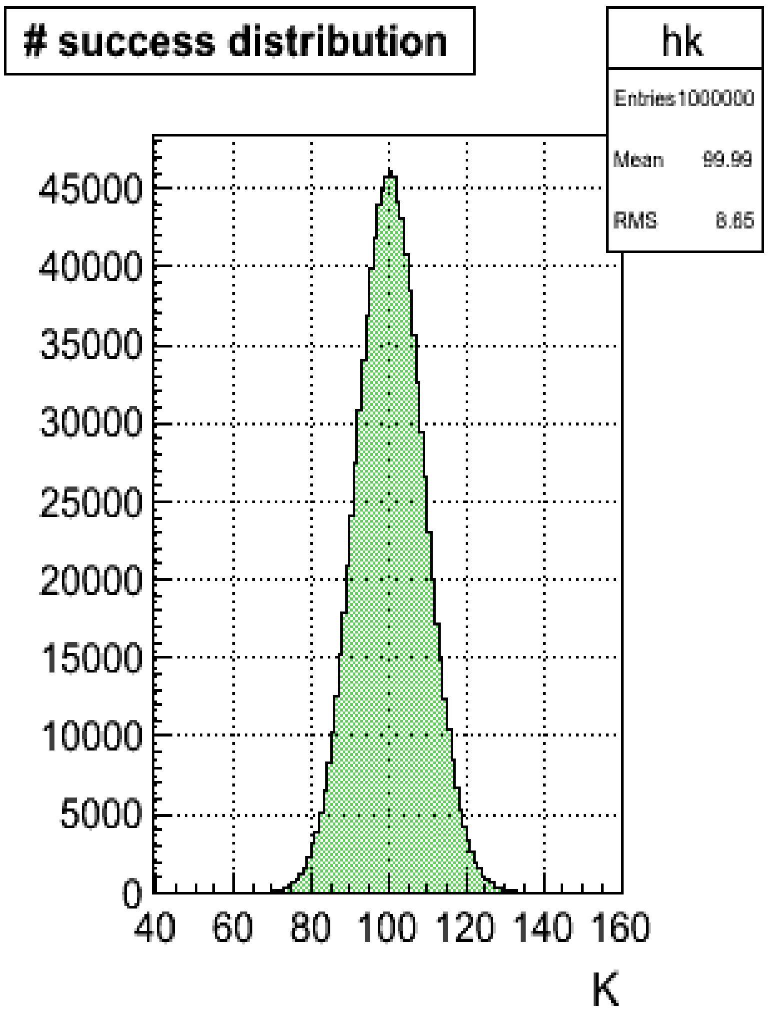
Averag

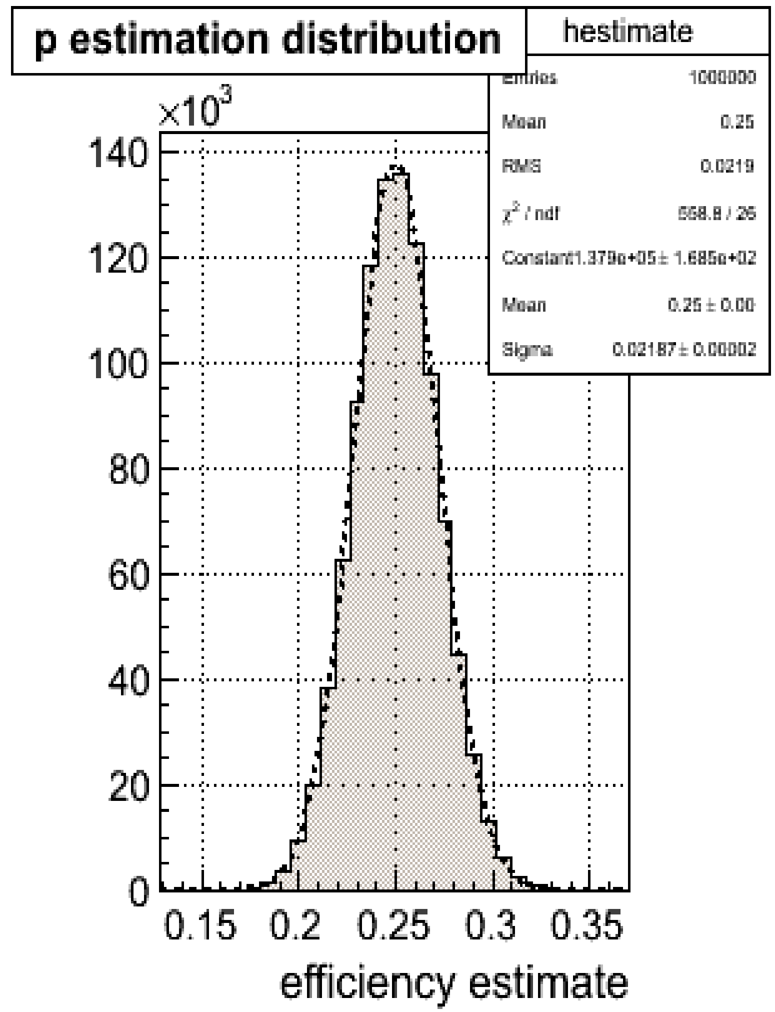
averag
stdev o

M=10^6 expériences de N=400 tirages

Macro uncertaintyOnEfficiency_simpleStudy.C

Toy MonteCarlo p=0.25





N=400 tria

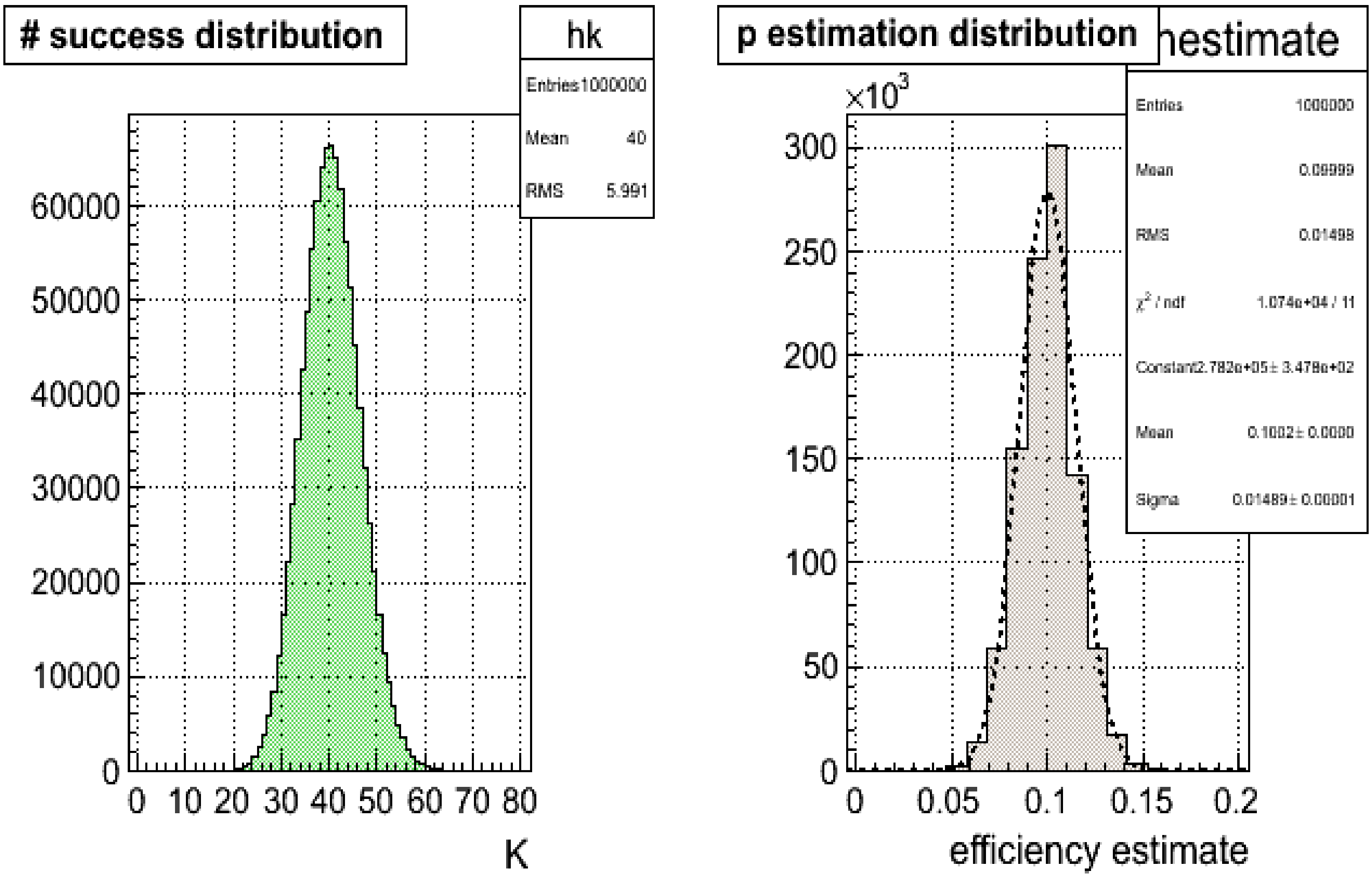
Success p

Average n

average p
stdev on p

M=10^6 expériences de N=400 tirages

Toy MonteCarlo p=0.1



N=400

Succes

Averag

averag
stdev o

M=10^6 expériences de N=400 tirages

Première leçon

- Incertitude = ecart-type de la distribution de l'estimateur
  - K est une variable aléatoire binomiale
    - Moyenne Np
    - Variance Np(1-p)
  - Estimateur de p = K/N → variance de p = variance(K)/N^2

