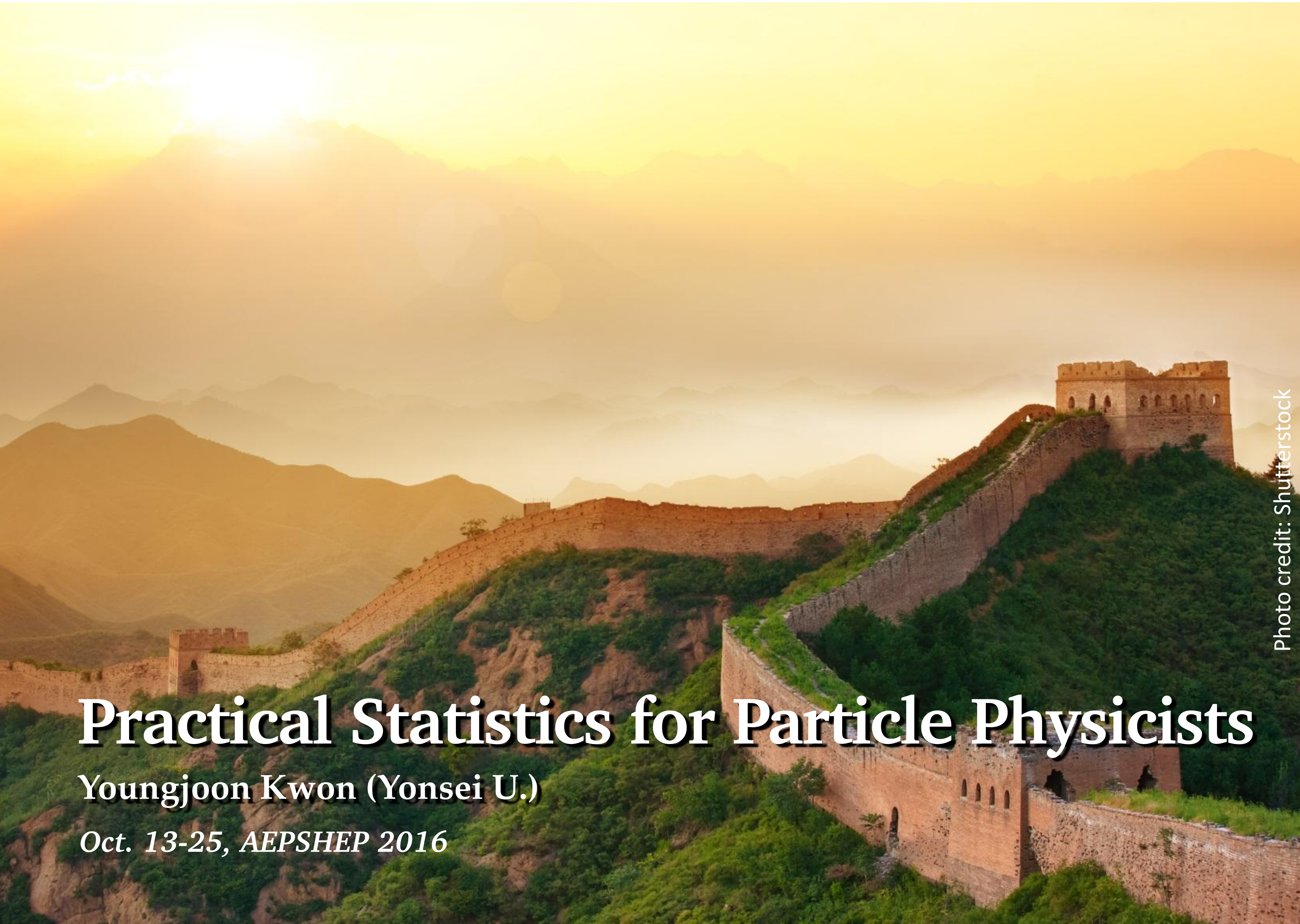


Practical Statistics for Particle Physicists

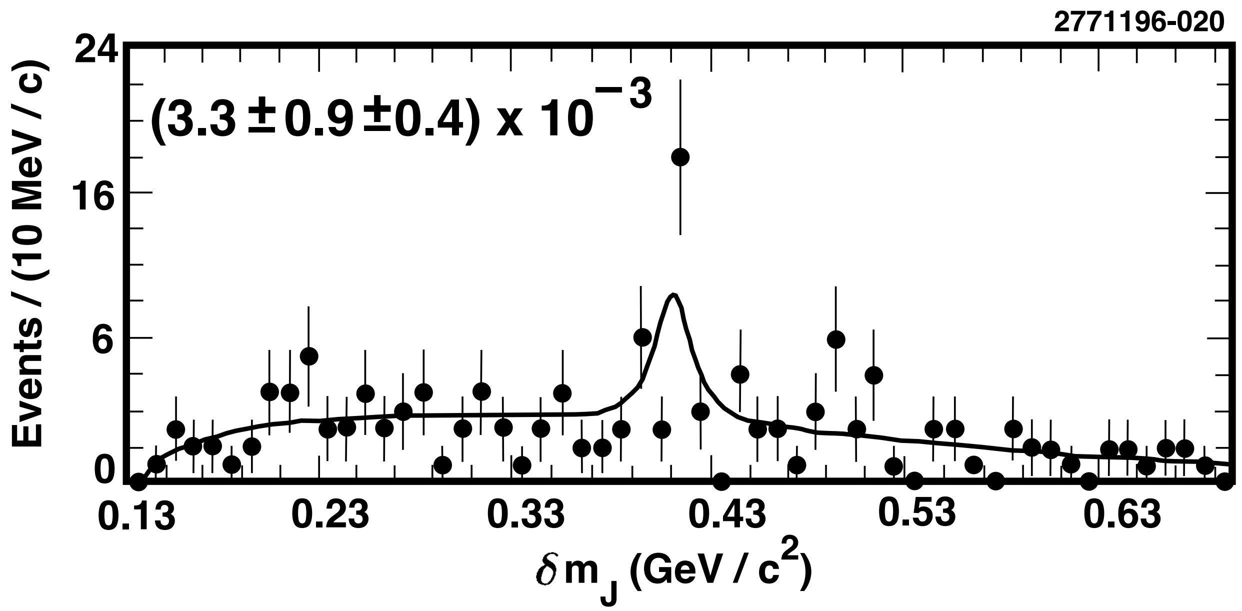
Youngjoon Kwon (Yonsei U.)

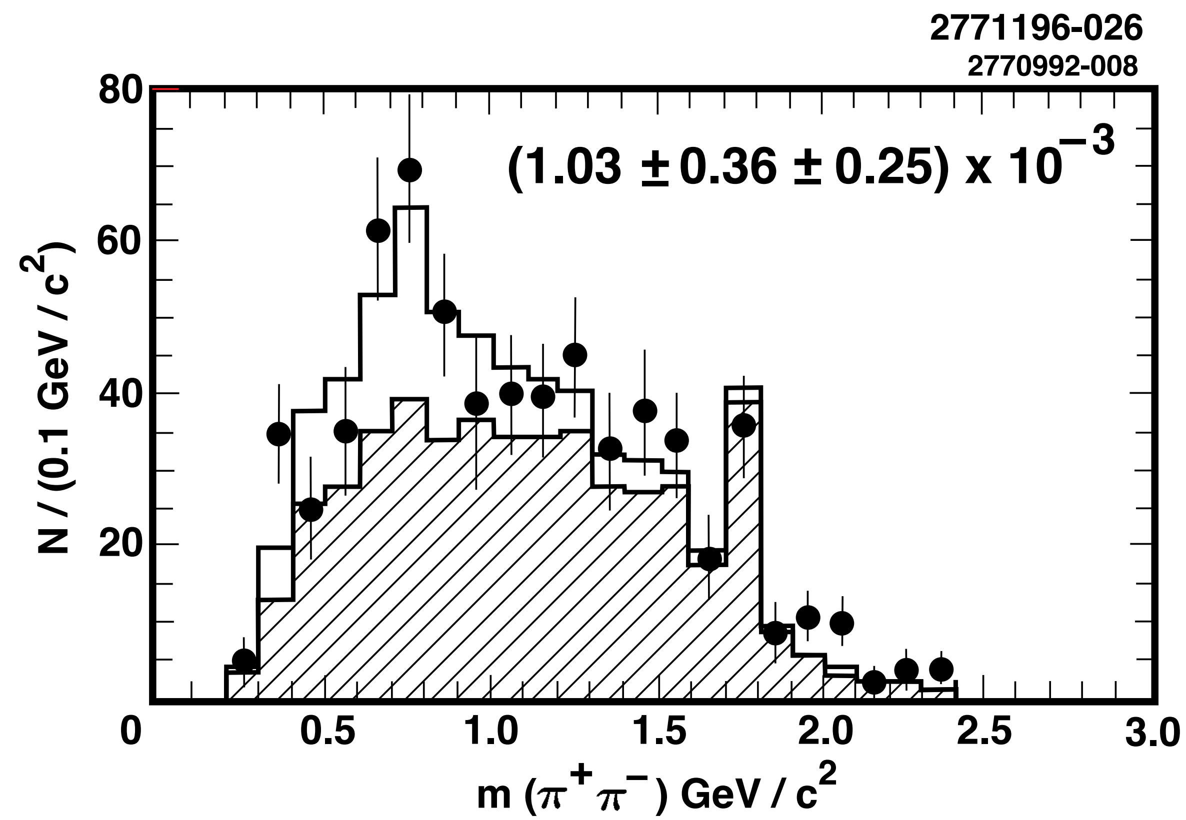
Oct. 13-25, AEPSHEP 2016

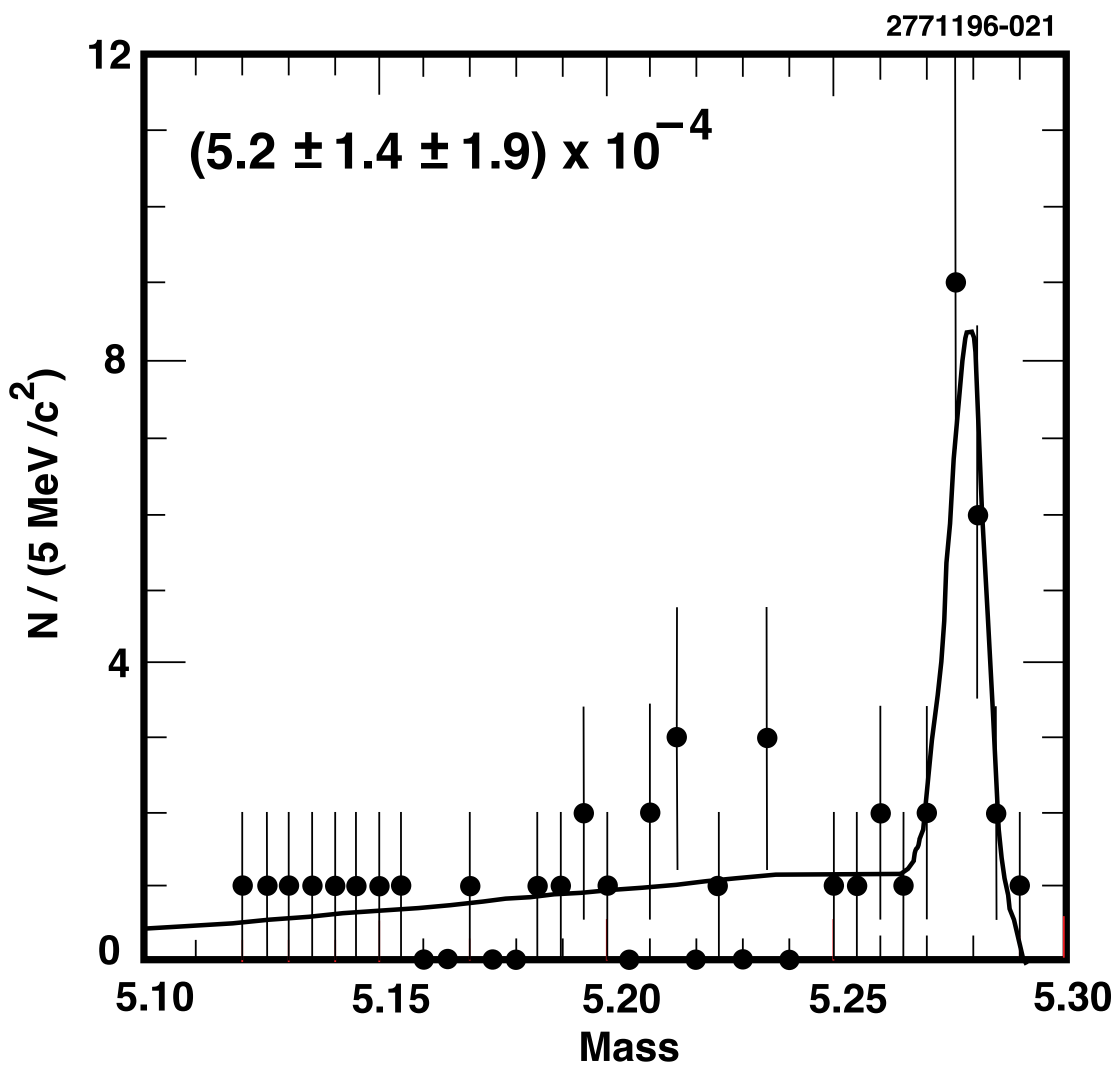
disclaimers

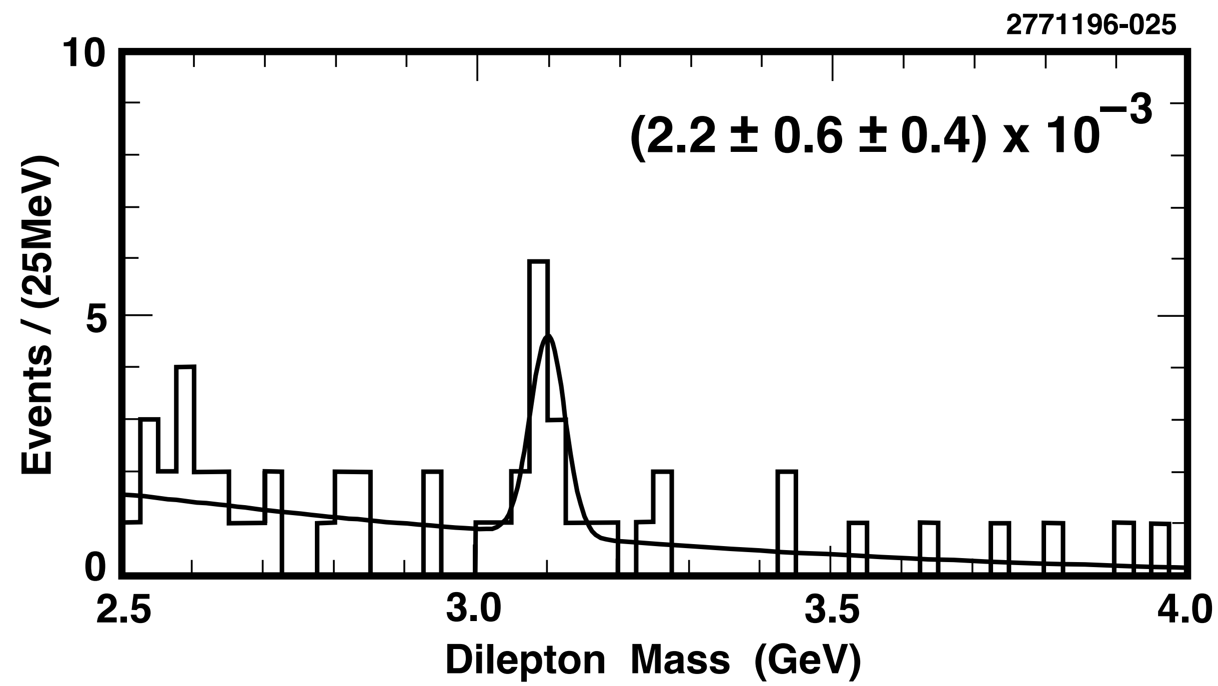
- freely taking from other people's lecture materials, w/o properly citing the references
  - just a rough list (from which I composed this lecture mostly) ...
- not paying attention to any mathematical rigor at all
- It will be impossible to cover "everything" even with the allocated time of 225 minutes…
  - so, I end up covering just a little fraction of the story, with a *subjective* choice of topics
- Please stop me any time if you don't follow the story, otherwise it will be merely a pointless series of slides.

Why bother with stat? How come not?