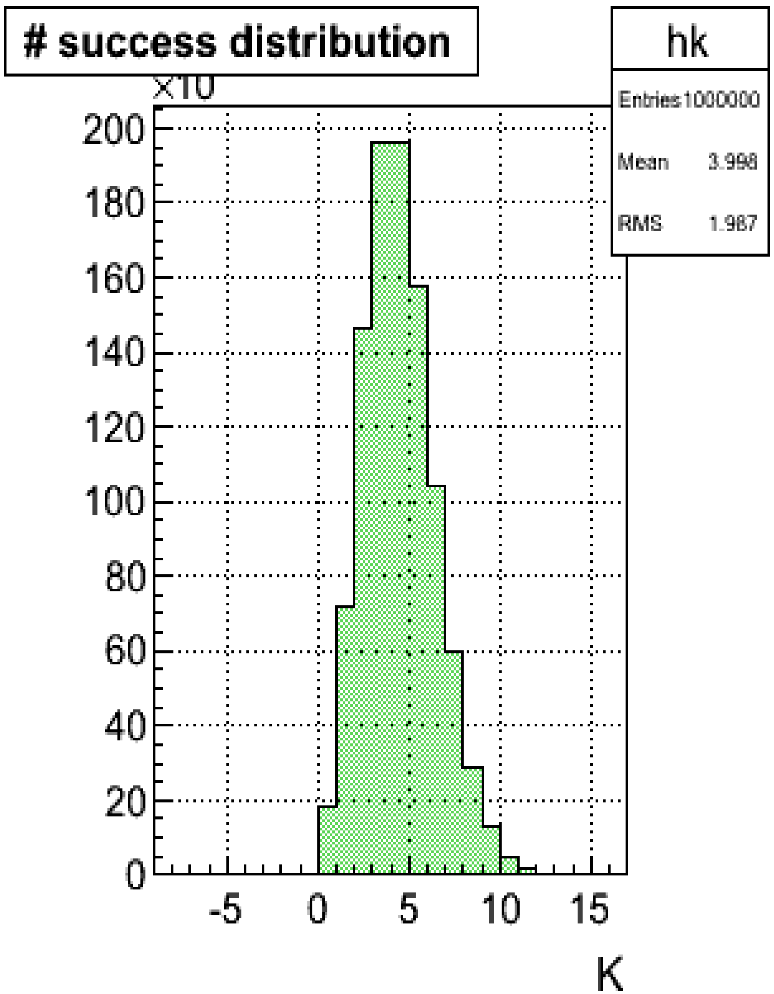
$$\sigma_p = \sqrt{\frac{p(1-p)}{N}}$$

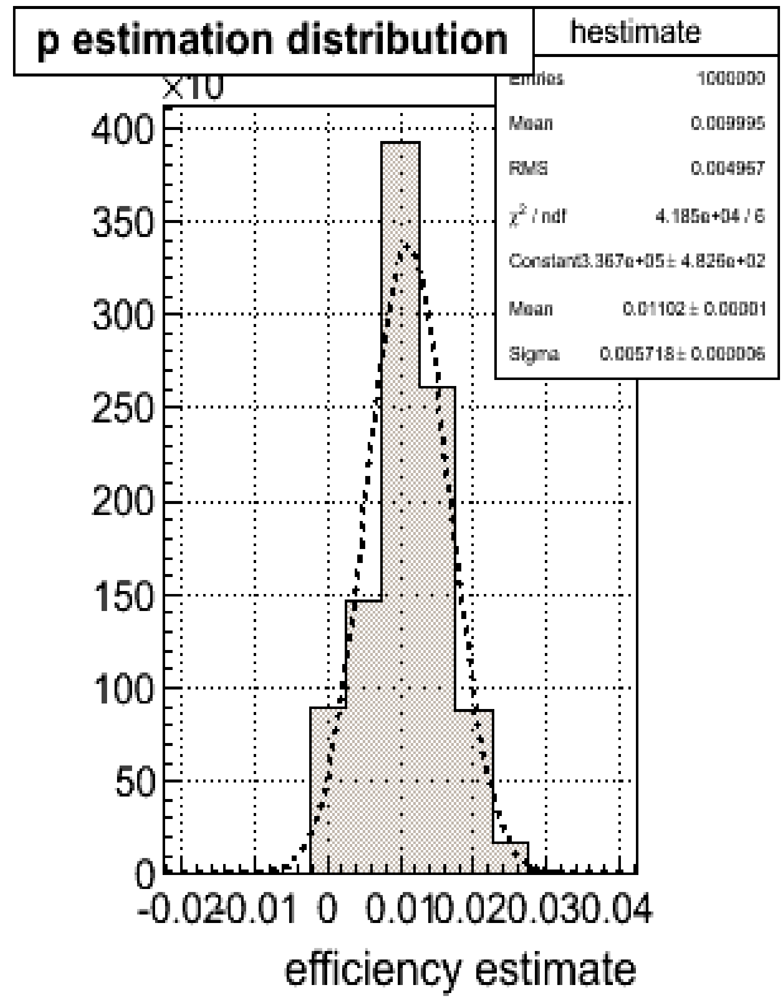
  - Sur les exemples précédents

p	RMS histo	(p)
0.5	0.025	0.025
0.25	0.022	0.022
0.1	0.015	0.015

So far, so good...

Toy MonteCarlo p=0.01





N=400 tri

Success

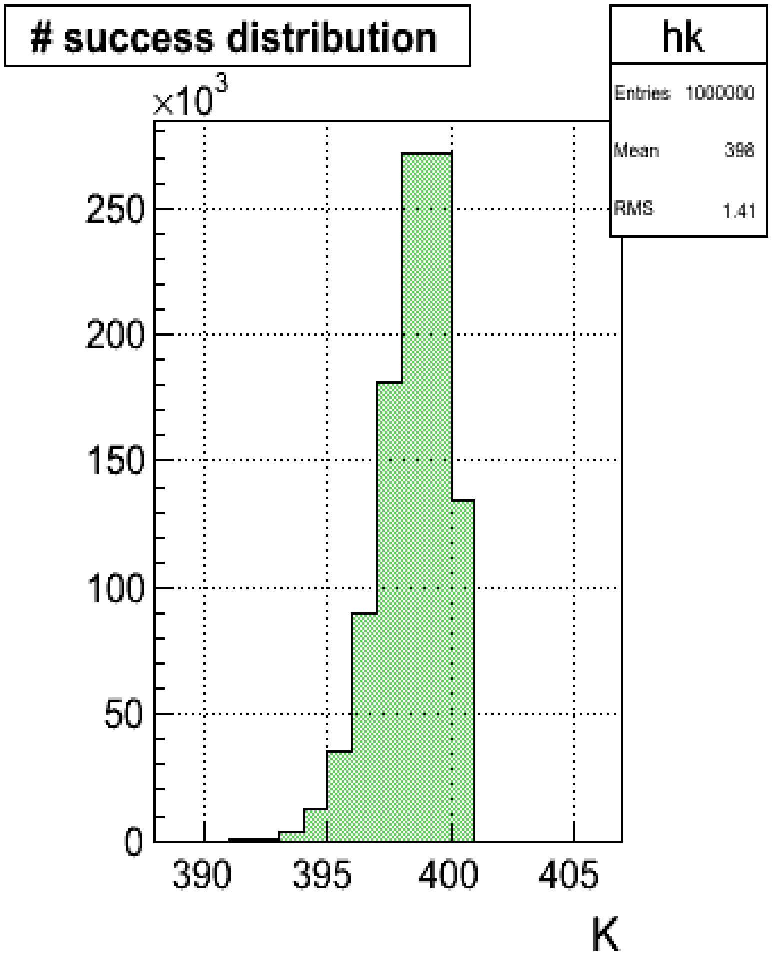
Average

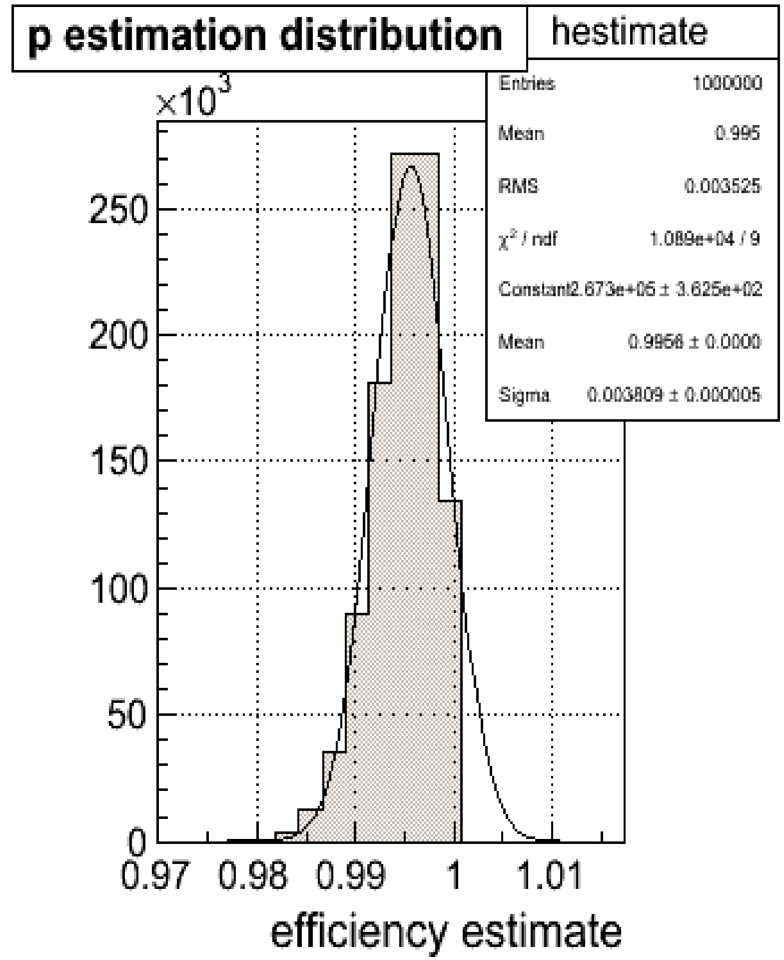
average |
stdev on

M=10^6 expériences de N=400 tirages

Moyenne ≠ 0.005
Variance prévue 0.005 !