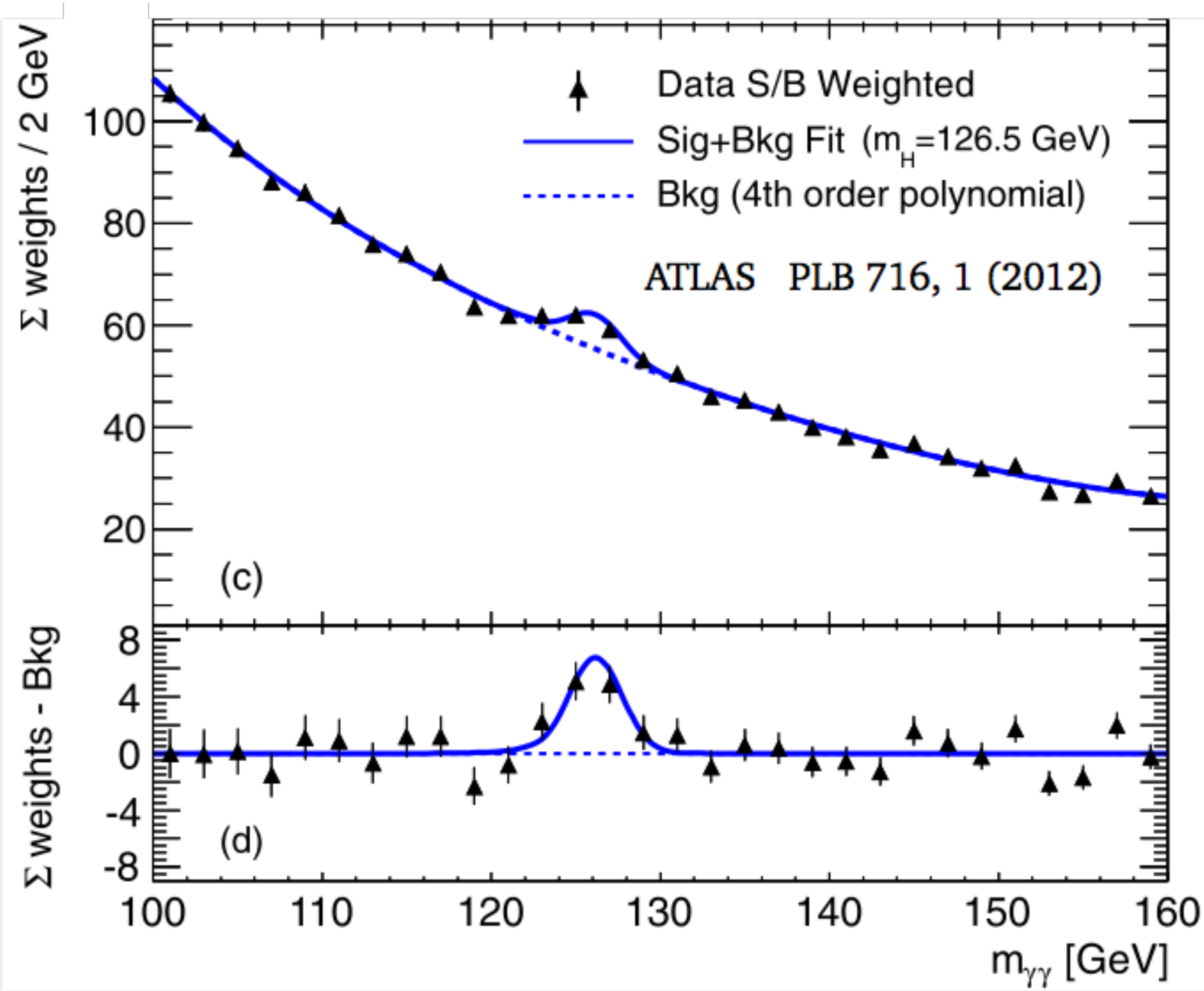




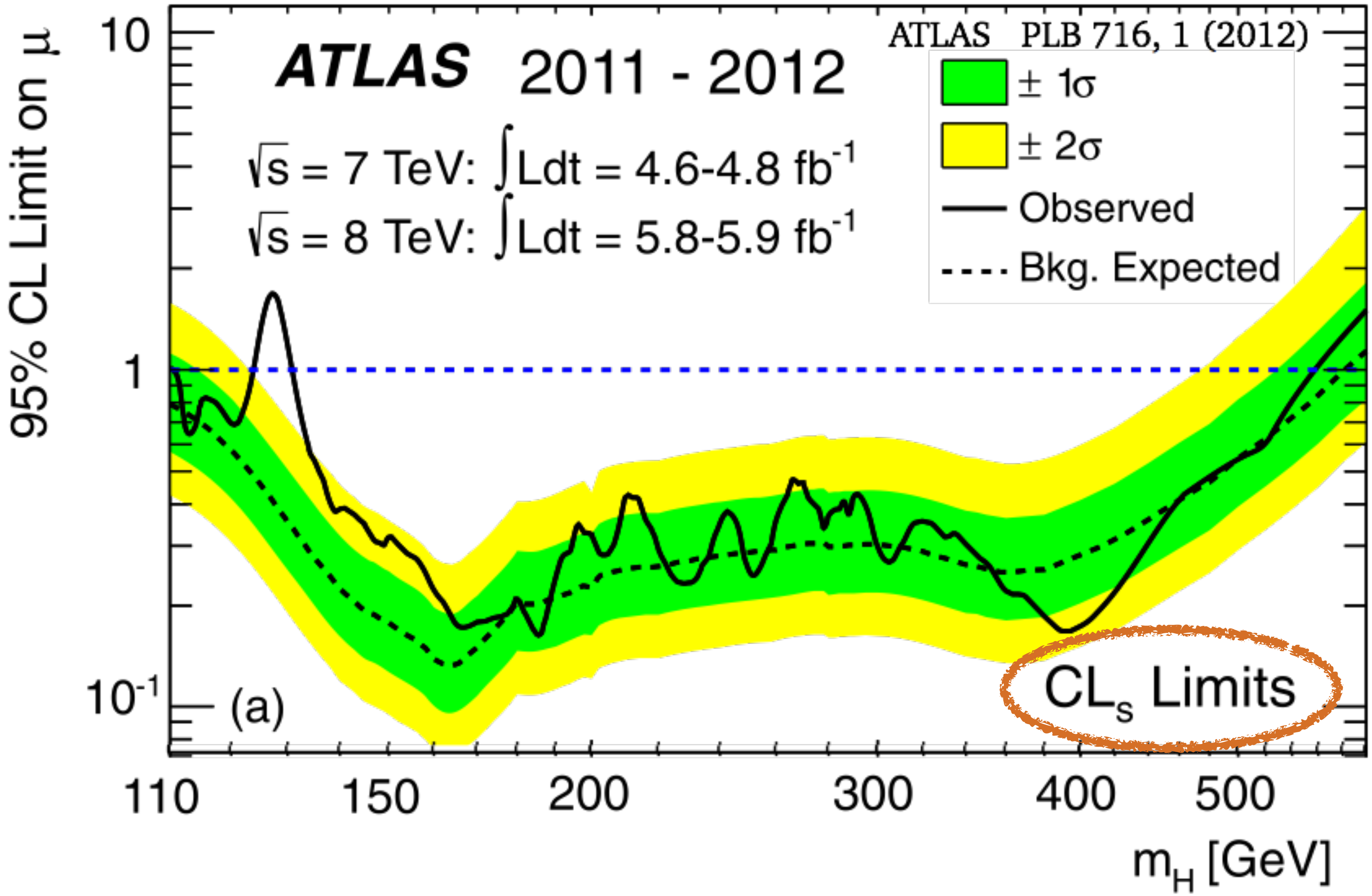


from 1997 TASI lecture by P. Drell

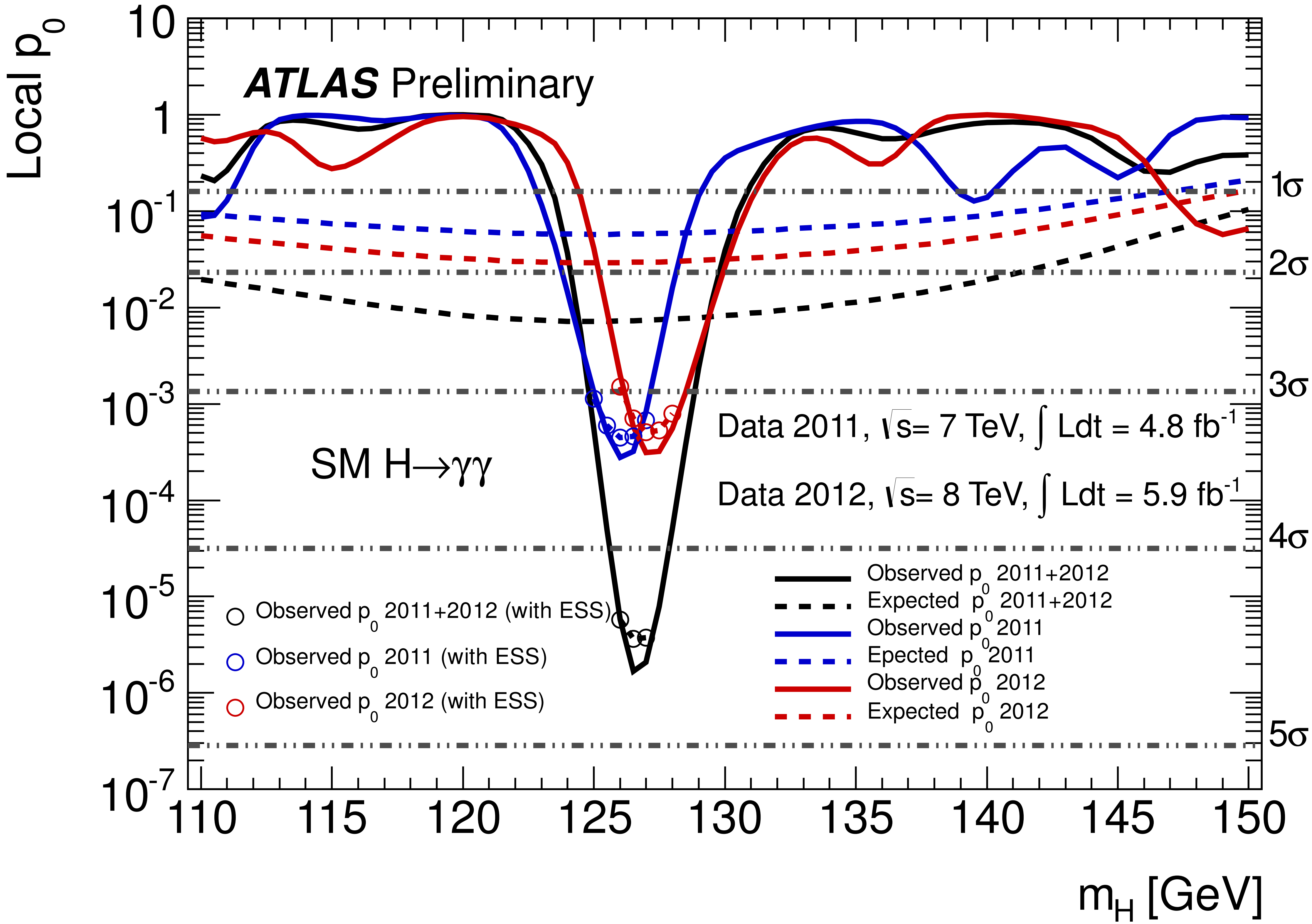
what to make sense of m_H plots, statistically



the green & yellow plots



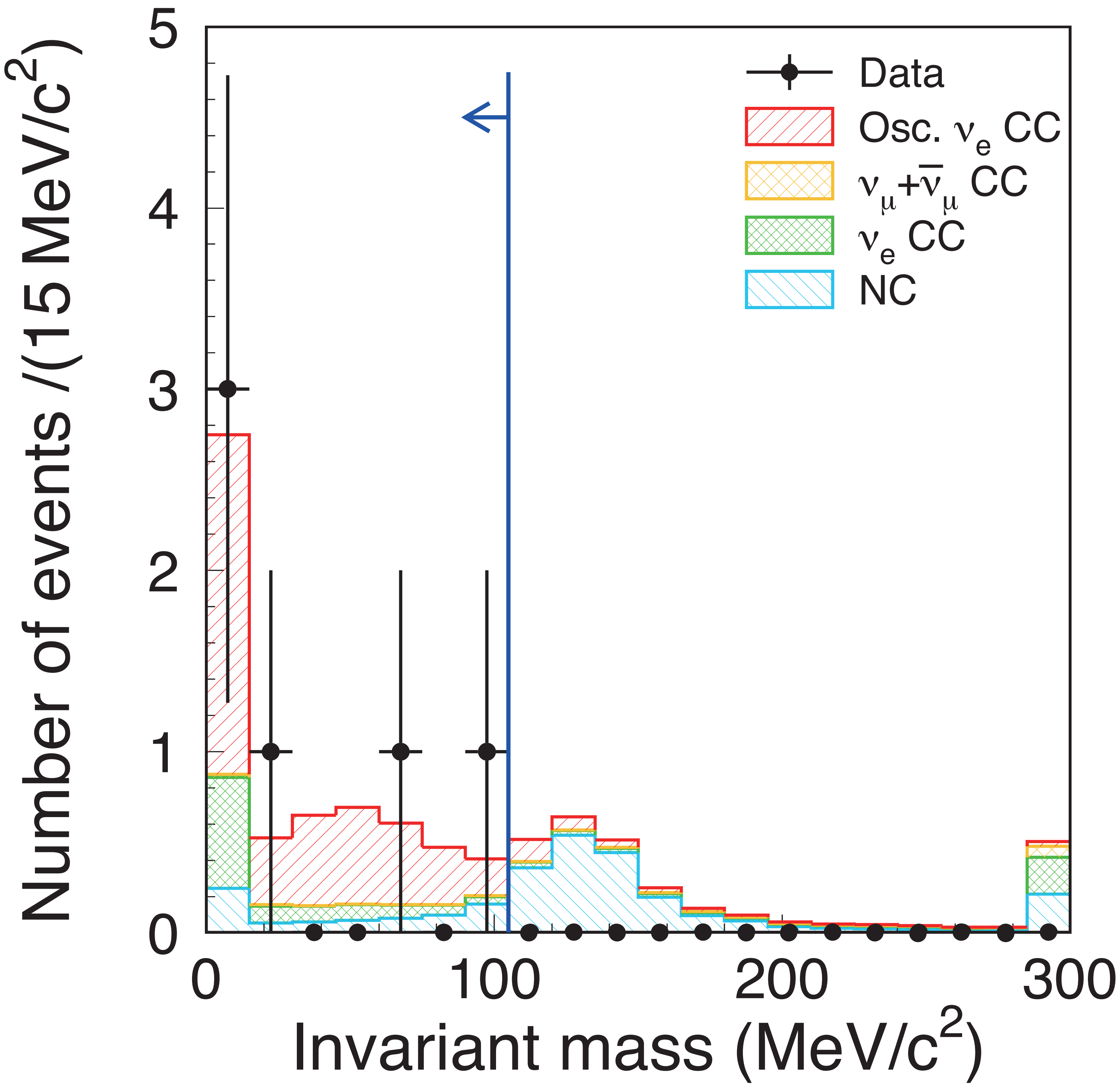
the p_0 plots



ATLAS Preliminary
Local p0
SM H→γγ
Data 2011, √s= 7 TeV, ∫ Ldt = 4.8 fb-1
Data 2012, √s= 8 TeV, ∫ Ldt = 5.9 fb-1
Observed p0 2011+2012
Expected p0 2011+2012
Observed p0 2011
Epected p0 2011
Observed p0 2012
Expected p0 2012
Observed p0 2011+2012 (with ESS)
Observed p0 2011 (with ESS)
Observed p0 2012 (with ESS)
1σ
2σ
3σ
4σ
5σ
10
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
110
115
120
125
130
135
140
145
150
mH [GeV]

(Example) T2K result

PRL 107, 041801 (2011)

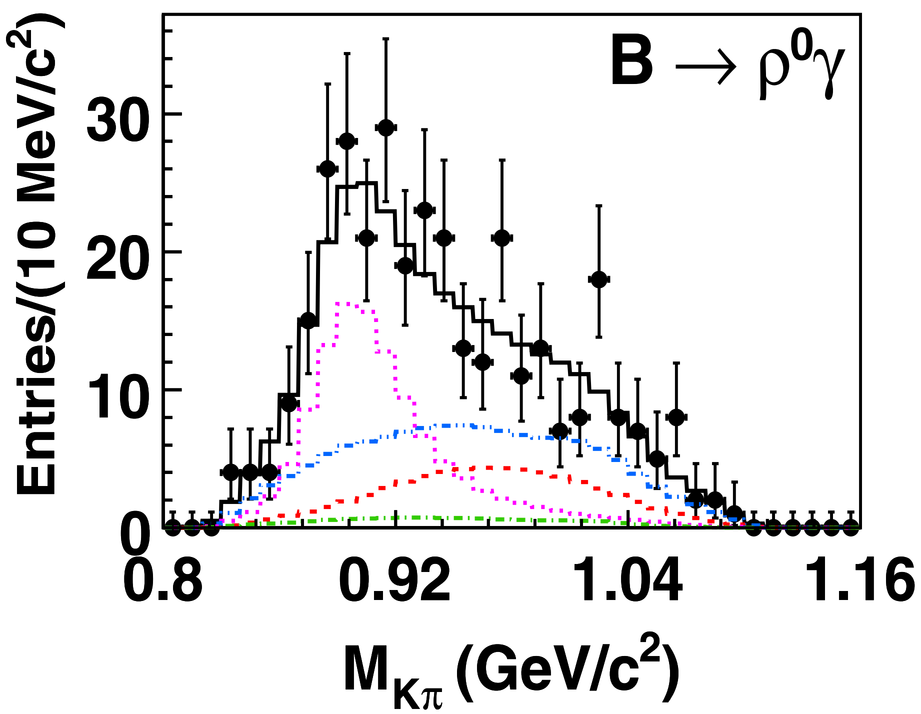


T2K observed 6 candidate events of $\nu_\mu \rightarrow \nu_e$ while a background of 1.5±03 events is expected.

- How significant is this signal?
- How to include the systematic uncertainty in the analysis?
- What is the relevant 'limit' from this result?

Particle ID and probabilities

Since $B \to K^* \gamma$ has much higher branching fraction than $B \to \rho\gamma$, the former can be a serious background to the latter. It is crucial to understand the “efficiency” and “fake rate” of K/π identification system of your experiment in this study. The figure below shows the $M_{K\pi}$ invarianbt mass distribution, where one of the pion mass (in $\rho^0 \to \pi^+\pi^-$ decay) is replaced by the Kaon mass, for the $B^0 \to \rho^0\gamma$ signal candidates (Belle, PRL 2008).



Express the following observables in Type-I & Type-II errors.

- $f_{\pi^+ \to K^+}$ = probability of misidentifying a π^+ as a K^+
- $f_{K^+ \to \pi^+}$ = probability of misidentifying a K^+ as a π^+
- ϵ_{K^+} = prob. of identifying a K^+ correctly as a K^+
- ϵ_{π^+} = prob. of identifying a π^+ correctly as a π^+

References *(a very rough list)*

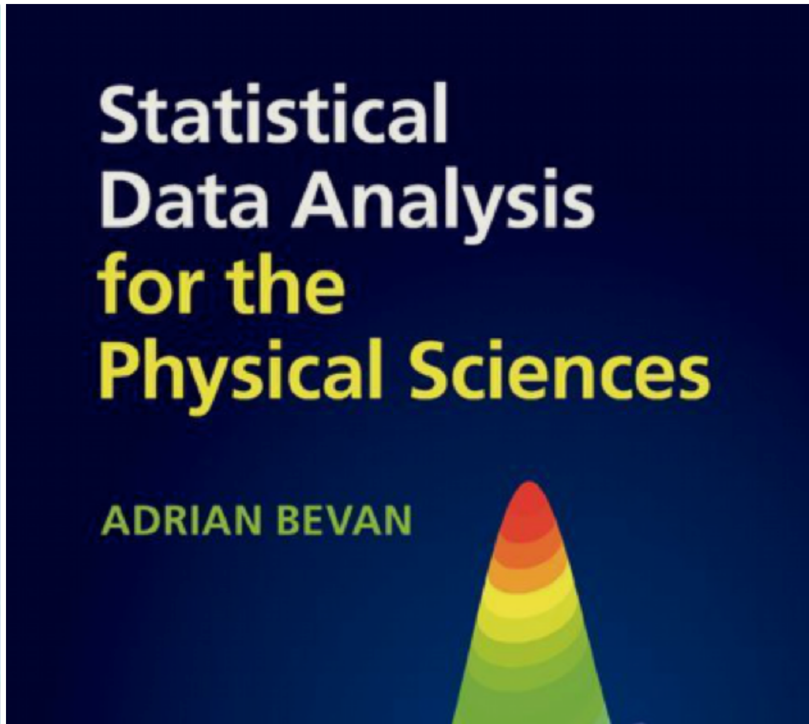
- "Statistical Data Analysis" by Glen Cowan

  http://www.pp.rhul.ac.uk/~cowan/stat_cern.html (lectures at CERN)

- "Statistical Data Analysis for the Physical Sciences" by Adrian Bevan (2013)
- Tom Junk @ TRIUMF, July 2009
- mini-reviews on Probability & Statistics in RPP (PDG)

  http://pdg.lbl.gov/2015/reviews/rpp2015-rev-statistics.pdf

- ...



Eur. Phys. J. C (2014) 74:3026
DOI 10.1140/epjc/s10052-014-3026-9

THE EUROPEAN PHYSICAL JOURNAL C

Review

The Physics of the B Factories

Received: 29 July 2014 / Accepted: 29 July 2014 / Published online: 19 November 2014

A. J. Bevan[*1], B. Golob[*2,3], Th. Mannel[*4], S. Prell[*5], B. D. Yabsley[*6],

K. Abe[§7], H. Aihara[§8], F. Anulli[§9,10], N. Arnaud[§11], T. Aushev[§12], M. Beneke[§13,14], J. Beringer[§15], F. Bianchi[§16,17], I. I. Bigi[§18], M. Bona[§16,17], N. Brambilla[§13], J. Brodzicka[§19], P. Chang[§20], M. J. Charles[§21], C. H. Cheng[§22], H.-Y. Cheng[§23], R. Chistov[§12], P. Colangelo[§24], J. P. Coleman[§25], A. Drutskoy[§12,26], V. P. Druzhinin[§27,28], S. Eidelman[§27,28], G. Eigen[§29], A. M. Eisner[§30], R. Faccini[§10,31], K. T. Flood[§22], P. Gambino[§16,17], A. Gaz[§32], W. Gradl[§33], H. Hayashii[§34], T. Higuchi[§35], W. D. Hulsbergen[§36], T. Hurth[§33], T. Iijima[§37,38], R. Itoh[§7], P. D. Jackson[§10,31], R. Kass[§39], Yu. G. Kolomensky[§15], E. Kou[§11], P. Križan[§2,3], A. Kronfeld[§40], S. Kumano[§41,42], Y. J. Kwon[§43], T. E. Latham[§44], D. W. G. S. Leith[§45], V. Lüth[§45], F. Martinez-Vid
R. Mussa[§16,17], M. Nakao[§7], S. Nishida[§7], J. Ocariz[§48], S. L. Olsen[§49], P. Pakhlov[§
A. Palano[§24,51], A. Pich[§52], S. Playfer[§53], A. Poluektov[§27,28], F. C. Porter[§22], S. H. R
A. Roodman[§45], Y. Sakai[§7], C. Schwanda[§56], A. J. Schwartz[§47], R. Seidl[§57], S. J. Sel
K. Sumisawa[§7], E. S. Swanson[§60], F. Tackmann[§61], K. Trabelsi[§7], S. Uehara[§7], S. Ur
G. Vasseur[§62], W. Verkerke[§63], R. Waldi[§64], M. Z. Wang[§20], F. F. Wilson[§65], J. Zupar

I. Adachi[¶7], J. Albert[¶55], Sw. Banerjee[¶55], M. Bellis[¶66], E. Ben-Haim[¶48], P. Biassoni[¶67,68], R. N. Cahn[¶15],
C. Cartaro[¶45], J. Chauveau[¶48], C. Chen[¶5], C. C. Chiang[¶20], R. Cowan[¶69], J. Dalseno[¶70], M. Davier[¶11]

* General Editor
§ Section Editor
¶ Additional Section Writer
† Deceased

Outline

- Basic elements
  - some vocabulary
  - Probability axioms
  - some probability distributions
- Two approaches: Frequentist vs. Bayesian

- Hypothesis testing
- Parameter estimation
- Other subjects — “nuisance”, “spurious”, “look elsewhere”

Part I

Parts II & III

Basic elements

some vocabulary

- statistics, probability
- random variables, PDF, CDF
- expectation values
- mean, median, mode
- standard deviation, variance, covariance matrix
- correlation coefficients
- weighted average and error
- ...

Statistics & Probability

Statistics is largely the inverse problem of probability.

- Probability:

  Know parameters of the theory $\Rightarrow$ predict distributions of possible experimental outcomes

- Statistics:

  Know the outcome of an experiment $\Rightarrow$ extract information about the parameters and/or the theory

  - Probability is the easier of the two – *more straightforward*.
  - Statistics is what we need as HEP analysts.
  - In HEP, the statistics issues often get very complex because we know so much bout our data and need to incorporate all of what we find.

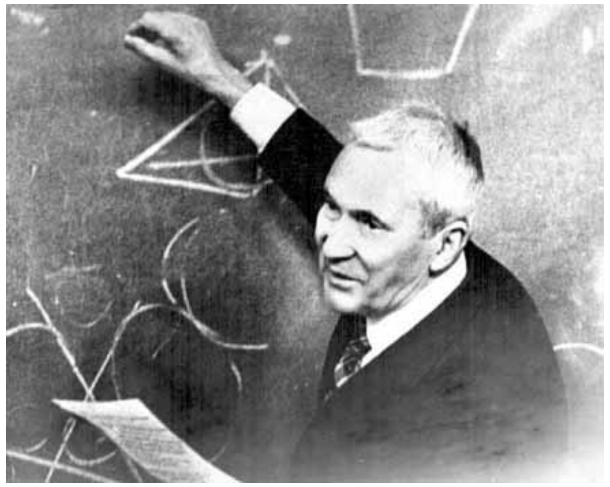
Probability Axioms

Consider a set S with subsets A, B, ...

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$



Kolmogorov (1933)

Also define conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note: $P(A|B) \neq P(B|A)$

Consider an extreme case (*with some made-up numbers*)

- Ω: all people
- $P(\text{woman}) = 50\%$
- $P(\text{pregnant}|\text{woman}) = 3\%$
- $P(\text{pregnant}) = 1.5\%$
- $P(\text{woman}|\text{pregnant}) = 100\%$

a consequence of $P(A|B) \neq P(B|A)$

$$P(\text{data}|\text{theory}) \neq P(\text{theory}|\text{data})$$

Random variables and PDFs

- A random variable is a numerical characteristic assigned to an element of the sample space; it can be discrete or continuous.
- Suppose outcome of experiments is continuous:

$$P(x \in [x, x + dx]) = f(x)dx$$

$\Rightarrow f(x)$ is the **probability density function** (PDF) with

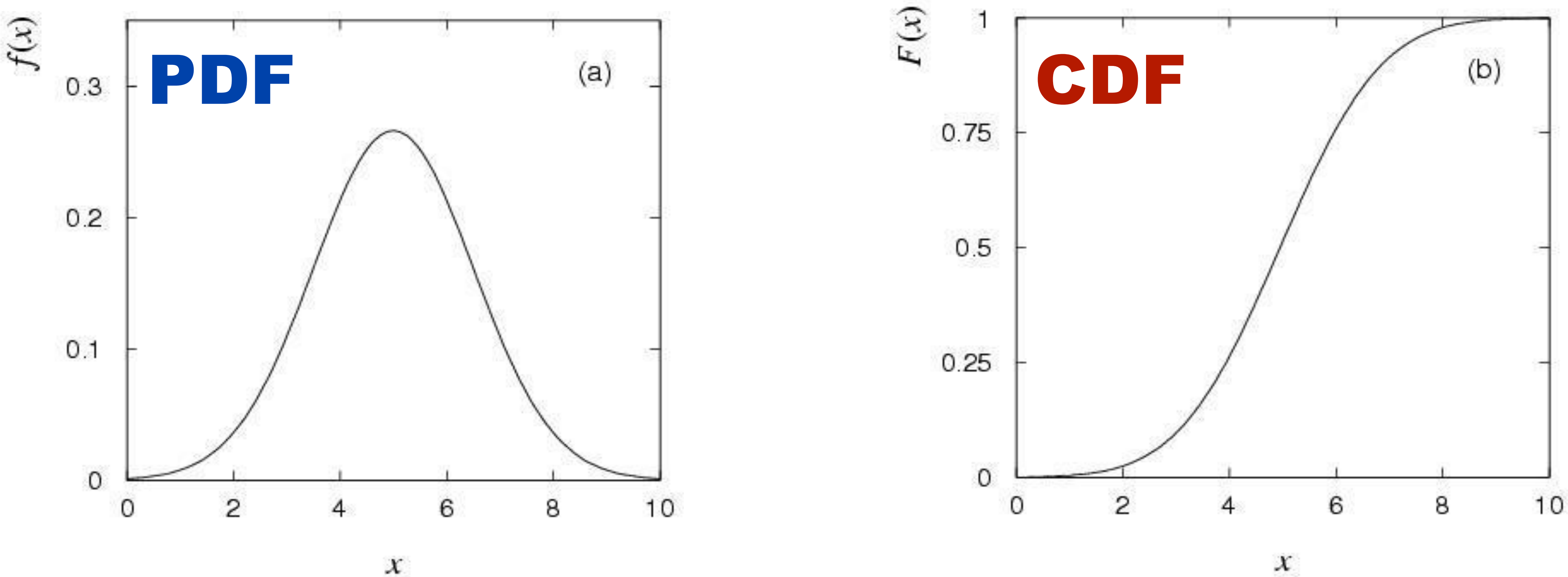
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

- Or, for discrete outcome x_i with e.g. $i = 1, 2, \cdots$
  * $P(x_i) = p_i$ "**probability mass function**"
  * $\sum_i P(x_i) = 1$

Cumulative distribution function (CDF)

- The probability $F(x)$ to have an outcome less than or equal to x is called the **cumulative distribution function** (CDF).

$$\int_{-\infty}^{x} f(x')dx' \equiv F(x) \ .$$



- Alternatively, we have $f(x) = \partial F(x)/\partial x$.

Expectation value

$g(X),\ h(X)$: functions of random variable X

- for discrete $X \in \Omega$

$$E(g) = \sum_{\Omega} P(X)\, g(X)$$

- for continuous $X \in \Omega$

$$E(g) = \int_{\Omega} dX\, f(X)\, g(X)$$

- E is a linear operator

$$E[\alpha g(X) + \beta h(X)] = \alpha E[g(X)] + \beta E[h(X)]$$

Examples of expectation values

- **mean** – expectation value for the PDF ($f(X)$ or $P(X_i)$)

$$\mu = \overline{X} = E(X) = \langle X \rangle = \int_\Omega dX\, f(X) X$$

- **variance** – it may not always exist!

$$\begin{aligned}\sigma^2 = V(X) &= E[(X-\mu)^2] \\ &= E(X^2) - [E(X)]^2 \\ &= \int_\Omega dX\, f(X)(X-\mu)^2\end{aligned}$$

sample mean & sample variance

- n measurements $\{x_i\}$ where x_i follows $N(\mu, \sigma)$
- sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

With more measurements, the estimation of the mean will become more accurate.

- sample variance

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

Sample variance approaches σ^2 for large n.

Mean and Variance in 2-D

- Expectation value in 2-D: (X, Y) as RV

$$E[g(X,Y)] = \iint_\Omega dX\, dY\, f(X,Y)\, g(X,Y)$$

$\Rightarrow$ Extension to higher dimension is straightforward!

- **mean** of X

$$\mu_X = E[X] = \iint_\Omega dX\, dY\, f(X,Y)\, X$$

- **variance** of X

$$\sigma_X^2 = E[(X-\mu_X)^2] = \iint_\Omega dX\, dY\, f(X,Y)\, (X-\mu_X)^2$$

Covariance matrix

- Given a n-dimensional random variable $\vec{X} = (X_1, \cdots, X_n)$, the covariance matrix C_{ij} is defined as:

$$\begin{aligned} C_{ij} &= E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= E[X_i X_j] - \mu_i \mu_j \end{aligned}$$

- more intuitive is the **correlation coefficient**, ρ_{ij}, given by

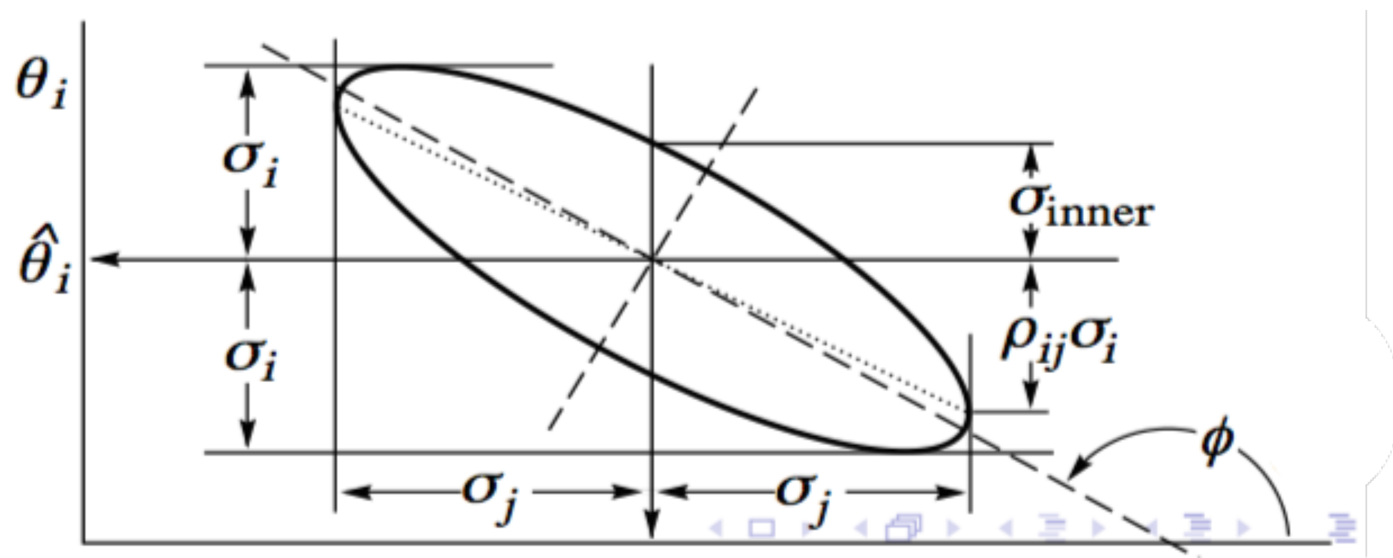
$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j}$$

properties of covariance matrix

- bounded by one: $-1 \leq \rho_{ij} \leq +1$
- for independent variables X, Y: $\rho(X, Y) = 0$

  But the reverse is not true! (e.g. $Y = X^2$)
- If $f(X_1, \cdots, X_n)$ is a multi-dim. Gaussian, then $cov(X_i, X_j)$ gives the *tilt* of the ellipsoid in (X_i, X_j)

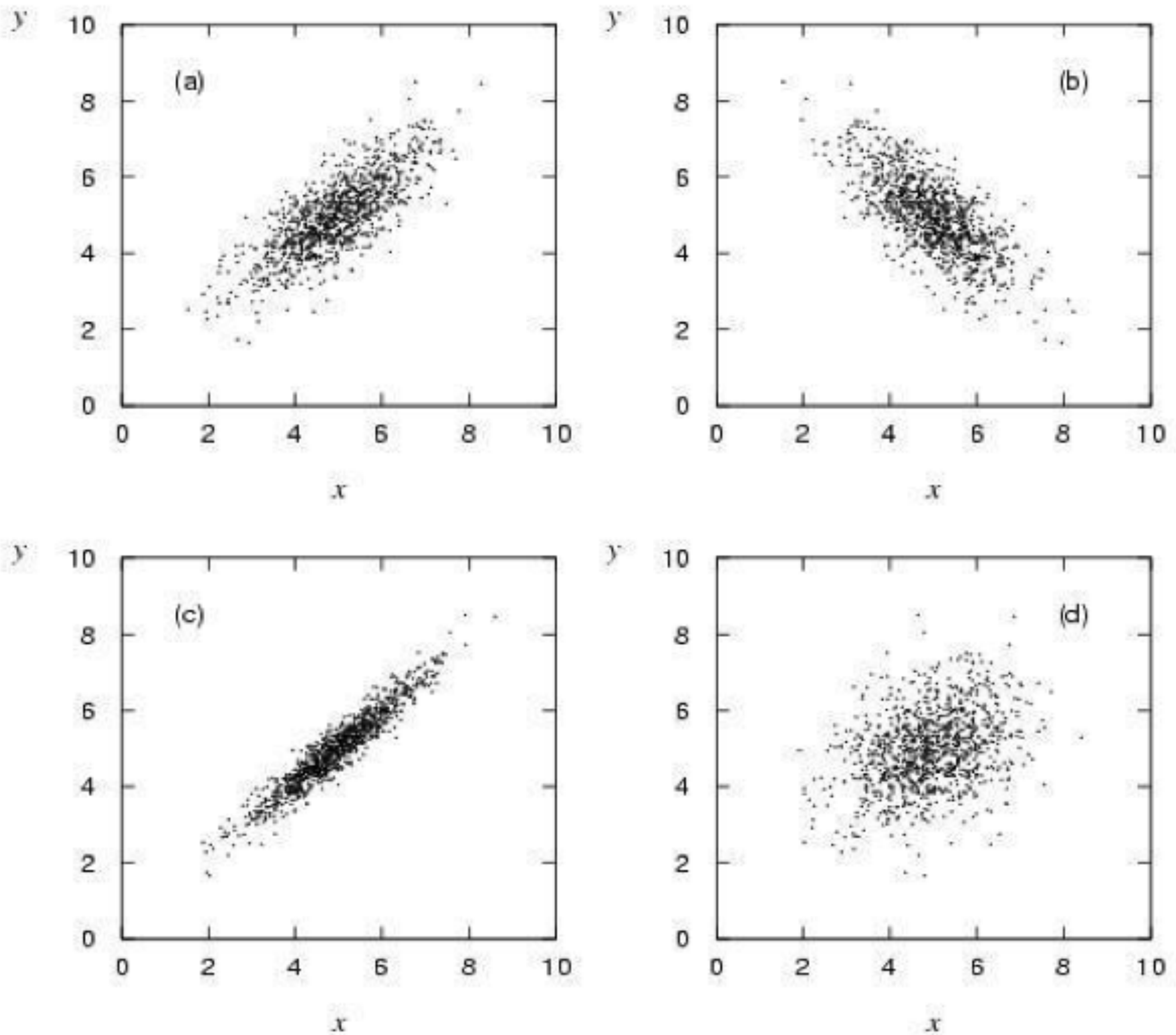
$$\tan 2\phi = \frac{2\,\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j)}{\sigma_j^2 - \sigma_i^2} = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}$$



Correlations - 2D examples

$\rho = 0.75$

$\rho = -0.75$

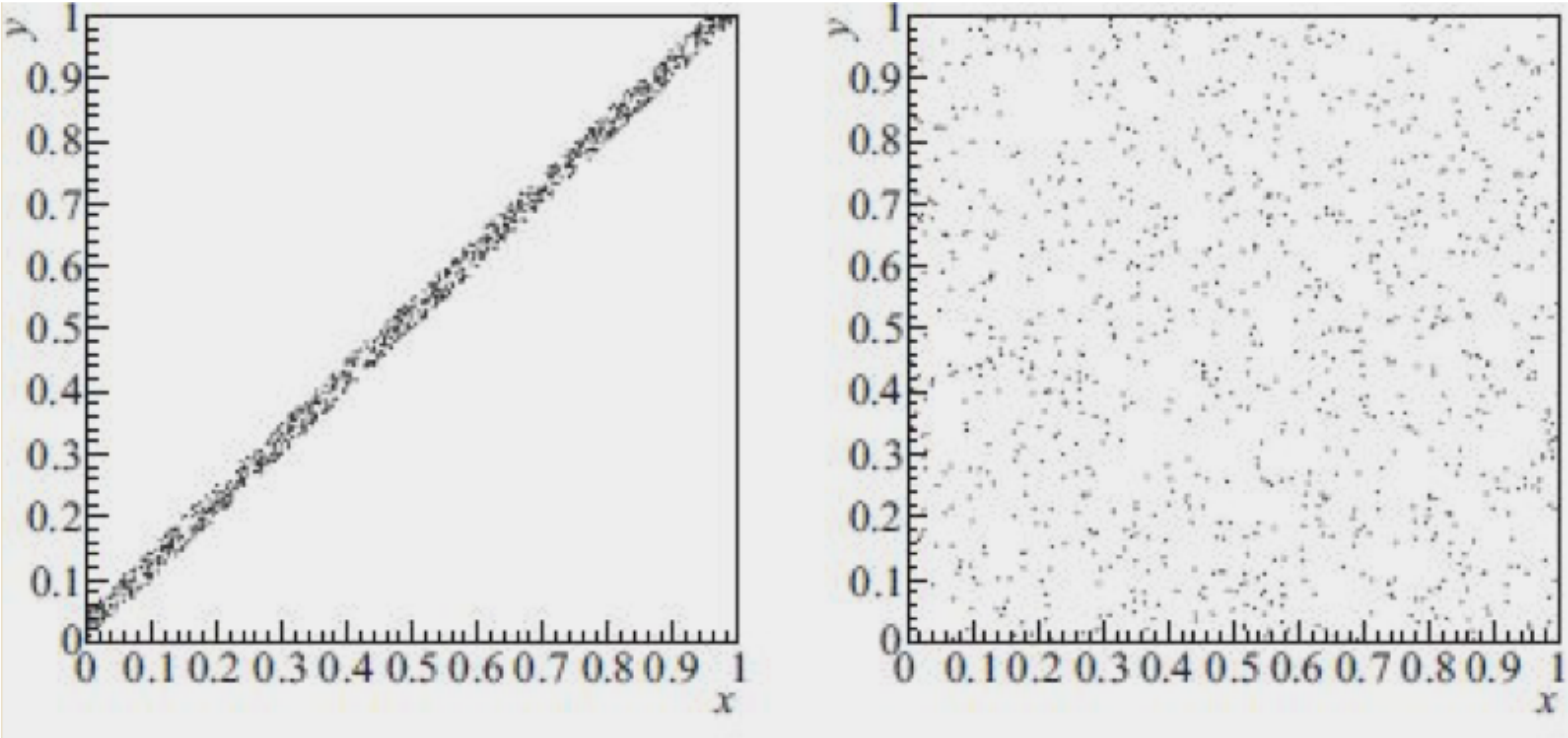


$\rho = 0.95$

$\rho = 0.25$

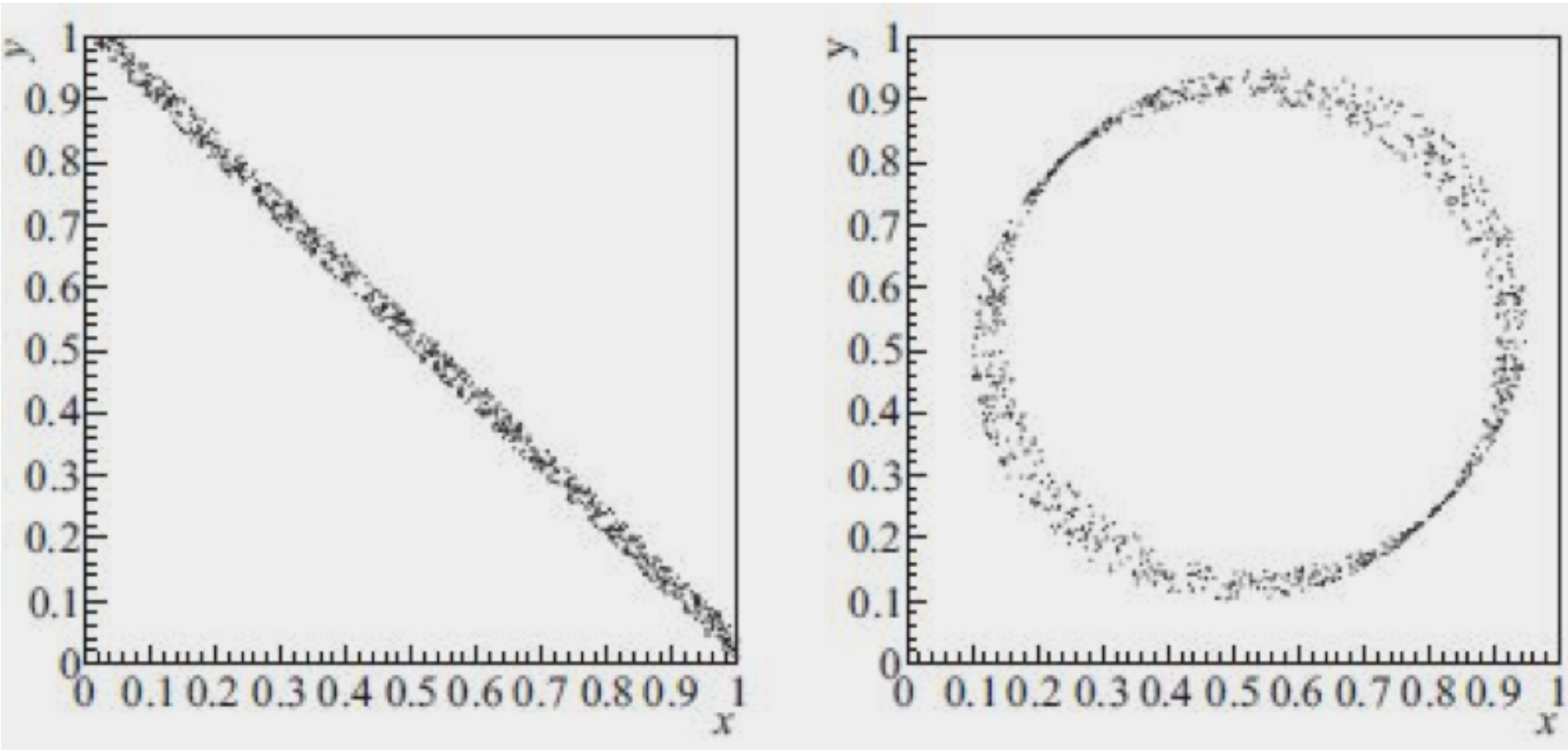
(Quiz time)