Toy MonteCarlo p=0.995





N=400 tria

Success p

Average n

average p
stdev on p

- La distribution devient asymétrique
  - ✗ La moyenne n'est plus le mode (valeur la plus probable)
  - ✗ L'erreur est ridiculement faible (prévue 0.0035)

Situation difficile

- Cas p~0 ou p~1
  - Il est possible d'observer K=0 ou N sans que p=0 ou 1 strictement
  - Formulation avec écart-type de la binomiale prédit dans ces cas :
    - ATTENTION, on utilise l'estimateur de p=K/N dans la formule
    - Incertitude (1-p)p = 0 ???

- Revenons à la loi Binomiale
  - Pour p=0 (ou 1), la seule variable possible de K est 0 (ou N)
  - Il est donc normal que la variance soir nulle

... On n'a rien compris !

La bonne question

- **La réponse que nous avons :**
  - ✗ Loi binomiale = probabilité que pour p fixé, on observe K (à N fixé)
    - ➔ Écart-type ~ intervalle dans lequel se situe 68% des K (pour N grand)

- **La bonne question :**
  - ✗ Pour K observé fixé, quelle est la loi de probabilité de p ?
    - ➔ Écart-type ~ intervalle dans lequel se situent 68% des p

- **L'hypothèse implicite que nous avons faite :**
  - ✗ Les deux questions précédentes sont équivalentes
    - ➔ Identification de
la loi de distribution de l'estimateur de p avec
la loi de distribution de la variable K
  - ✗ Aux limites p=0, ou 1 cette hypothèse est mise en défaut !

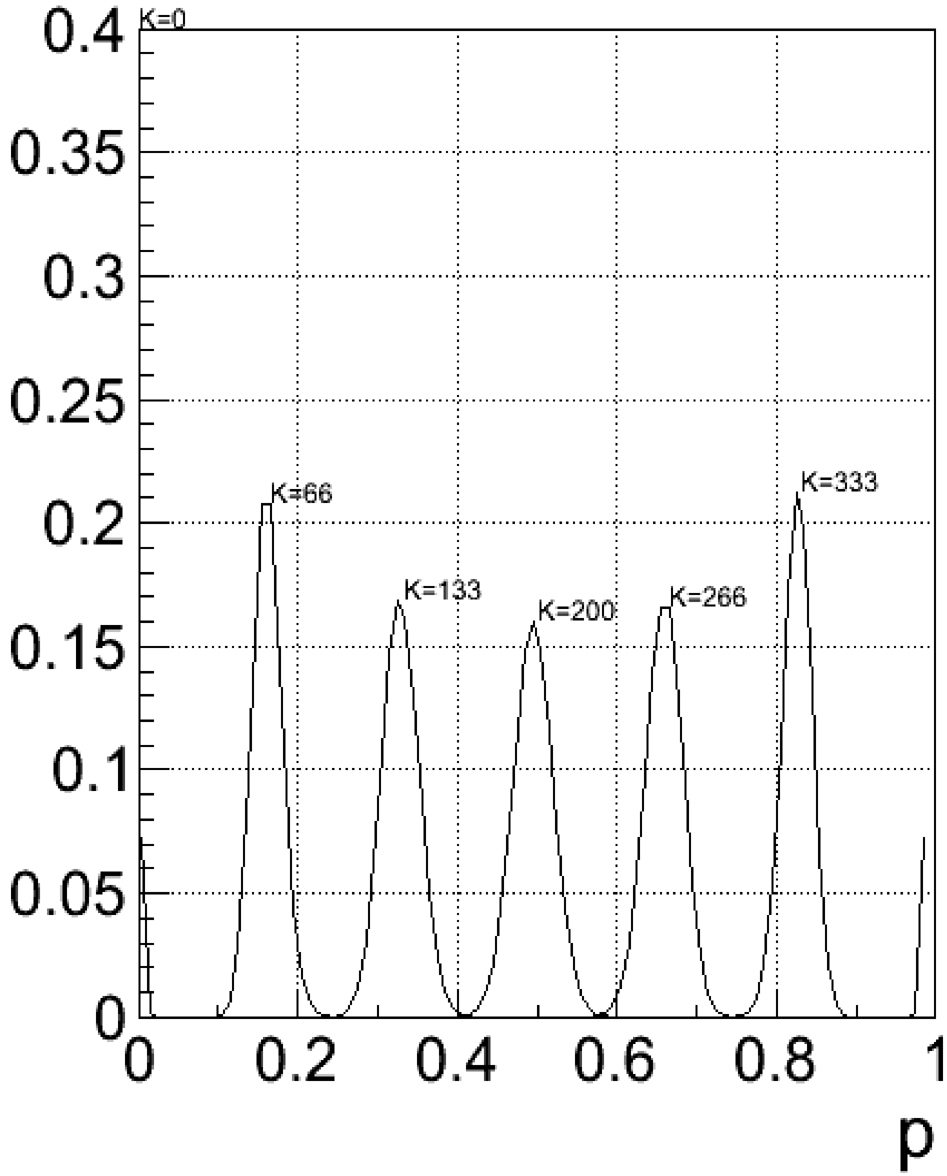
Back to Bayes

- La solution = probabilité conditionnelle
  - Probabilité de p sachant que K est observé = P(p|K)

$$P(p|K)=\frac{P(K|p)*P(p)}{Z}$$

  - P(K|p) est la probabilité binomiale, connue
  - P(p) est une distribution *a priori*, le fameux prior des Bayesiens !
    - Choix "naturel" P(p) uniforme=1 dans [0,1], nulle ailleurs
  - Z est une constante de normalisation (0 < P(p|K) < 1)
    - Vérifie $$1=\int P(p|K)=\frac{1}{Z}\frac{N!}{K!K-N!}\int p^K(1-p)^{N-K}dp$$
    - On obtient $$\boxed{P(p|K)=\frac{N+1!}{K!K-N!}p^K(1-p)^{N-K}}$$