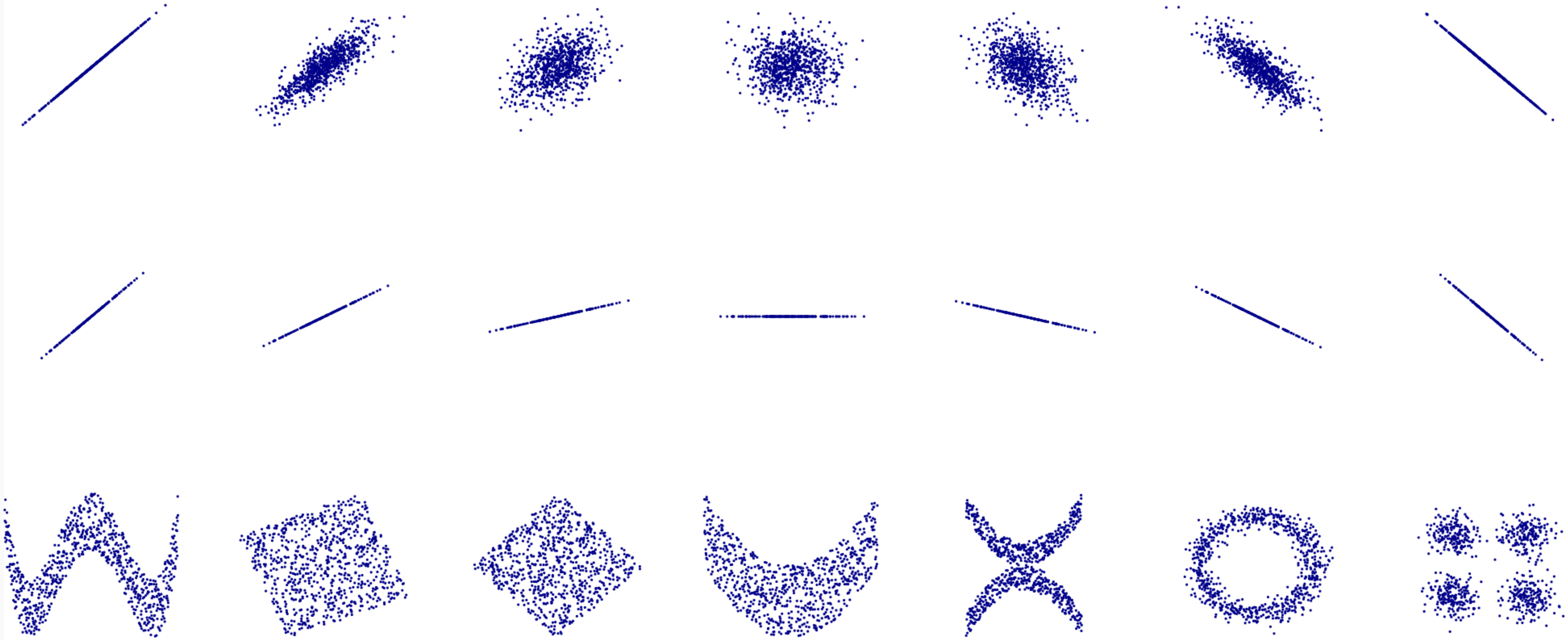
- ρ =?
- Are x and y correlated?



(Quiz time)

- ρ =?
- Are x and y correlated?





from https://en.wikipedia.org/wiki/Correlation_and_dependence

Error propagation on *f(x,y)*

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix}$$

(Q) What if x and y are independent?

(HW) Obtain the error on $f(x,y) = C\,x/y$

If *x* and *y* are uncorrelated (independent),

$$\Rightarrow \sigma_f^2 = (f_x \ \ f_y)\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}\begin{pmatrix} f_x \\ f_y \end{pmatrix}$$

$f_x \equiv \partial f / \partial x$, etc. $\qquad = \sigma_x^2 \left(\frac{\partial f}{\partial x}\right)^2 + \sigma_y^2 \left(\frac{\partial f}{\partial y}\right)^2$

$$f(x, y) = Cx/y \ \rightarrow \ \delta f / f = \sqrt{(\delta x/x)^2 + (\delta y/y)^2}$$

If *x* and *y* are 100% (+) correlated, e.g. $y = \alpha x$

$$\sigma_f^2 = (f_x \ \ f_y)\begin{pmatrix} \sigma_x^2 & \alpha\sigma_x^2 \\ \alpha\sigma_x^2 & \sigma_y^2 (= \alpha^2\sigma_x^2) \end{pmatrix}\begin{pmatrix} f_x \\ f_y \end{pmatrix}$$

$$= \sigma_x^2 \left(\frac{f}{x} - \frac{f}{y}\alpha\right)^2 = 0$$

$\delta y = \alpha\, \delta x$. $(\alpha > 0)$

$\rho_{ij} = V_{ij} / \sigma_i \sigma_j$

$+1 = \dfrac{V_{ij}}{\sigma_x \ \sigma_y}$, $\sigma_y = \alpha\,\sigma_x$

Weighted average and error

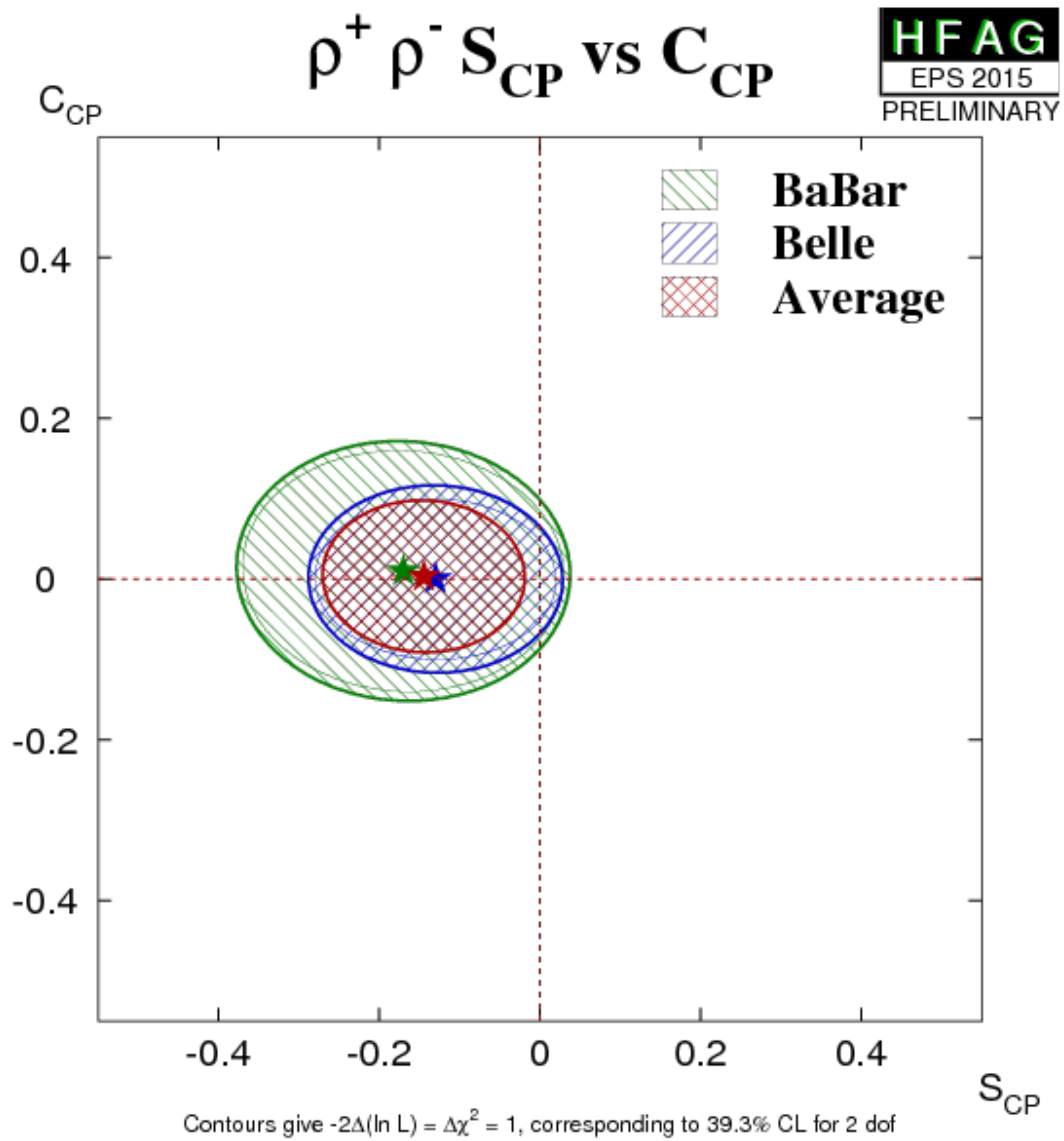
- **How to combine uncorrelated measurements (x_j, σ_j) with different amount of errors?**

$$\overline{x} \pm \sigma_x = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \pm \left(\sum_{i=1}^{n} 1/\sigma_i^2\right)^{-1/2}$$

- **What will happen if the measurements are correlated?**

$$\overline{x} = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1} \cdot \left[\sum_{j=1}^{M} V_j^{-1} x_j\right] \qquad V = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1}$$

(Ex) to measure S and C from $B^0 \to \rho^+\rho^-$



(Ex) to measure S and C from $B^0 \to \rho^+\rho^-$

Exp	S	C	V$_{ij}$
BaBar	-0.17 ±0.21	-0.01 ±0.16	-0.0012
Belle	-0.13 ±0.16	0.00 ±0.12	0.00033

Let $j = 1$ for BaBar, $= 2$ for Belle

- Obtain $\vec{X}^{(j)}$ where $X_1 = S$, $X_2 = C$
- Obtain $V = [V_1^{-1} + V_1^{-1}]^{-1}$
- Calculate the weighted average of S and C, and their errors

some useful distributions

Distribution	Probability density function f (variable; parameters)	Characteristic function $\phi(u)$	Mean	Variance σ^2
Uniform	$f(x;a,b) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{e^{ibu} - e^{iau}}{(b-a)iu}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Binomial	$f(r;N,p) = \dfrac{N!}{r!(N-r)!}\; p^r q^{N-r}$ $r = 0,1,2,\ldots,N\ ;\quad 0 \le p \le 1\ ;\quad q = 1-p$	$(q + pe^{iu})^N$	Np	Npq
Poisson	$f(n;\nu) = \dfrac{\nu^n e^{-\nu}}{n!}\ ;\quad n = 0,1,2,\ldots\ ;\quad \nu > 0$	$\exp[\nu(e^{iu} - 1)]$	ν	ν
Normal (Gaussian)	$f(x;\mu,\sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}}\ \exp(-(x-\mu)^2/2\sigma^2)$ $-\infty < x < \infty\ ;\quad -\infty < \mu < \infty\ ;\quad \sigma > 0$	$\exp(i\mu u - \frac{1}{2}\sigma^2 u^2)$	μ	σ^2
Multivariate Gaussian	$f(\boldsymbol{x};\boldsymbol{\mu},V) = \dfrac{1}{(2\pi)^{n/2}\sqrt{\lvert V\rvert}}$ $\times \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$ $-\infty < x_j < \infty;\quad -\infty < \mu_j < \infty;\quad \lvert V\rvert > 0$	$\exp\left[i\boldsymbol{\mu}\cdot\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T V\boldsymbol{u}\right]$	$\boldsymbol{\mu}$	V_{jk}
χ^2	$f(z;n) = \dfrac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)}\ ;\quad z \ge 0$	$(1-2iu)^{-n/2}$	n	$2n$

Binomial distribution

- **Given a repeated set of N trials, each of which has probability p of "success" (hence $1-p$ of "failure"), what is the distribution of the number of successes if the N trials are repeated over and over?**

$$\mathrm{Binom}(k \mid N, p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{Np(1-p)}$$

where k is the number of success trials

- (Ex) events passing a selection cut, with a fixed total *N*

$$\epsilon = \frac{N_{\mathrm{pass}}}{N}$$

$$\sigma_\epsilon = \sigma_{N_{\mathrm{pass}}}/N = \sqrt{Np(1-p)}/N = \sqrt{p(1-p)/N}$$

Binomial error: an example

- **What is the uncertainty σ_A on an asymmetry given by $A = (N_1 - N_2)/(N_1 + N_2)$, where $N_1 + N_2 = N$ is the (fixed) total # of events obtained in the counting experiment? Take, e.g., $N_1 = 80$ and $N_2 = 20$.**

Poisson distribution

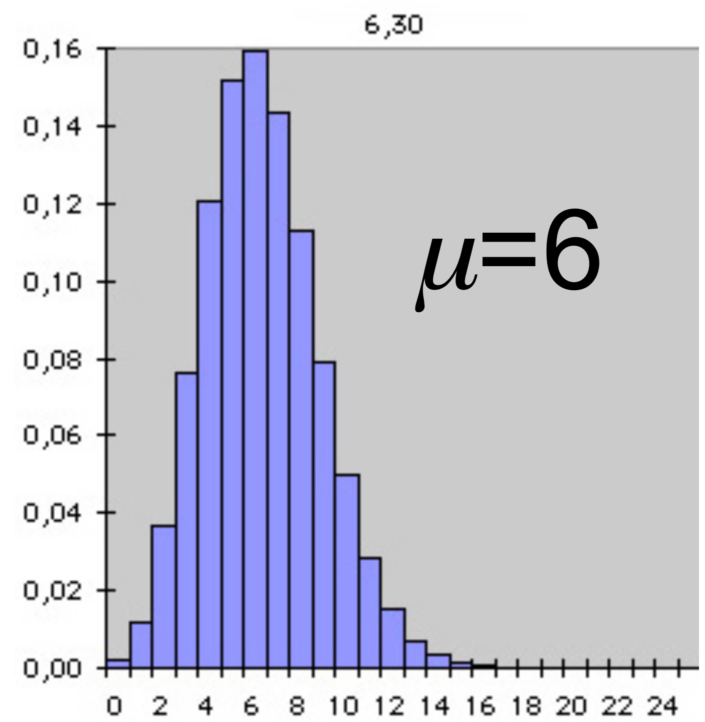
- Limit of Binomial when $N \to \infty$ and $p \to 0$ with $Np = \mu$ being finite and fixed $\Rightarrow$ Poisson distribution

$$f_P(k|\mu) = \frac{e^{-\mu}\mu^k}{k!}, \quad \sigma(k) = \sqrt{\mu}$$

Normalized in two different ways:

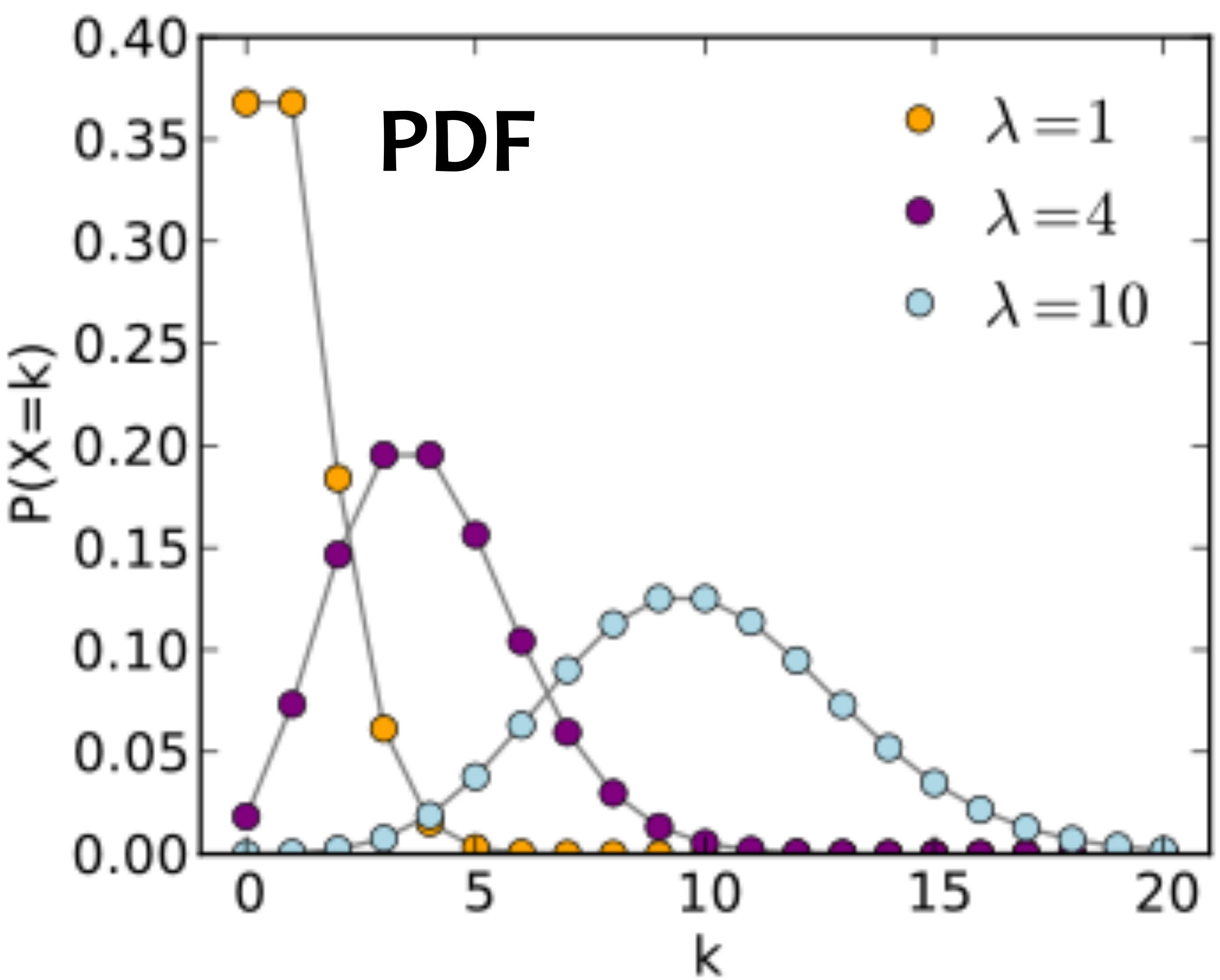
$$\sum_{k=0}^{\infty} f_P(k|\mu) = 1, \ \forall\mu$$

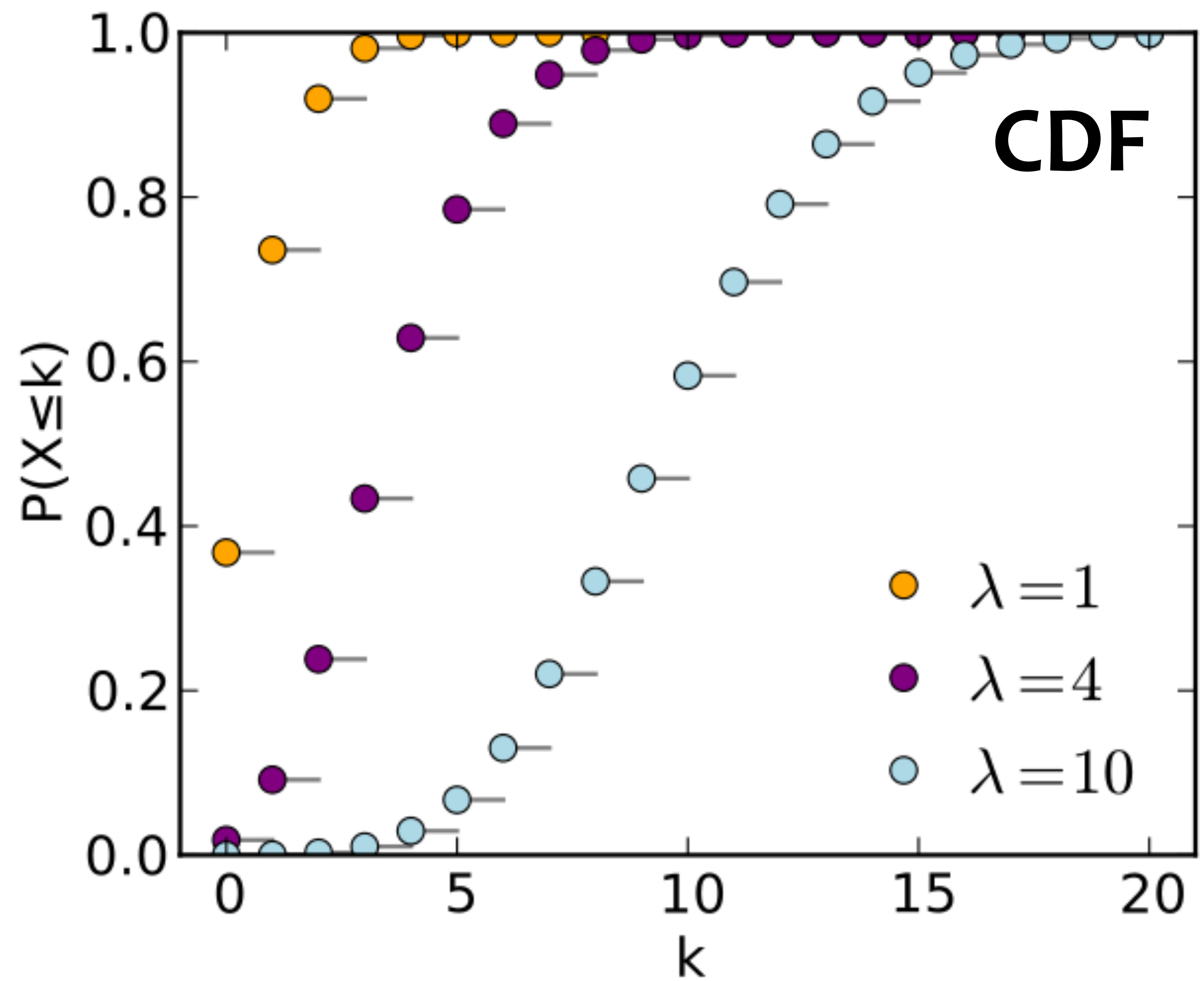
$$\int_0^{\infty} f_P(k|\mu) d\mu = 1, \ \forall k$$