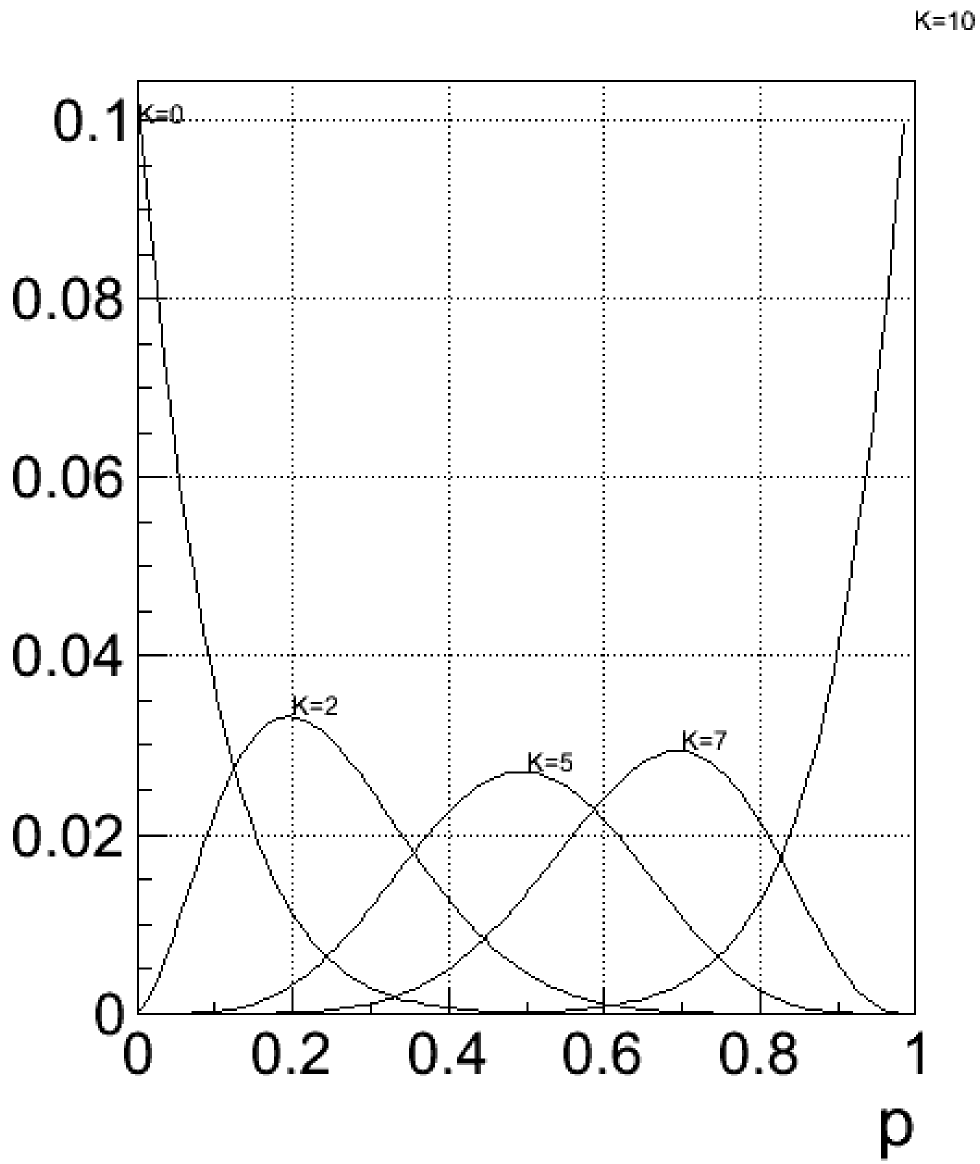
La distribution P(p|K)



- Pour N(=400) grand, l'asymétrie de la distribution reste peut visible sauf à très petit (grand) K

Macros :
- distributionEfficiencyFromOccurence.C
- distributionOfEfficiency.C

La distribution P(p|K)



- Pour N(=10) faible, l'asymétrie de la distribution apparaît rapidement
- Le mode vaut K/N
- La moyenne vaut (K+1)/(N+2)   mode !
- La probabilité est nulle en p=0 ou 1 sauf si K=0 ou N