All counting results in HEP are assumed to be Poisson-distributed

Poisson distribution

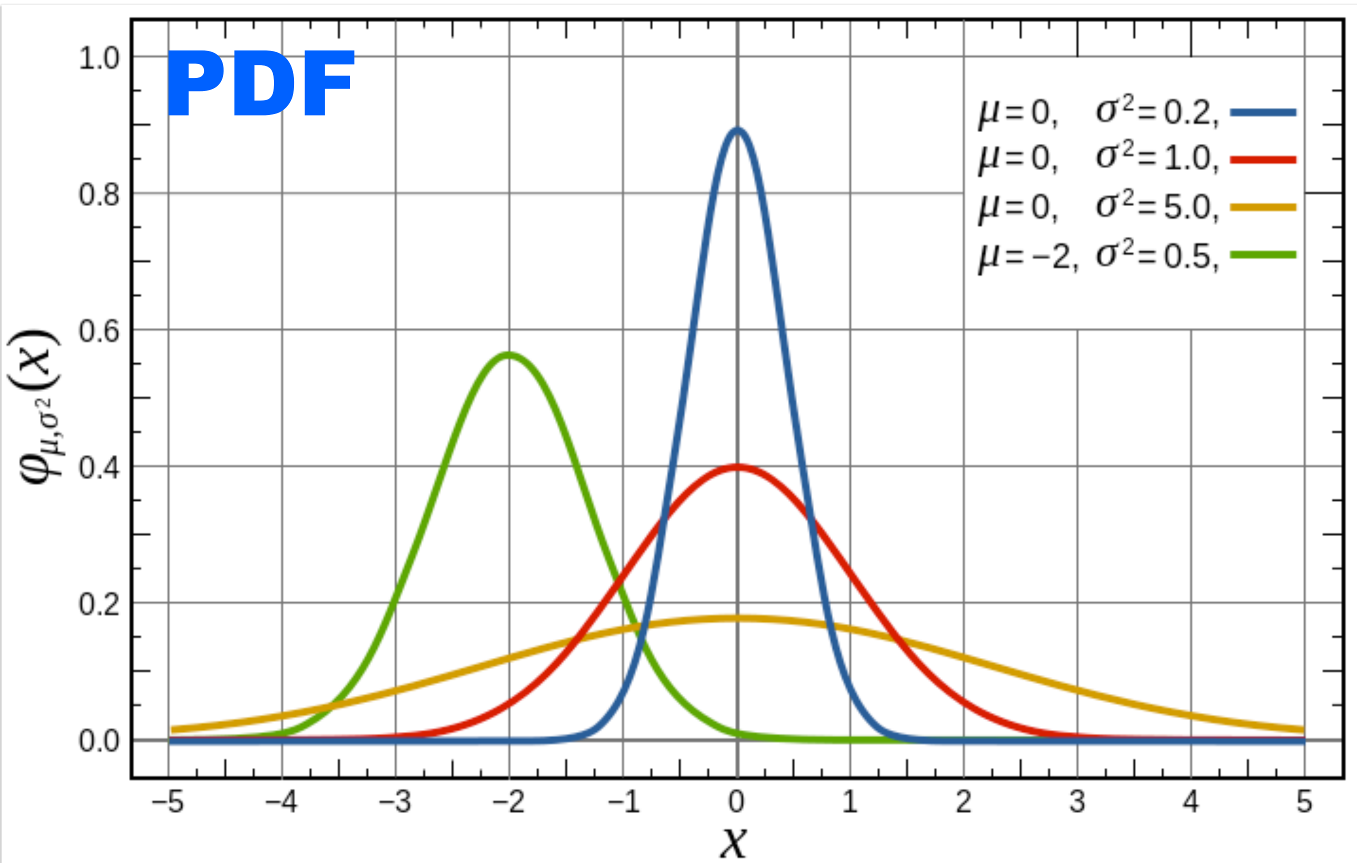


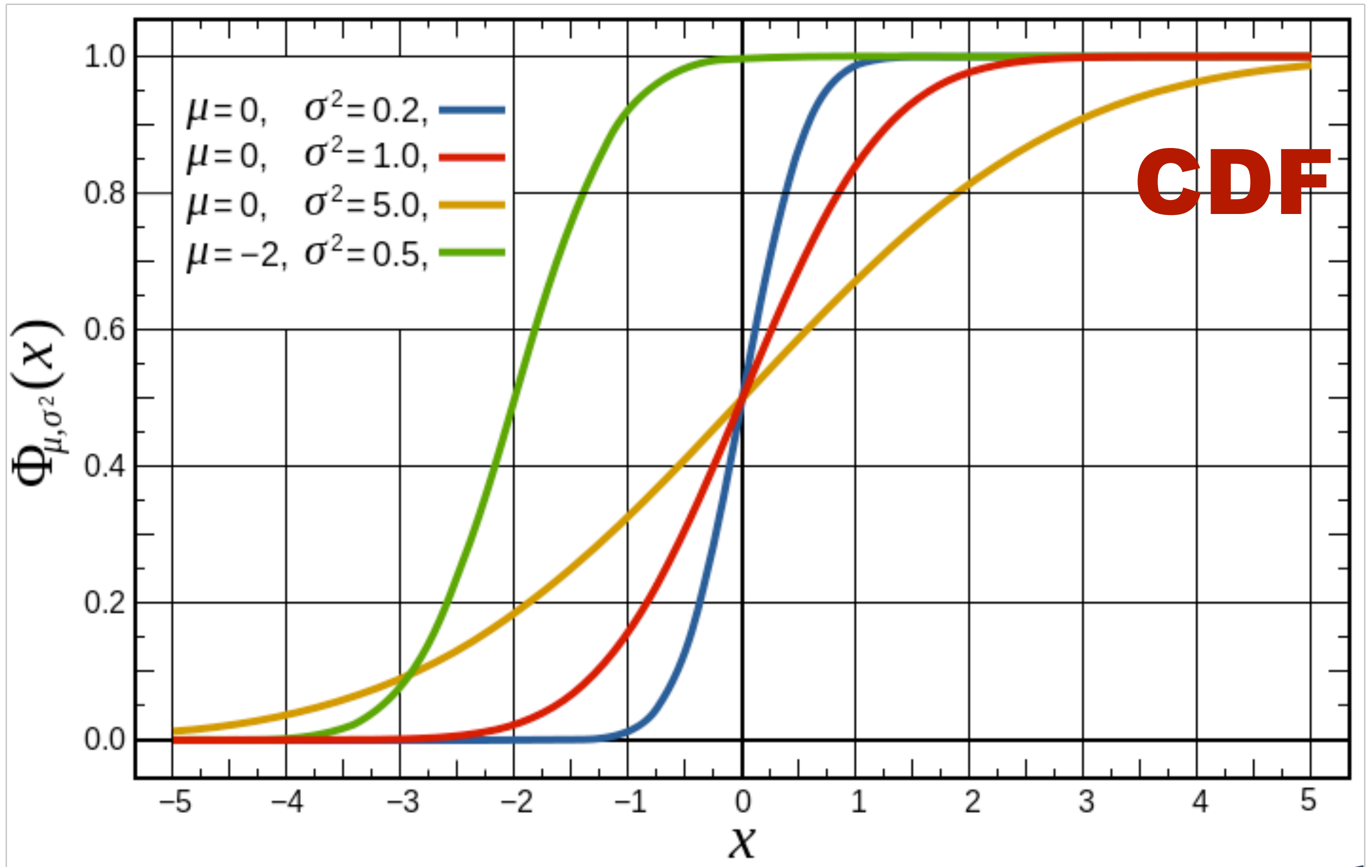


Gaussian (Normal) distribution

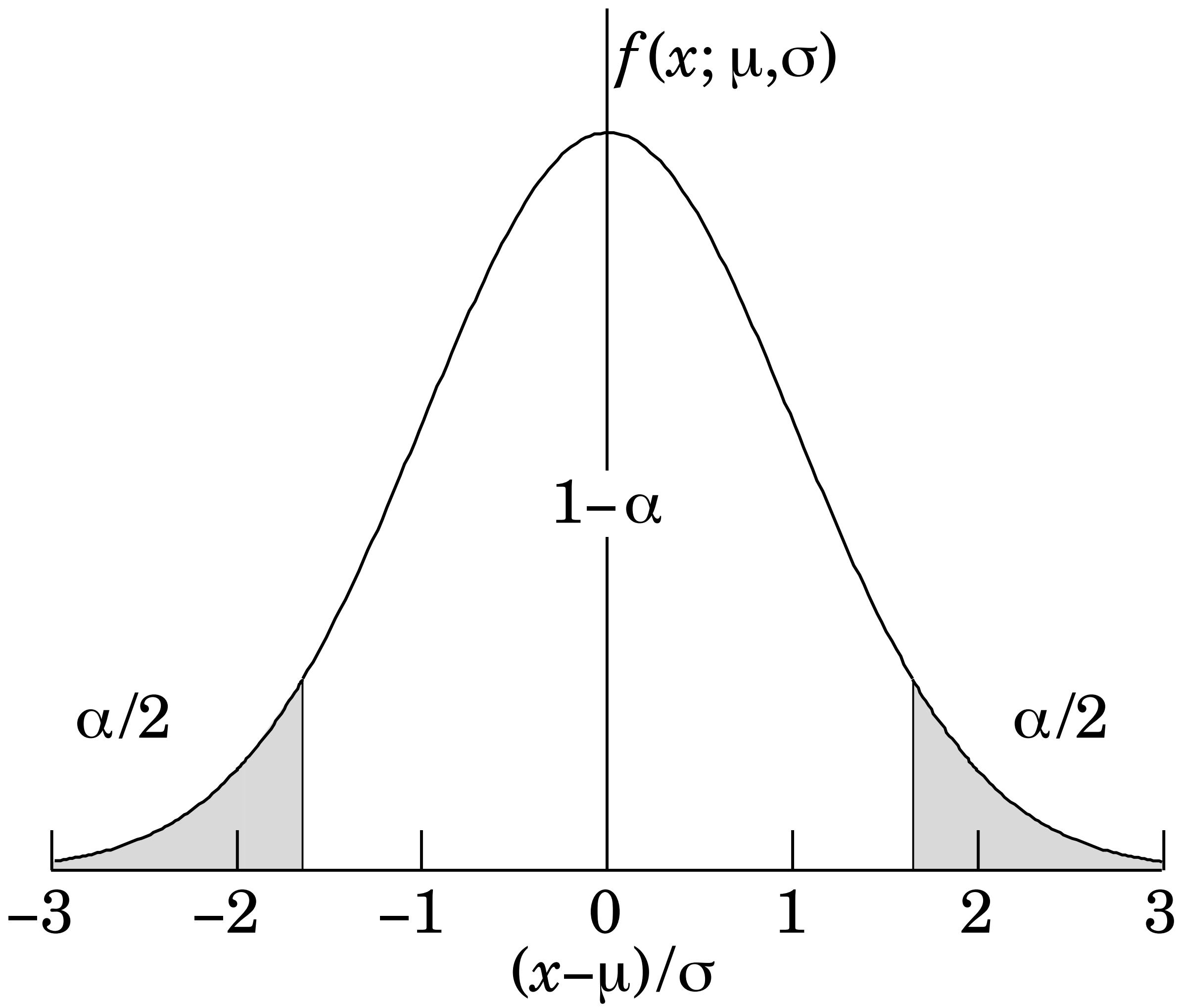
$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp(-(x-\mu)^2/2\sigma^2)$$

$$\int_{-\infty}^{x} f(x)dx = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$





Gaussian (Normal) distribution

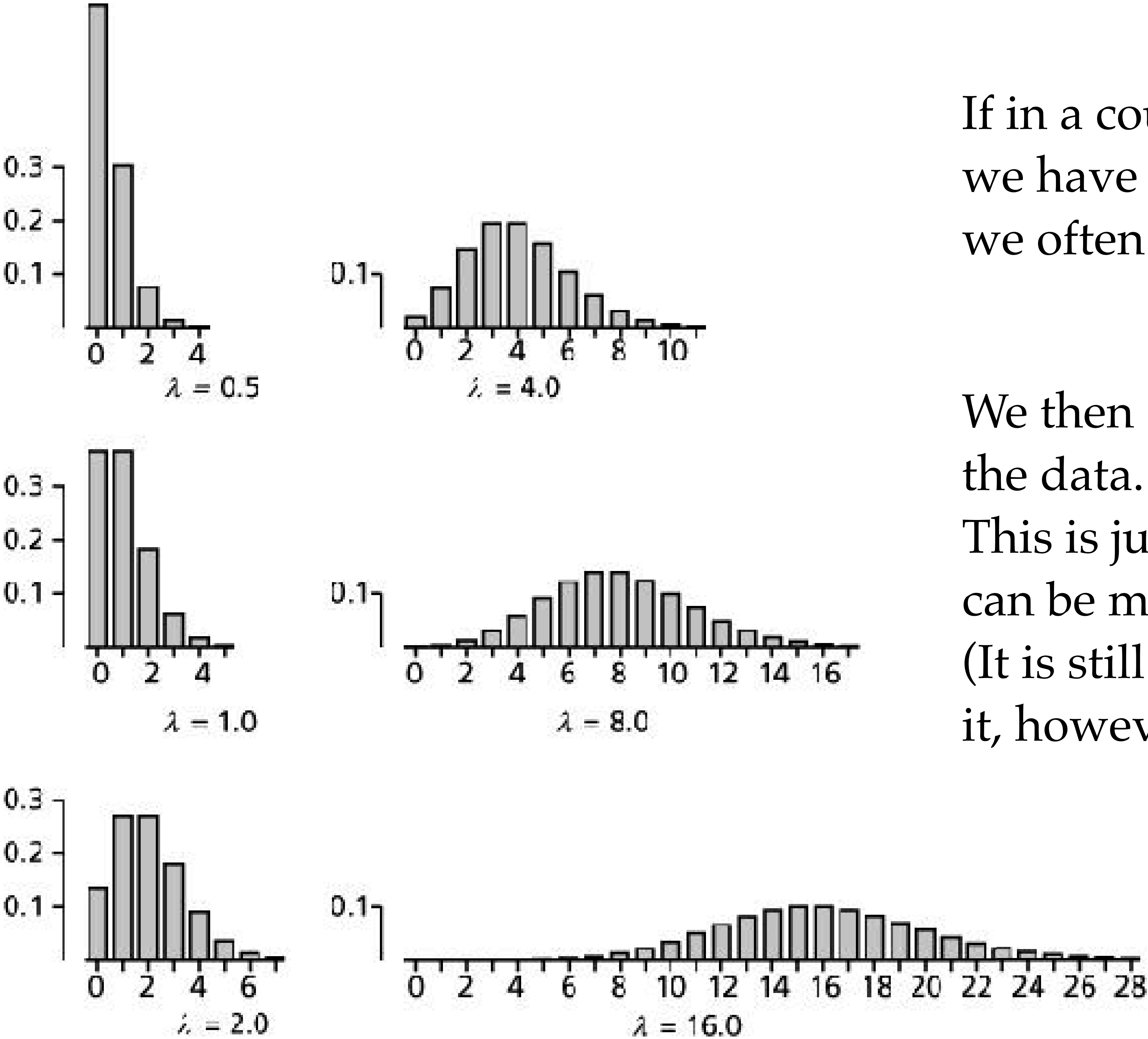


TMath::Prob(δ²,1)

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

Table 36.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

Poisson for large μ is approximately Gaussian of width $\sigma = \sqrt{\mu}$



If in a counting experiment all we have is a measurement n, we often use this to estimate μ.

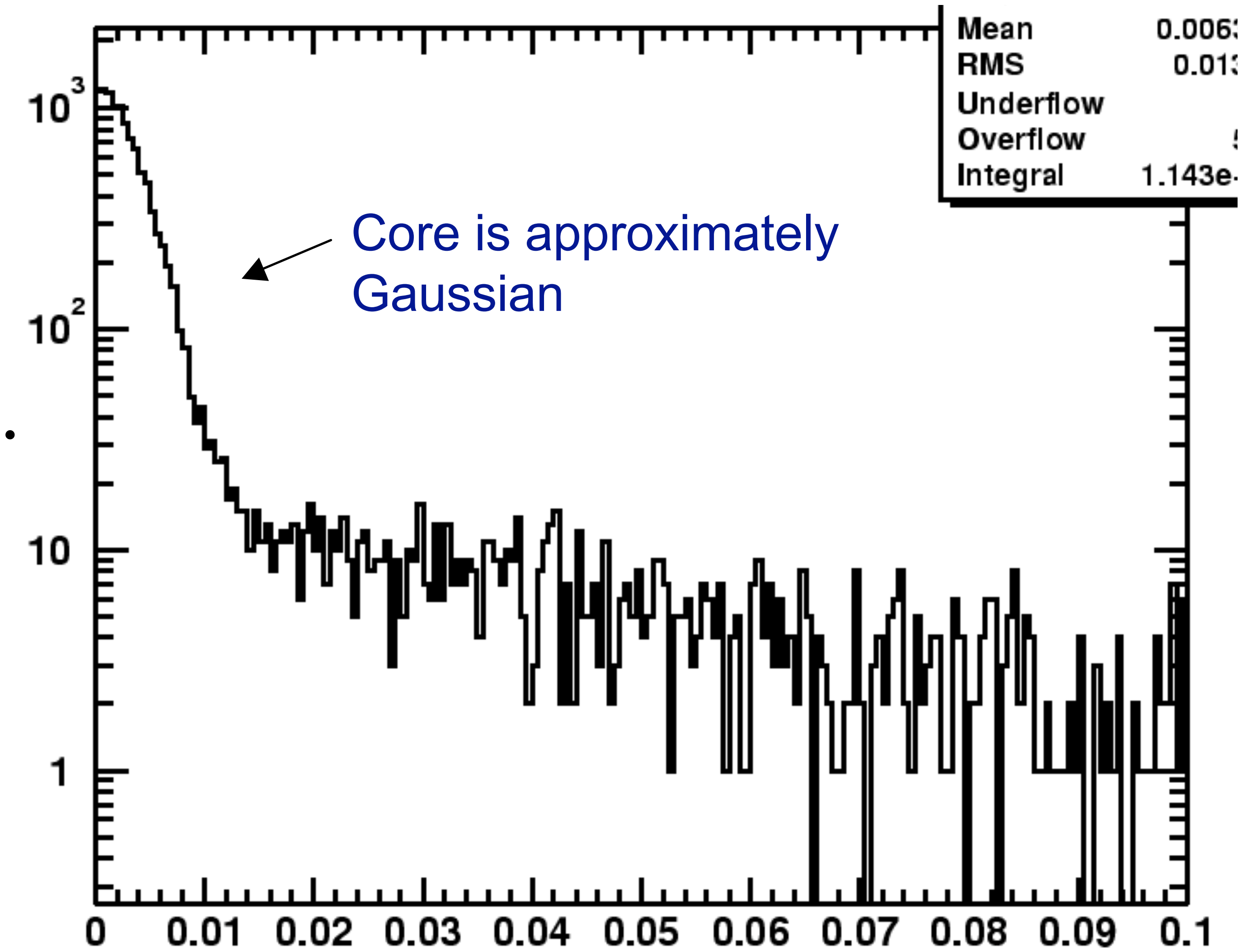
We then draw $\sqrt{n}$ error bars on the data.
This is just a convention, and can be misleading.
(It is still recommended you do it, however.)

Not all distributions are Gaussian

(Ex) track impact parameter distributions

∃ **multiple scattering**

- dominant Gaussian core
- rare large scatters, including heavy-quark decays, nuclear interactions, etc.



"All models are wrong, but some are useful." from Box & Draper (1987)

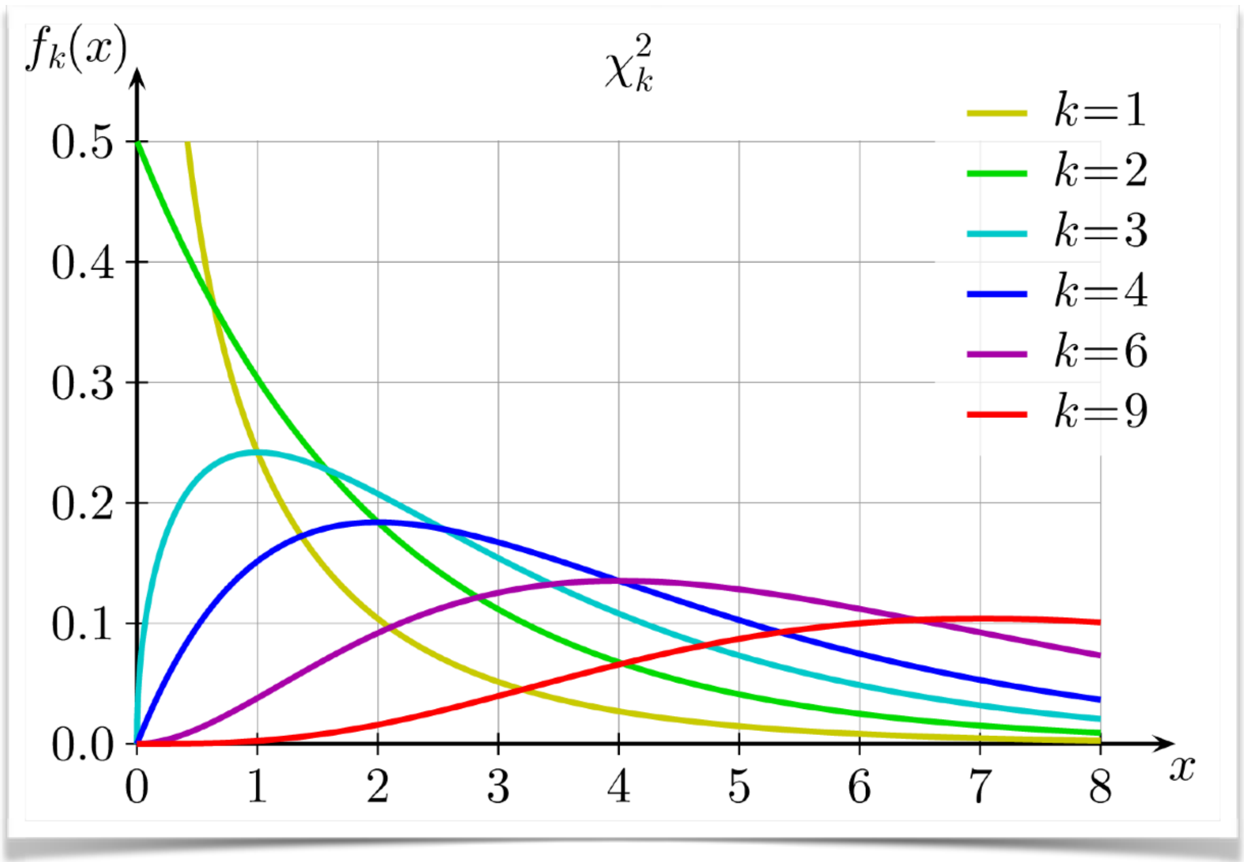
Chi-square(χ^2) distribution

The χ^2 pdf $f(z;n)$ for continuous random variable $z(\geq 0)$ with n deg. of freedom:

$$f(z;n) = \frac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)}$$

$$n = 1, 2, \cdots = \#(\text{d.o.f.})$$

$$E[z] = n, \quad V[z] = 2n$$



- For independent Gaussian r.v. $x_i (i = 1, \cdots, n)$ each with mean μ_i and variance σ_i^2, $z = \sum_{i=1}^{n}(x_i - \mu_i)^2/\sigma_i^2$ follows χ^2 pdf with n dof.
- Useful for *goodness-of-fit* test with method of least squares.

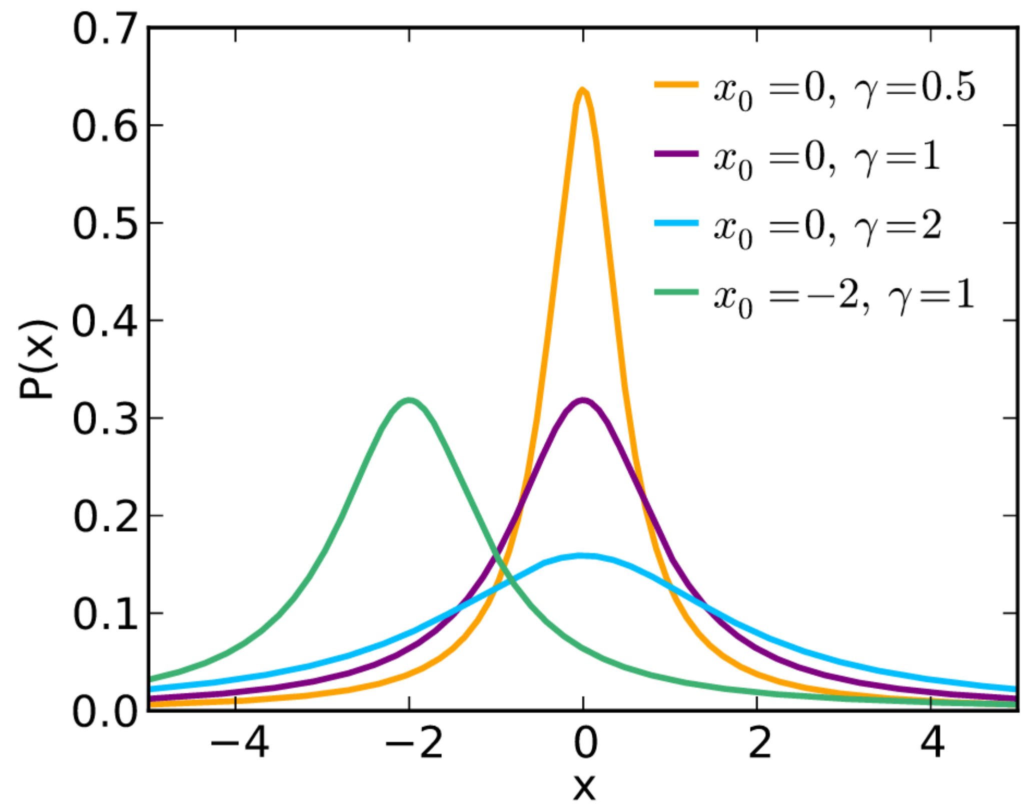
Cauchy (Breit-Wigner) distribution

$$f_{\mathrm{BW}}(x;\Gamma,x_0) = \frac{1}{\pi}\frac{\Gamma/2}{(x-x_0)^2+(\Gamma/2)^2}$$

$E(x)$, $V(x)$: not well-defined

x_0 = mode, median

Γ = full width at half-maximum



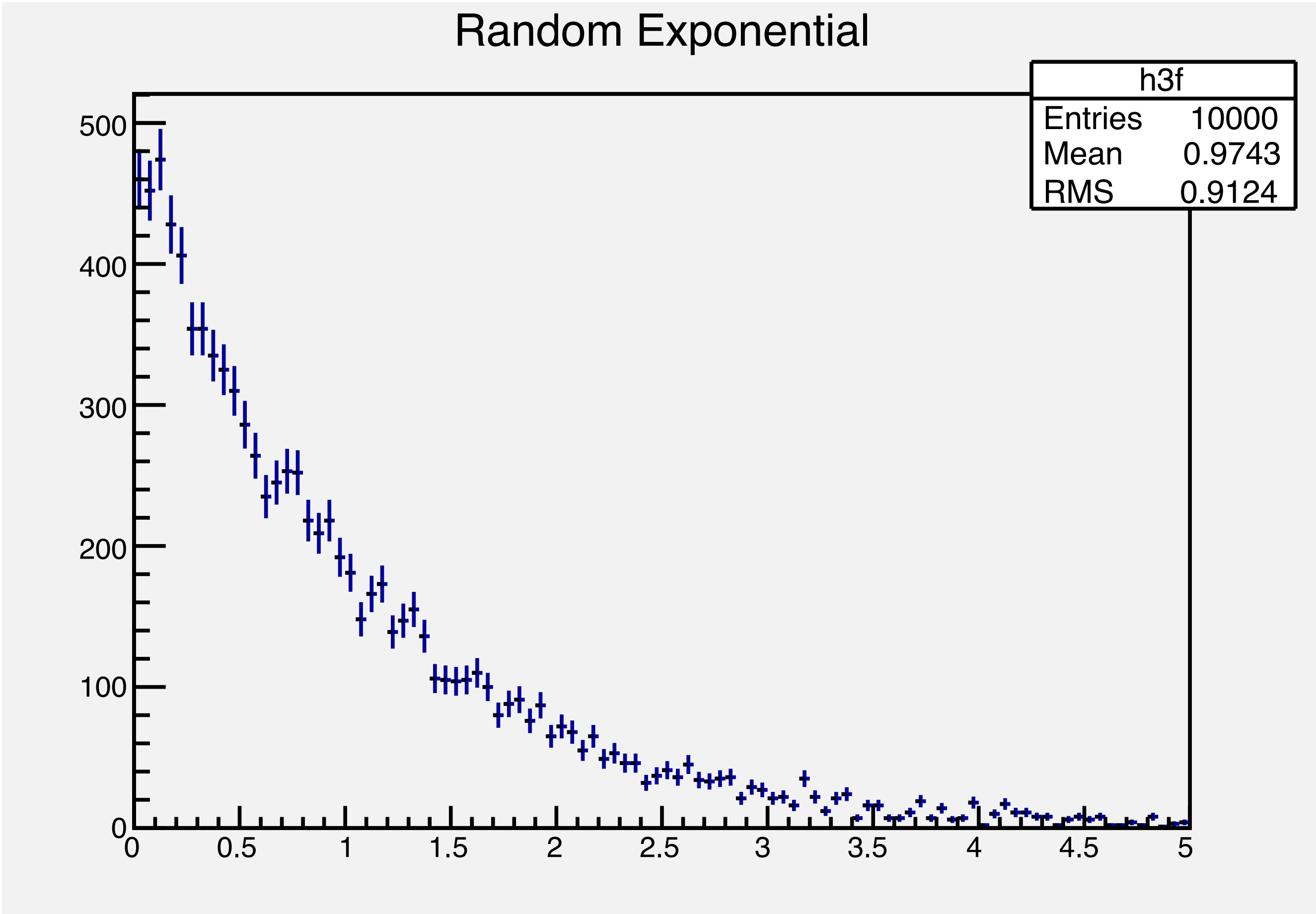
- (Ex) invariant mass distribution of strongly-decaying hadrons, e.g. $\rho,\ K^*,\ \phi$, with $\Gamma(=1/\tau)$ being the decay rate

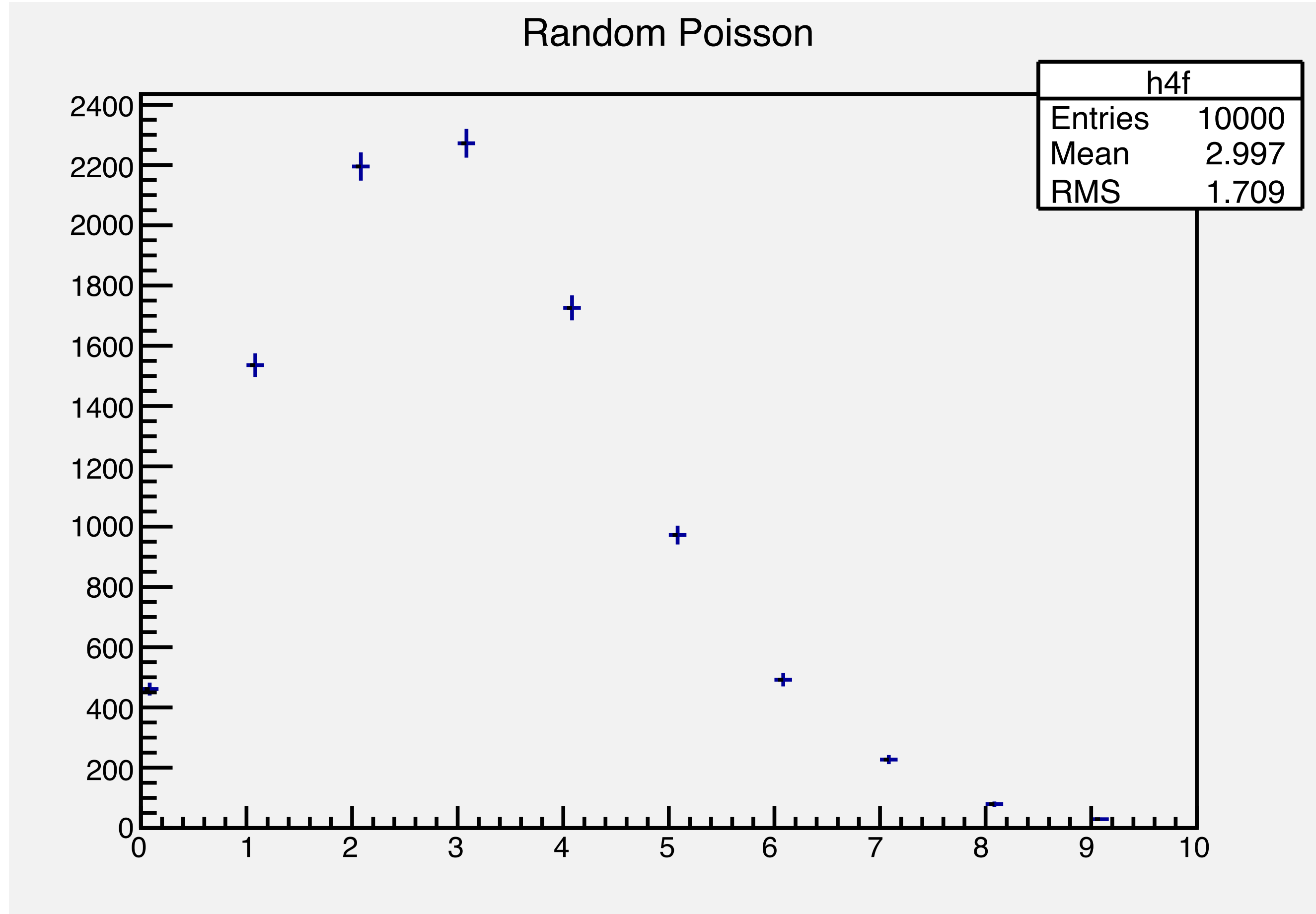
Why not make your own random variables?

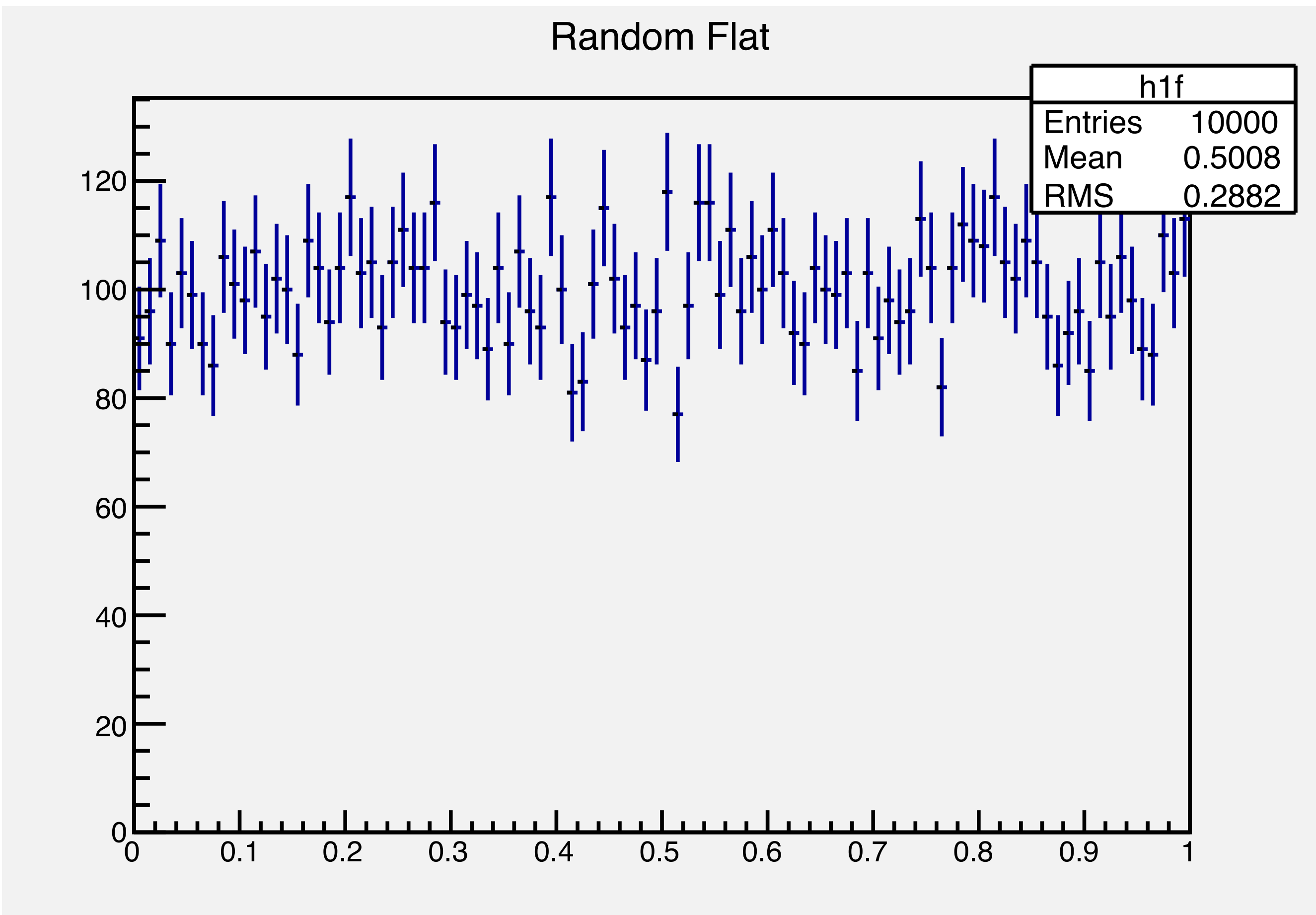
- **a free & powerful utility: ROOT http://root.cern.ch/**
- **some frequently used random variables by ROOT**
  - flat on [0,1]
  - Gaussian
  - Exponential
  - Poisson

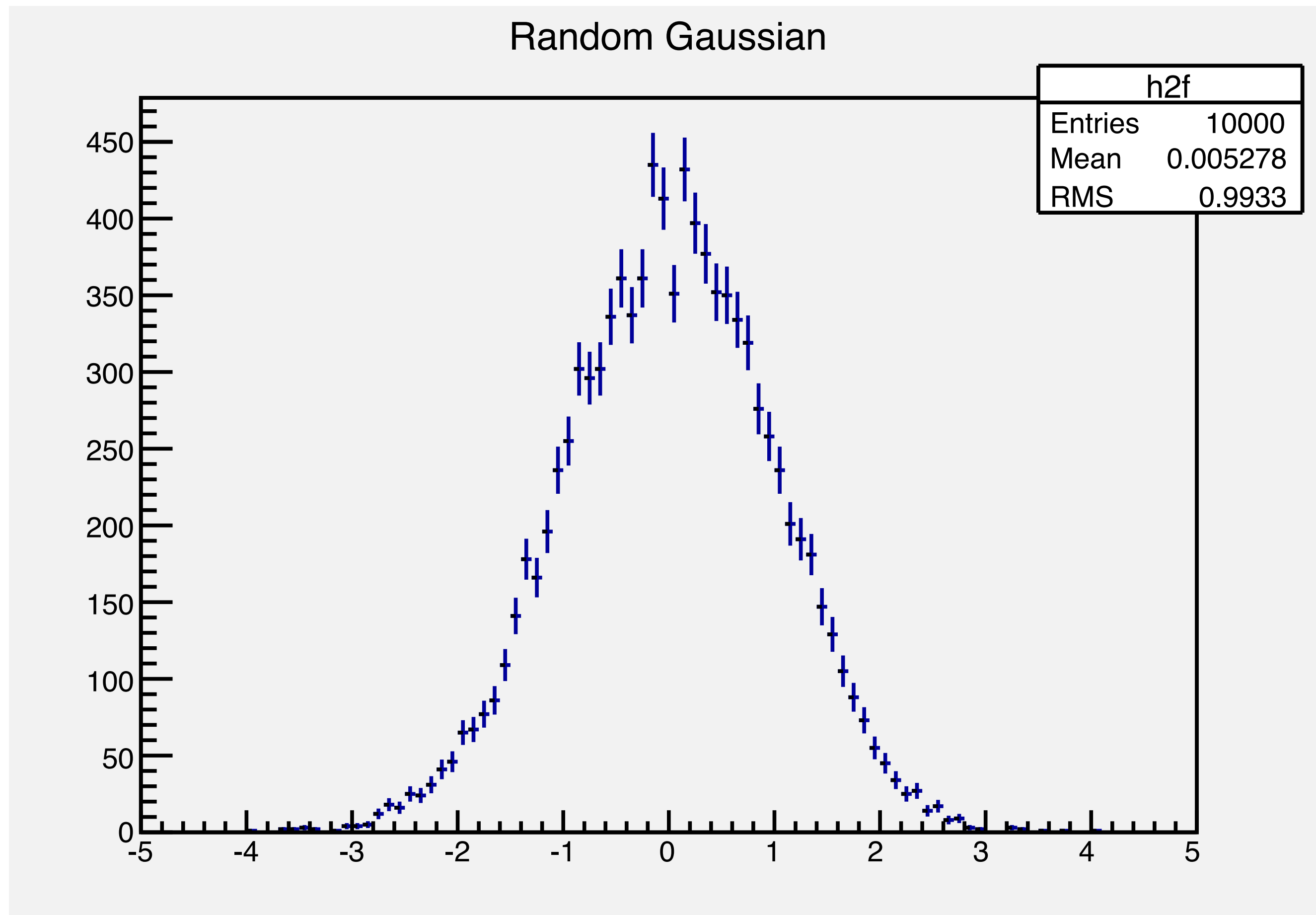
  and so on…

```
x1 = r1.Rndm();
x2 = r2.Gaus(0.0,1.0);
x3 = r3.Exp(1.0);
x4 = r4.Poisson(3.0);
```


Random Exponential
h3f
Entries 10000
Mean 0.9743
RMS 0.9124


Random Poisson
h4f
Entries 10000
Mean 2.997
RMS 1.709


Random Flat
h1f
Entries 10000
Mean 0.5008
RMS 0.2882


Random Gaussian
h2f
Entries 10000
Mean 0.005278
RMS 0.9933

some theorems, laws…

the Law of Large Numbers

- Suppose you have a sequence of indep't random variables x_i
  - with the same mean μ
  - and variances σ_i^2
  - but otherwise distributed "however"
  - the variances are not too large

$$\lim_{N\to\infty}(1/N^2)\sum_{i=1}^{N}\sigma_i^2 = 0 \tag{1}$$

Then the average $\bar{x}_N = (1/N)\sum_i x_i$ converges to the true mean μ

- (Note) What if the condition (1) is finite but non-zero?