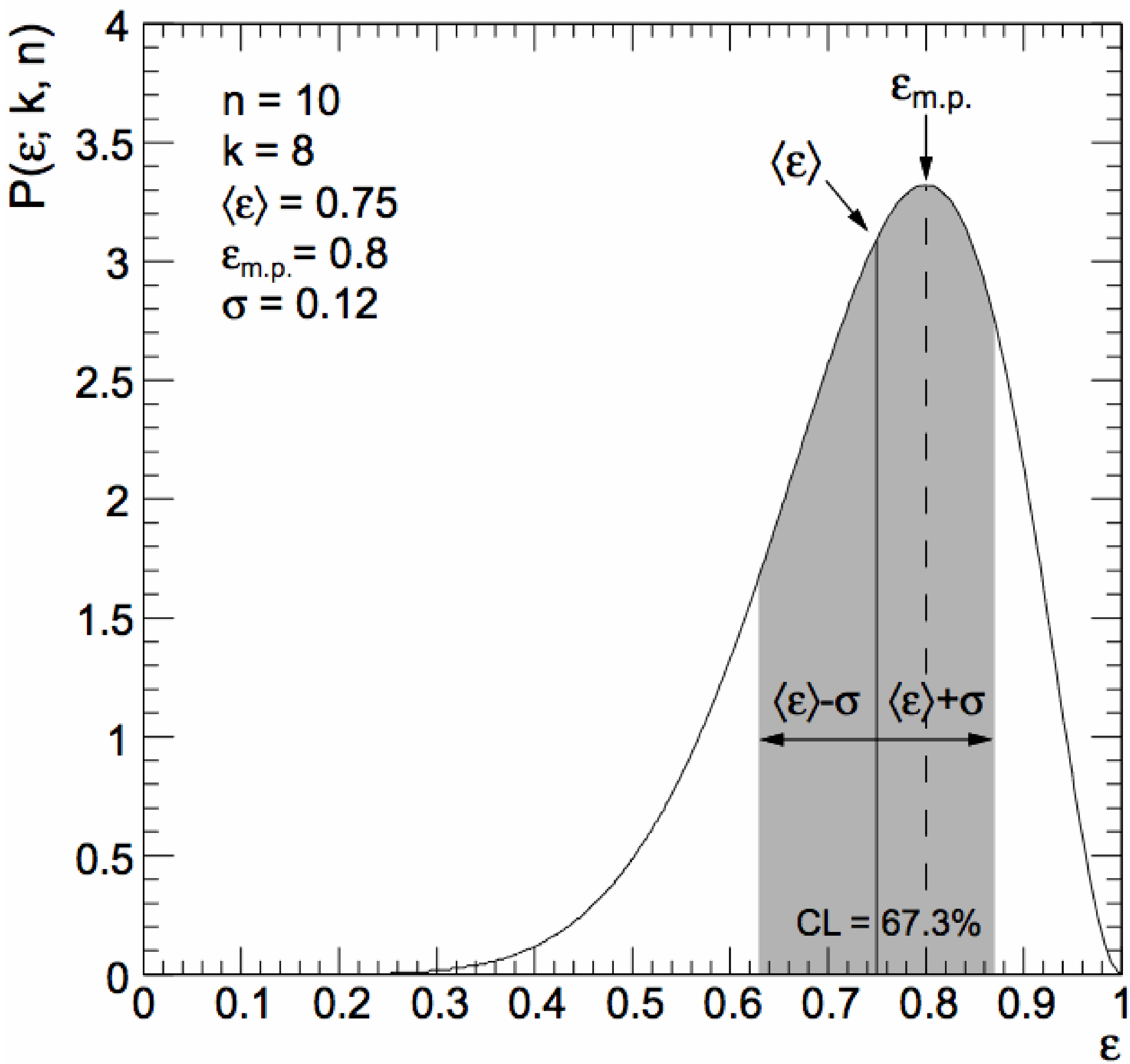
Estimation de l'incertitude

- Méthode 1 : prendre l'écart-type de P(p|K)
  - Calul donne :

$$\boxed{\sigma_p = \sqrt{\frac{(K+1)(K+2)}{(N+2)(N+3)} - \frac{(K+1)^2}{(N+2)^2}}}$$

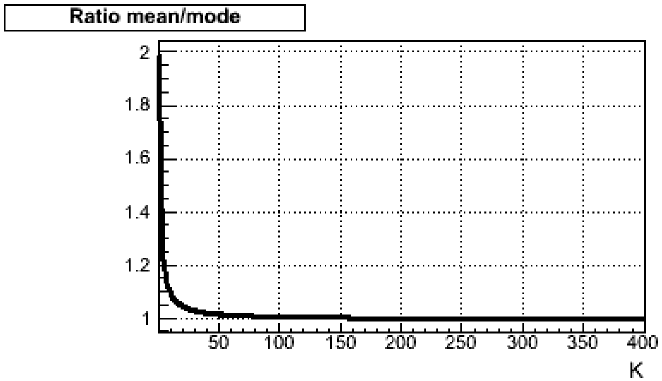
- Méthode 2 : intervalle avec couverture de 68%
  - Méthode numérique proposée par M.Paterno
  - Voir la classe TgraphAsymmError dans Root

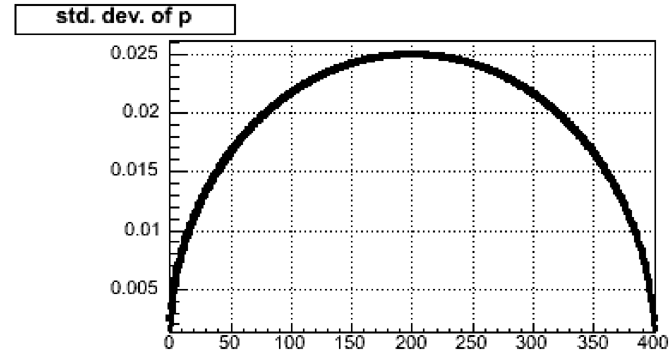
Estimation de l'incertitude

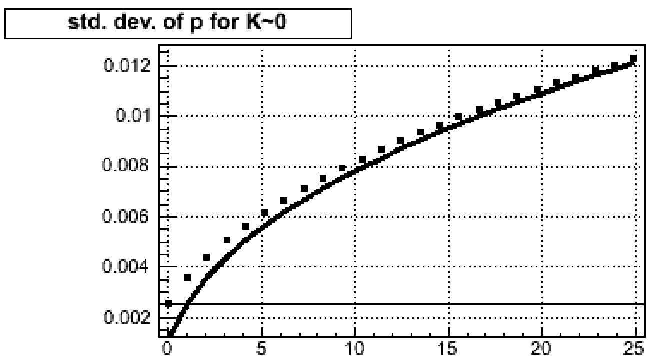


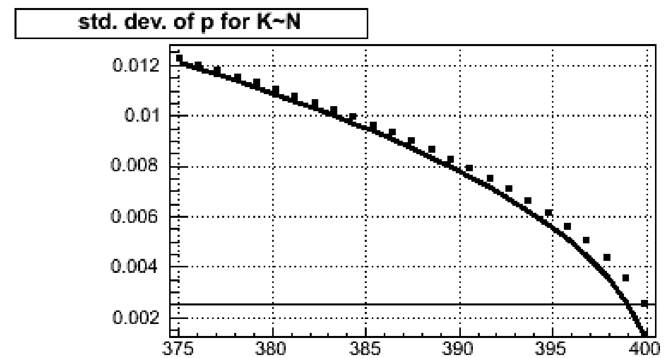
Comparons les incertitudes

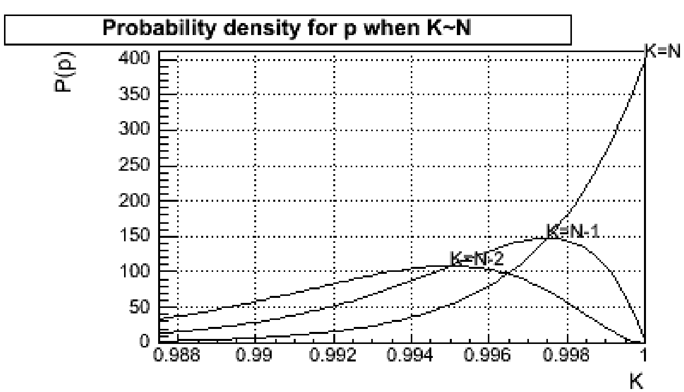
Macro uncertaintyOnEfficiency_study.C











N=400 trials

K is the number of success

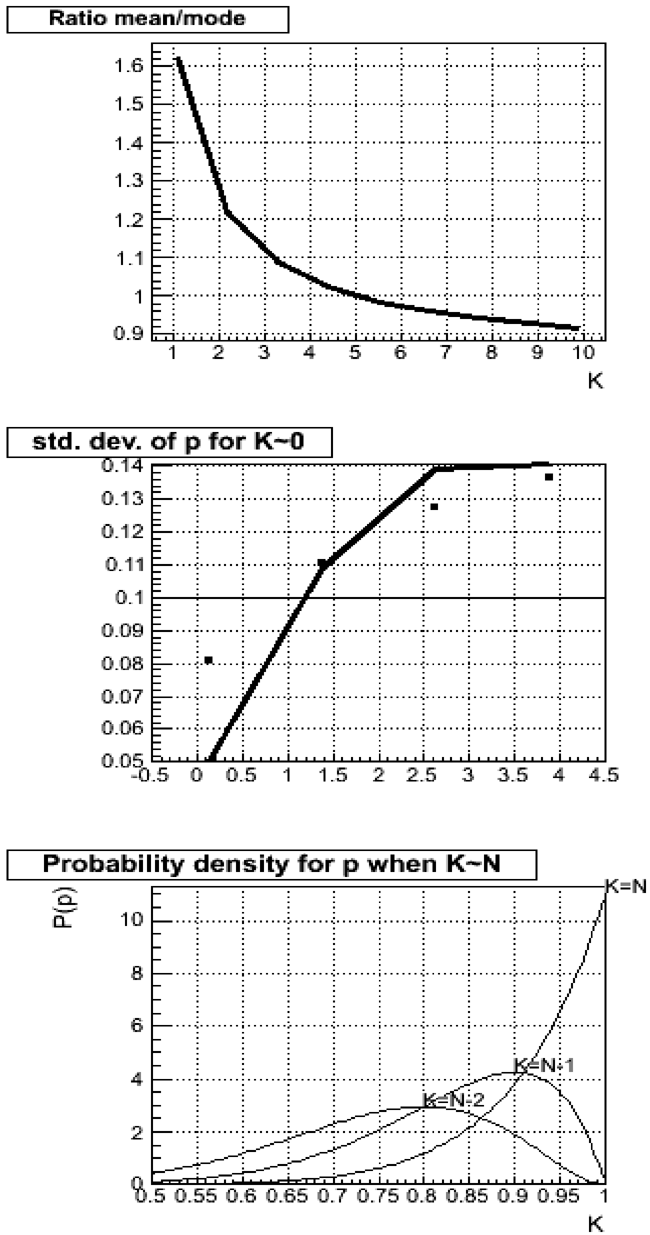
p is the probability for a success

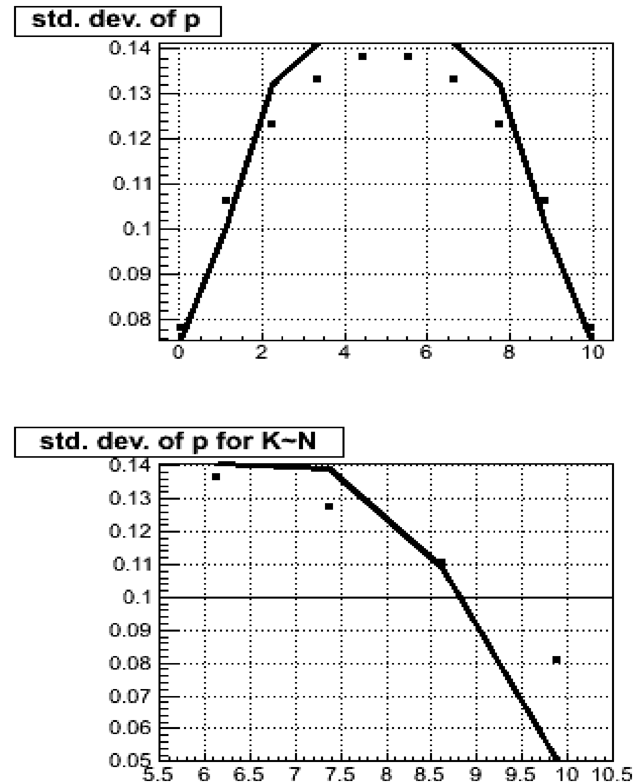
- full line stands for the estimation:
  p = K/N
  uncertainty on epsilon = sqrt(p(1-p)/N)

- Dots stand for the exact p probability distribution:
  <p> = (K+1)/(N+2)
  sqrt(<p^{2}>) = sqrt((K+1)(N-K+1)/(N+3)(N+2)^{2})

- when K~N or 0,
  then #sigma_{p} = 1/N = Poisson statistics (line)

Comparons les incertitudes





N=10 trials

K is the number of success

p is the probability for a success

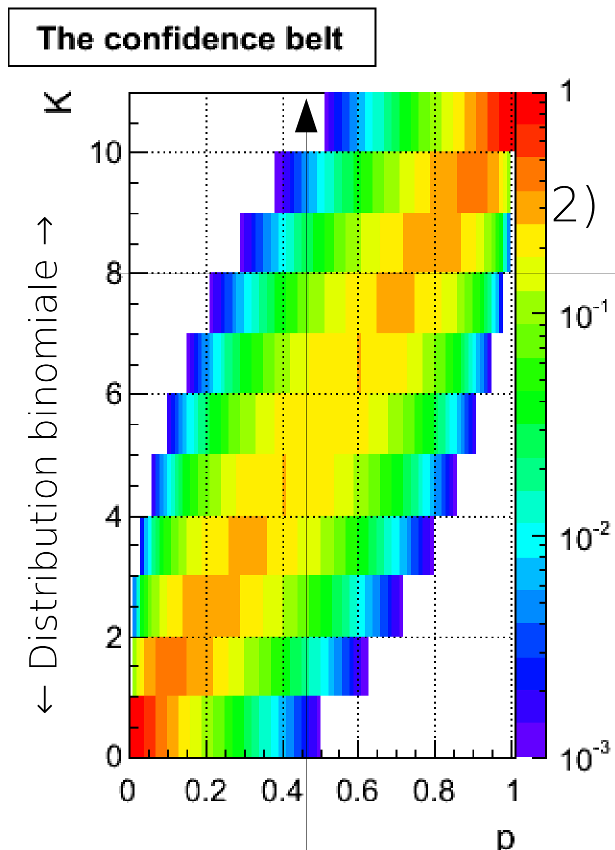
- full line stands for the estimation:
  p = K/N
  uncertainty on epsilon = sqrt(p(1-p)/N)