$\Rightarrow$ the convergence is "almost certain" (*i.e.* the failures have measure zero)

In short, if you try many times, eventually you get the true mean!

the Central Limit Theorem

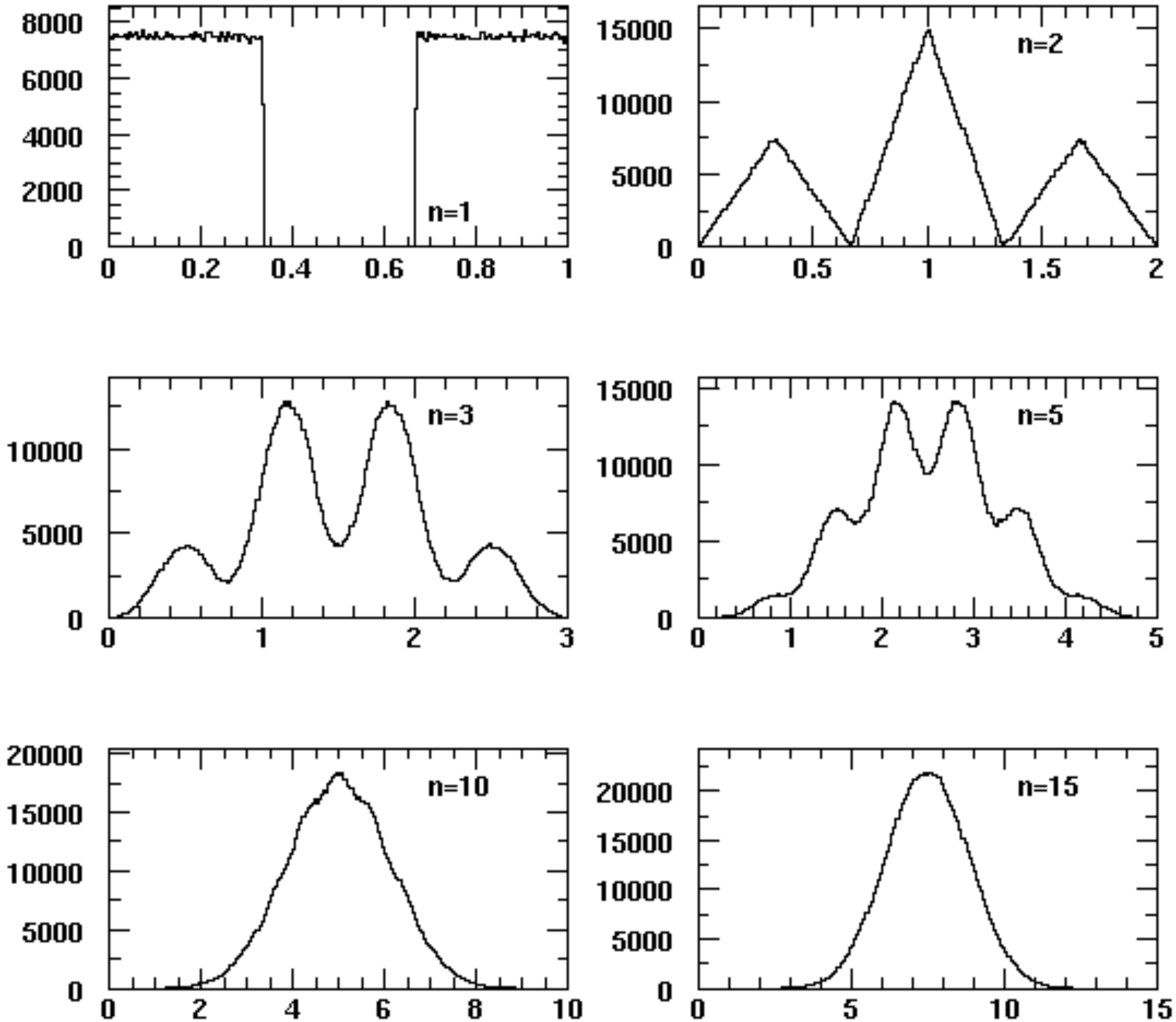
- Suppose you have a sequence of indep't random variables x_i
  - with means μ_i and variances σ_i^2
  - but otherwise distributed "however"
  - and under certain conditions on the variances

The sum $S = \sum_i x_i$ converges to a Gaussian

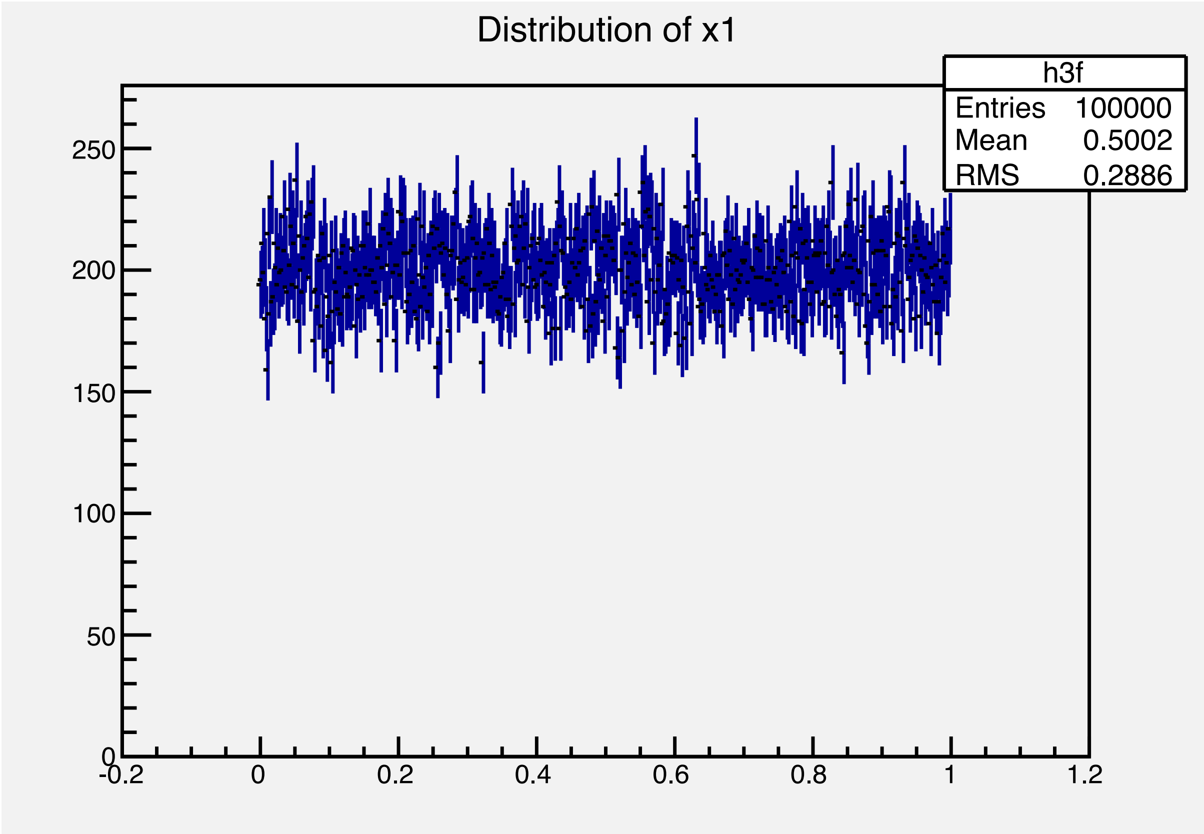
$$\lim_{N\to\infty} \frac{S - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \to \mathcal{N}(0, 1) \tag{2}$$

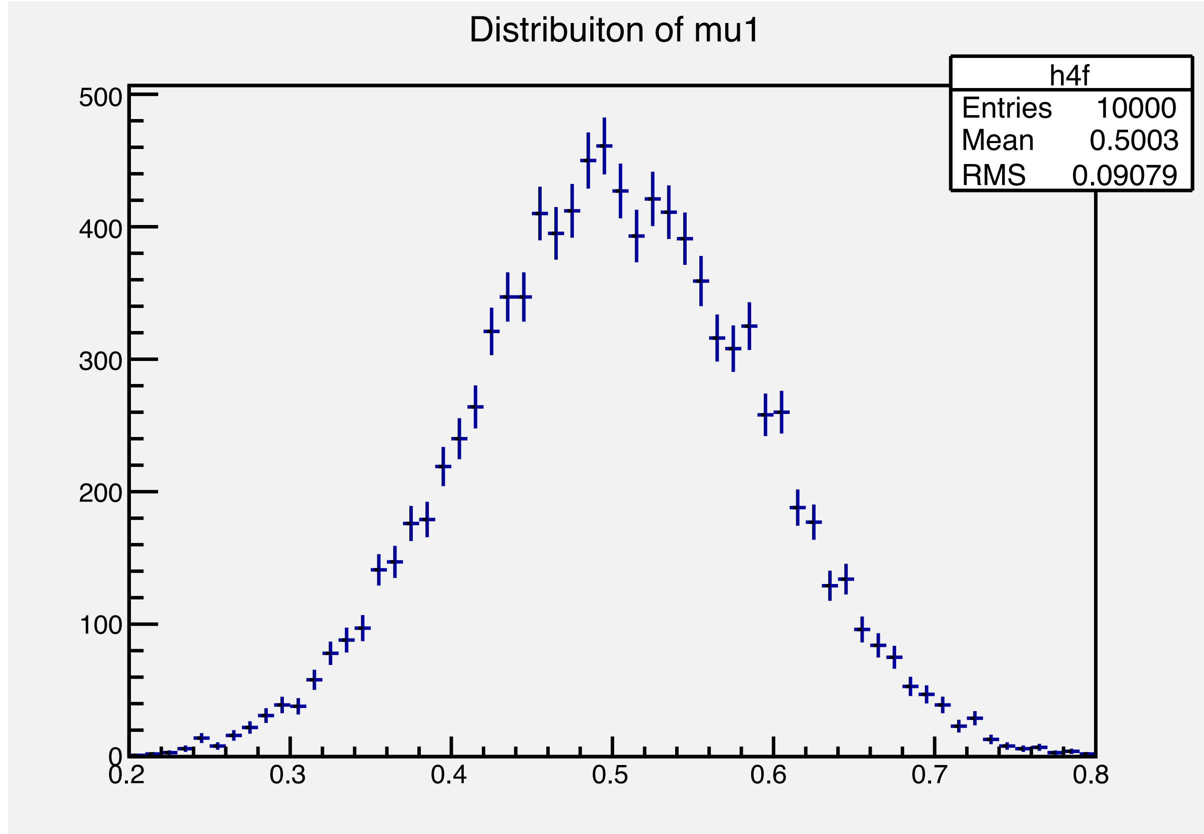
- (Note) important not to confuse LLN with CLT
  - **LLN**: with enough samples, the average $\to$ the true mean
  - **CLT**: if you put enough random numbers into your processor, the distribution of their average $\to \mathcal{N}(0, 1)$