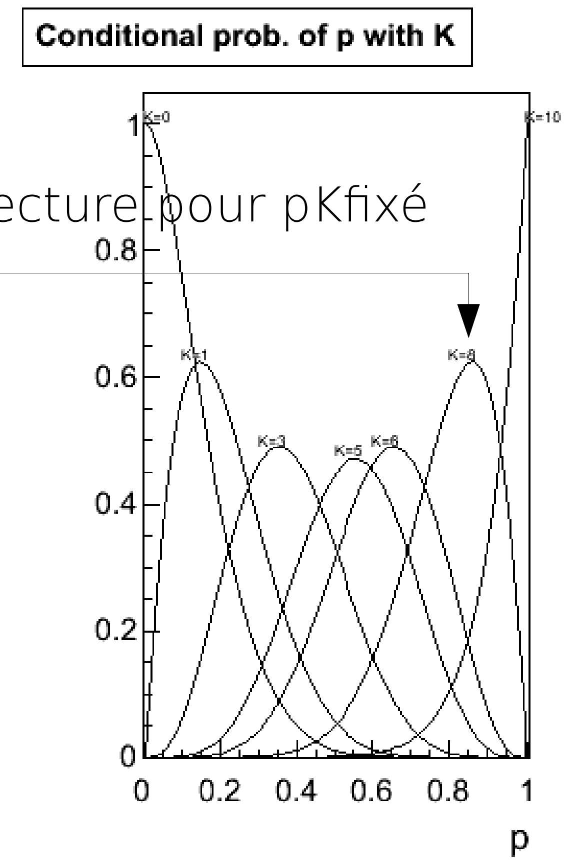
- Dots stand for the exact p probability distribution:
  <p> = (K+1)/(N+2)
  sqrt(<p^{2}>) = sqrt((K+1)(N-K+1)/(N+3)(N+2)^{2})

- when K~N or 0,
  then #sigma_{p} = 1/N = Poisson statistics (line)

Méthode numérique pour P(p|K)

- **Inversion graphique**
  - De P(K|p) vers P(p|K)
  - Macro uncertaintyOnEfficiency_Inversion.C



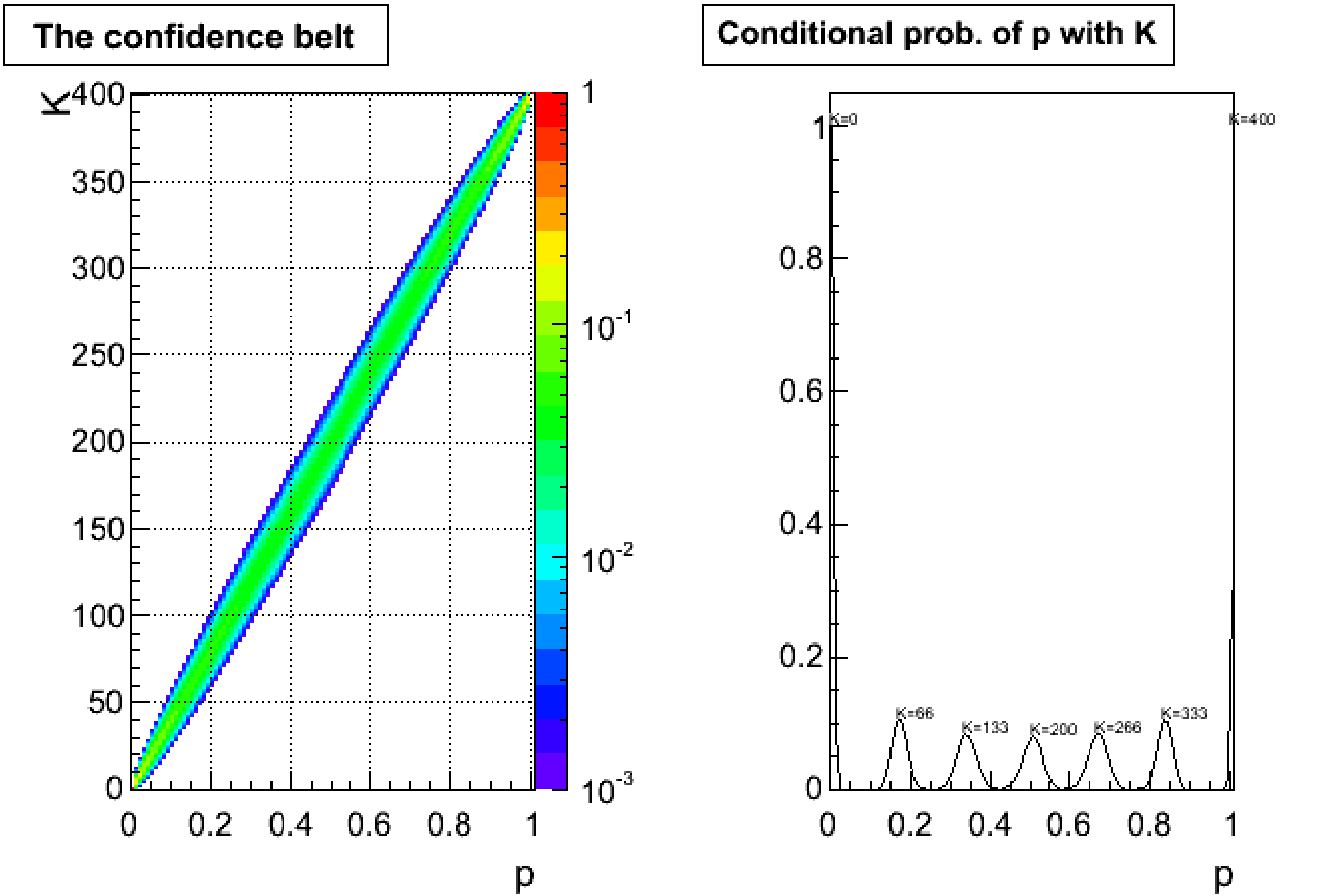


2) Lecture pour pKfixé

1) Génération à p fixé

Méthode numérique pour P(p|K)

N=400



School Of Statistics

Conclusion

- Sur le problème de l'incertitude associée à l'estimation de l'efficacité :
  - ✗ Méthode de la variance binomiale correcte dans la plupart des cas
  - ✗ Pour les cas limites (p~0, 1) faire attention !

- En général
  - ✗ Le point crucial consiste à se poser la bonne question