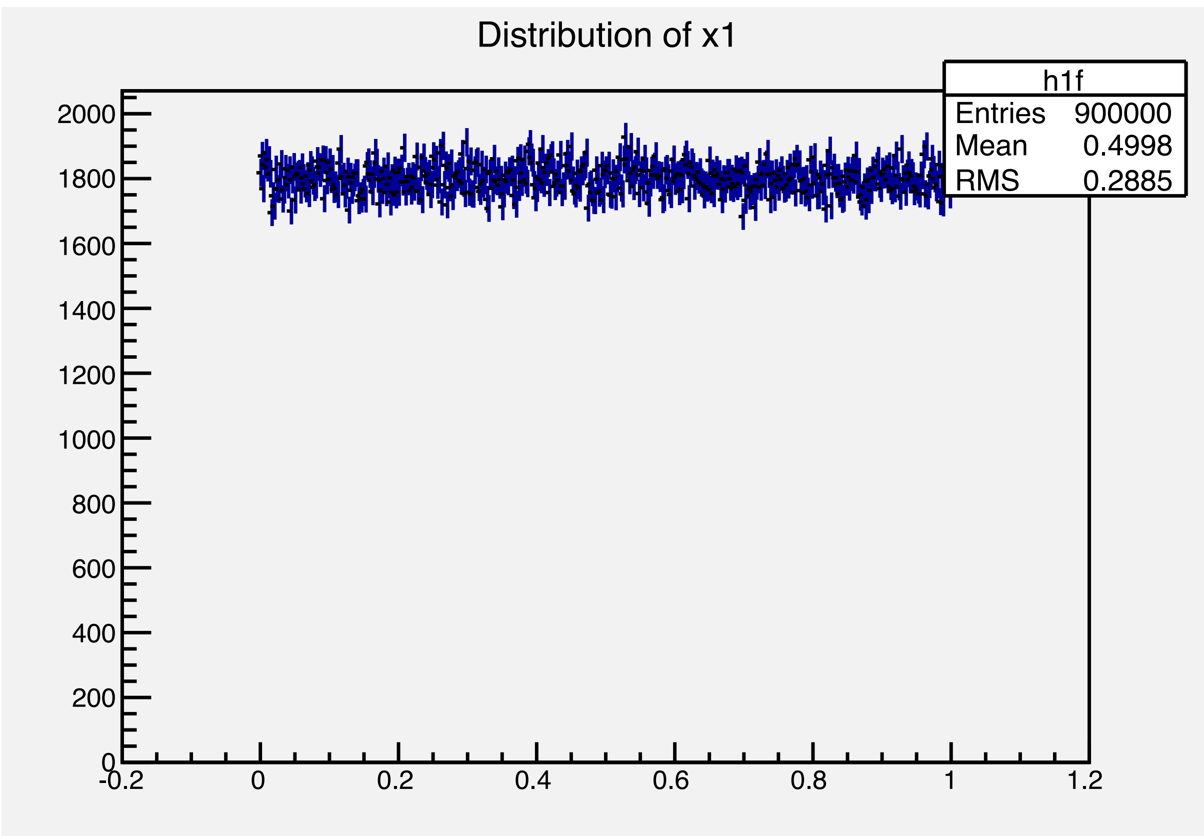
an example of the CLT at work

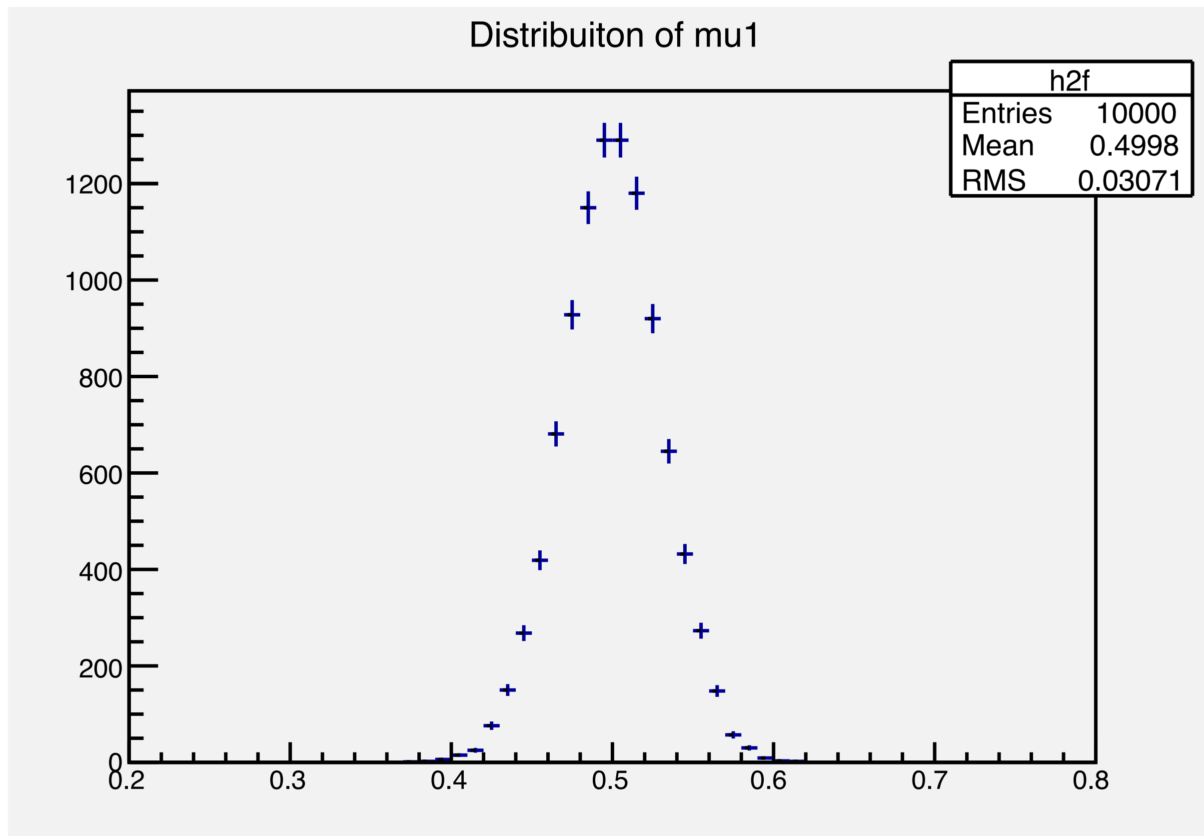


more examples of CLT at work

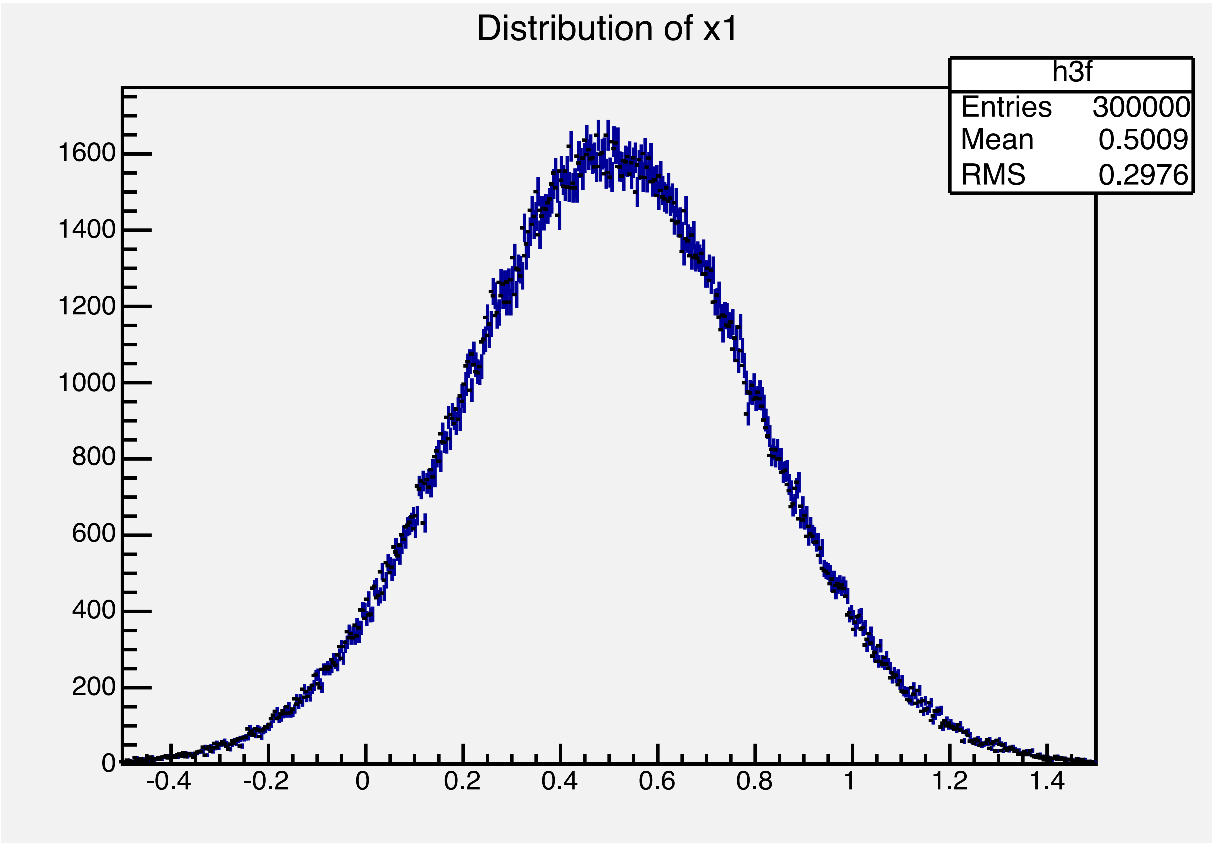


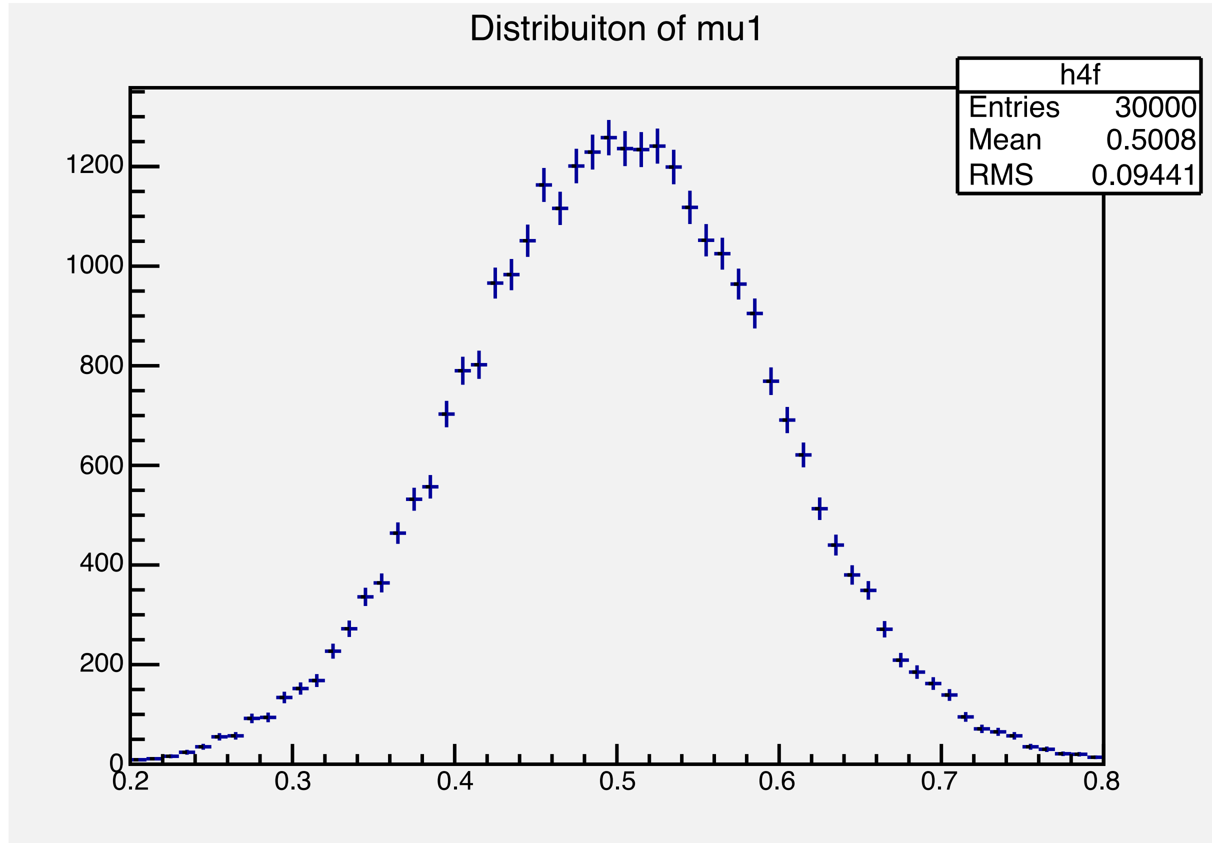


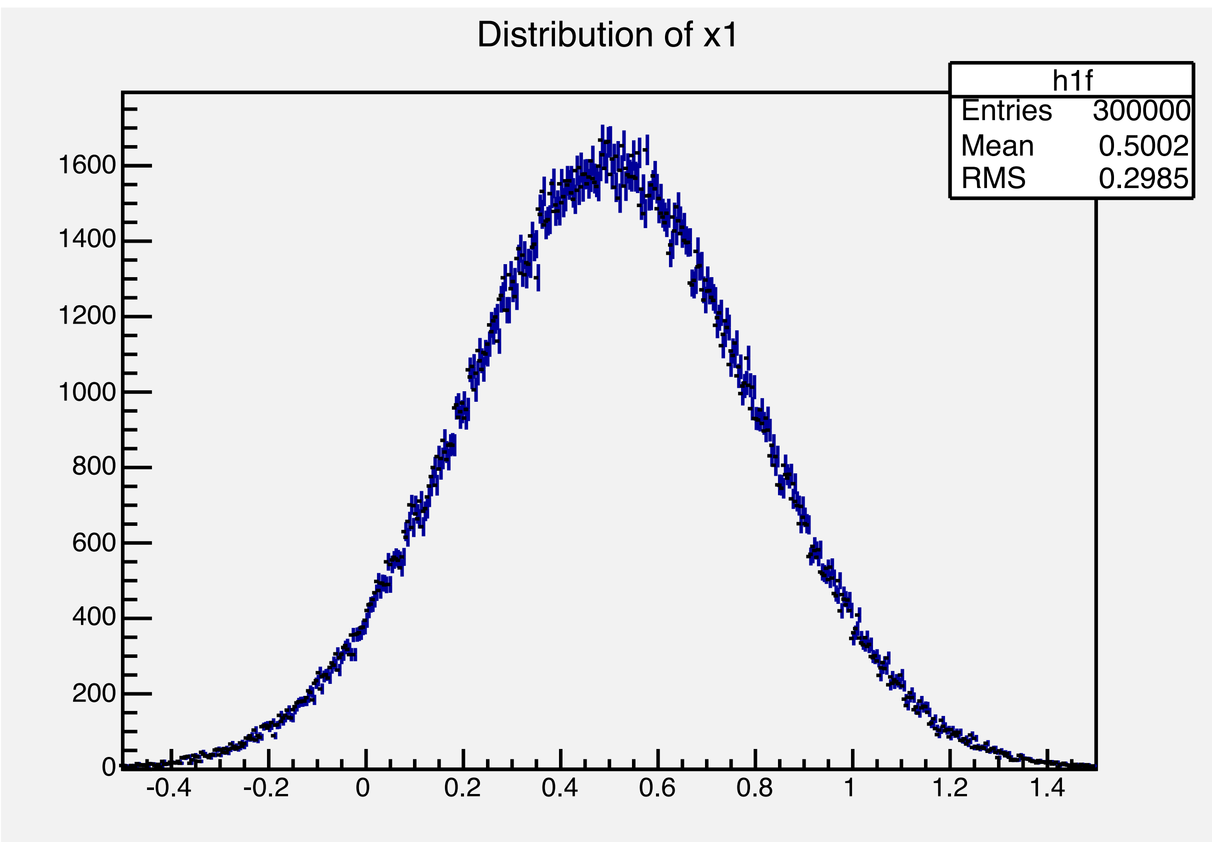


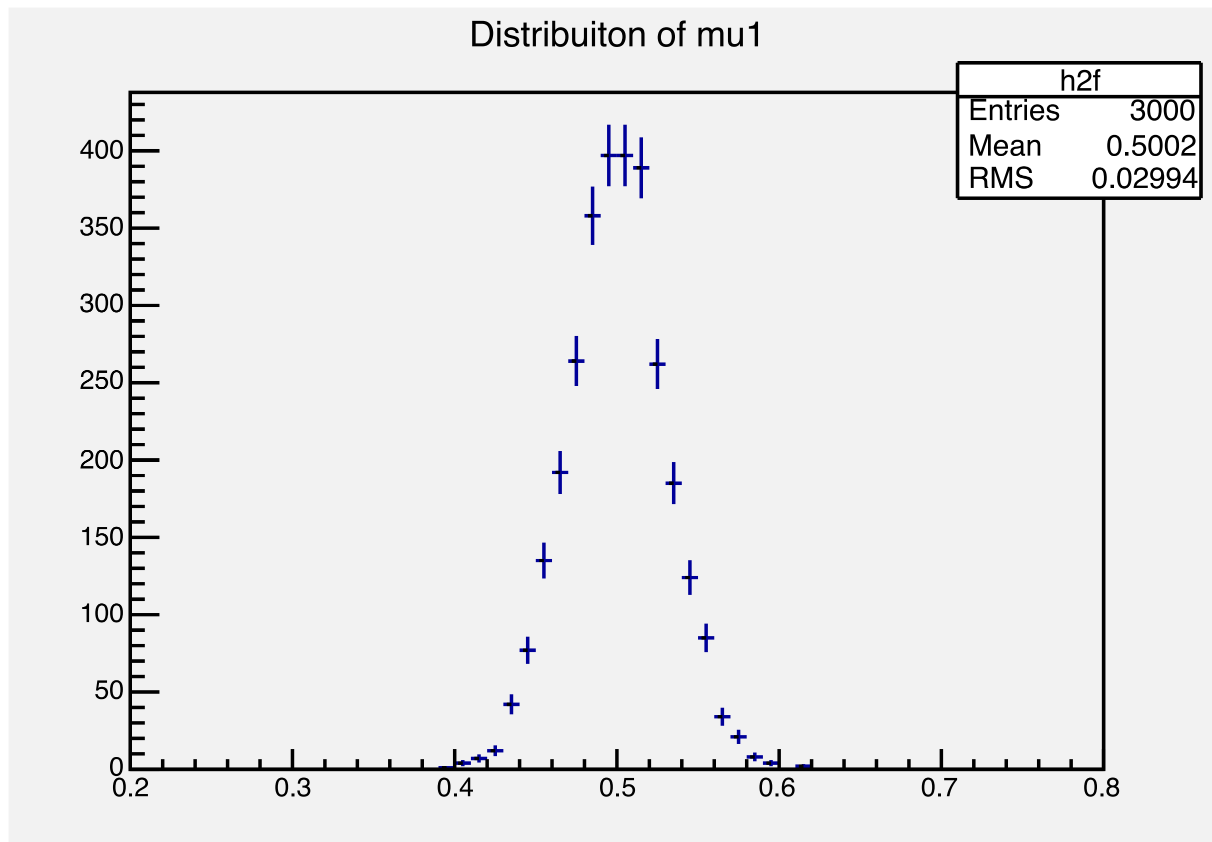


more examples of CLT at work









the Neyman-Pearson Lemma

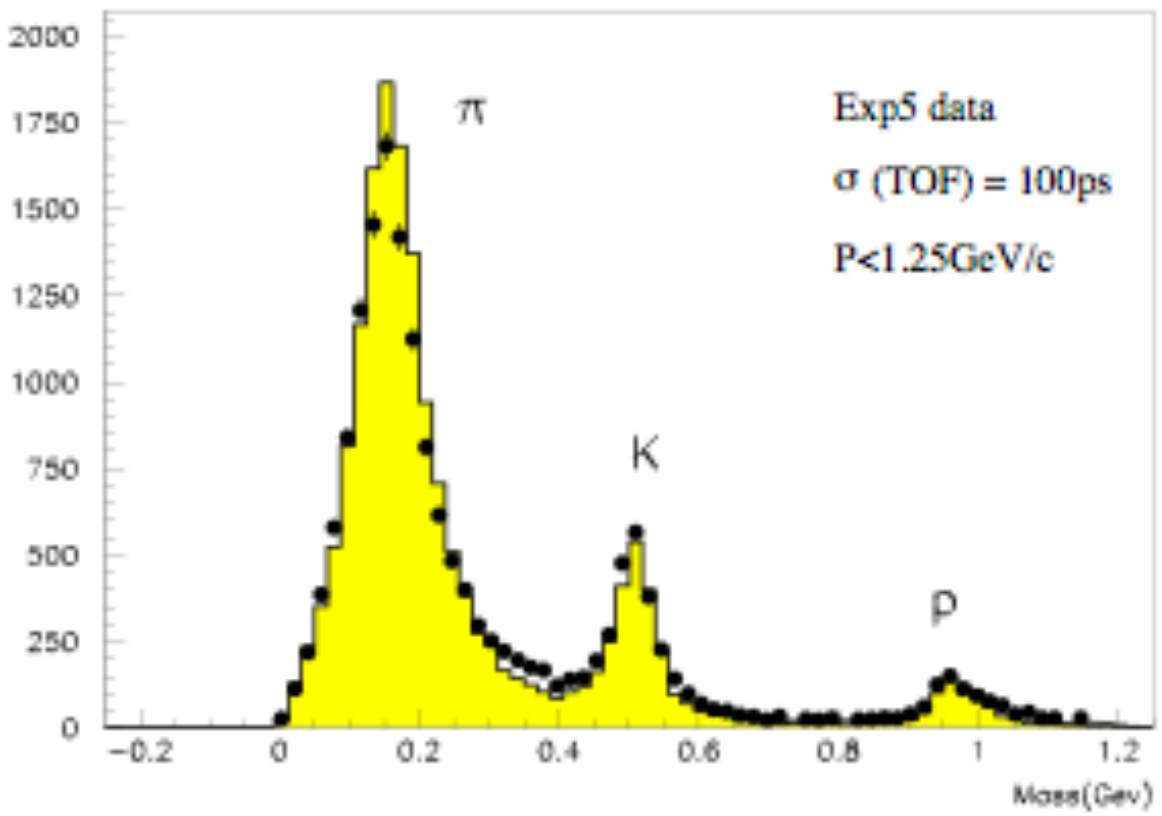
For a test of size α of the simple hypothesis H_0,
to obtain the highest power w.r.t. the simple alternative H_1,
choose the critical region w such that the likelihoot ratio satisfies

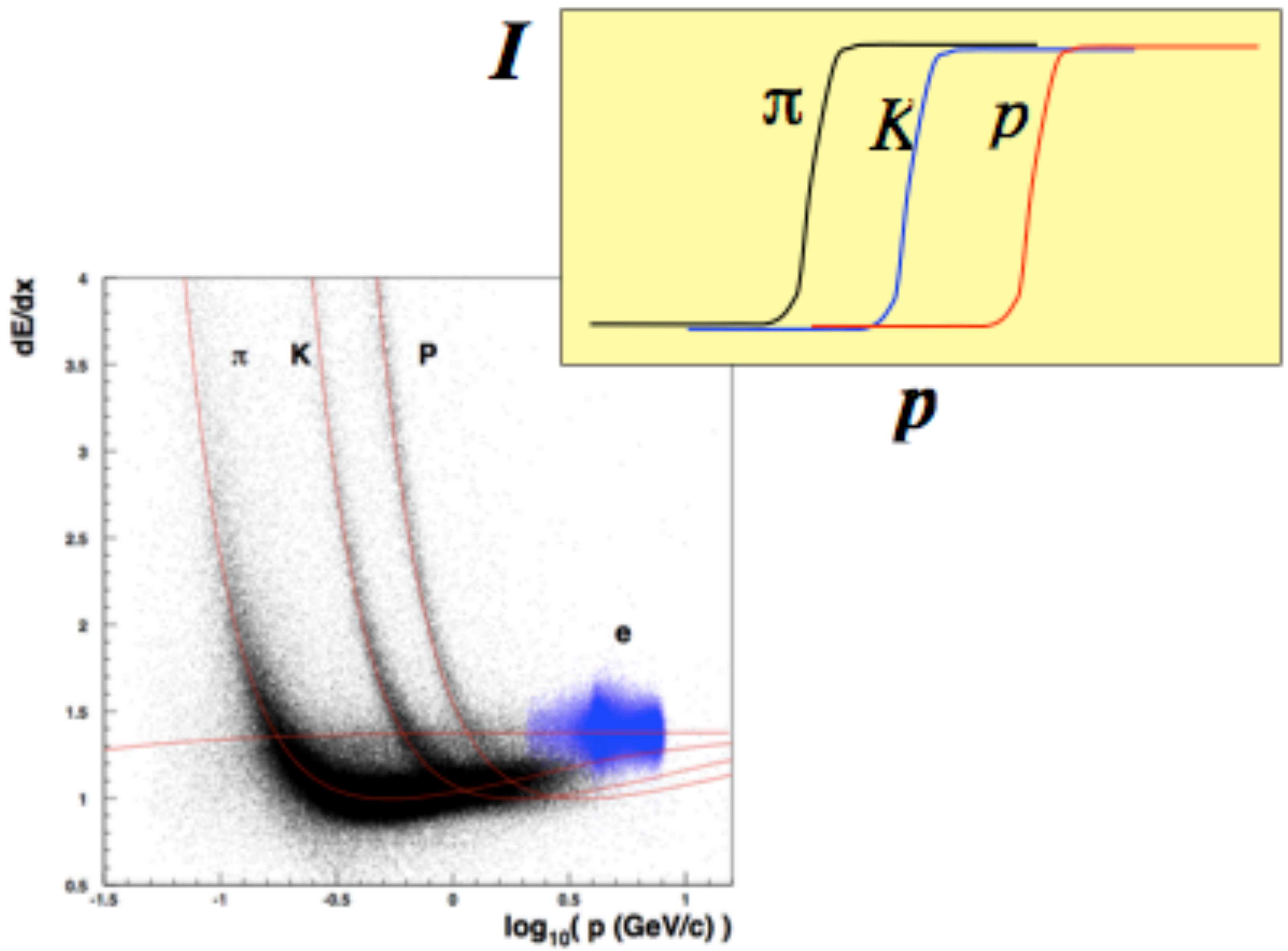
$$\frac{P(\vec{x}|H_1)}{P(\vec{x}|H_0)} \geq k$$

everywhere in w and is $< k$ elsewhere,
where k is a constant chosen for each pre-determined size α.

the Neyman-Pearson Lemma

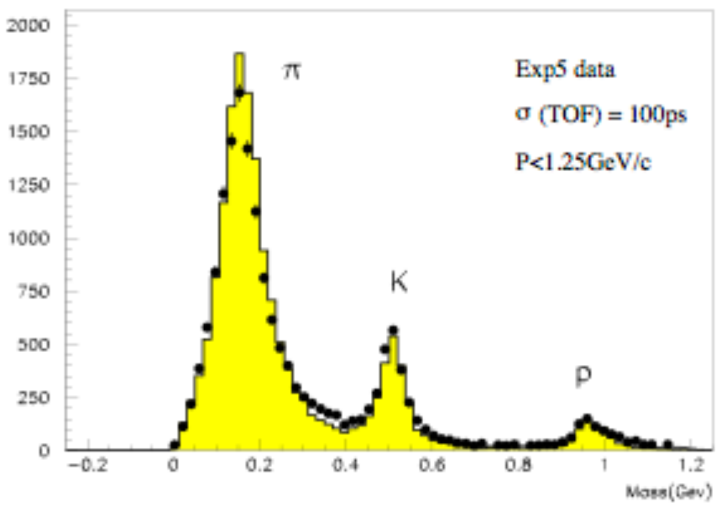
- ***Application in Belle particle ID***

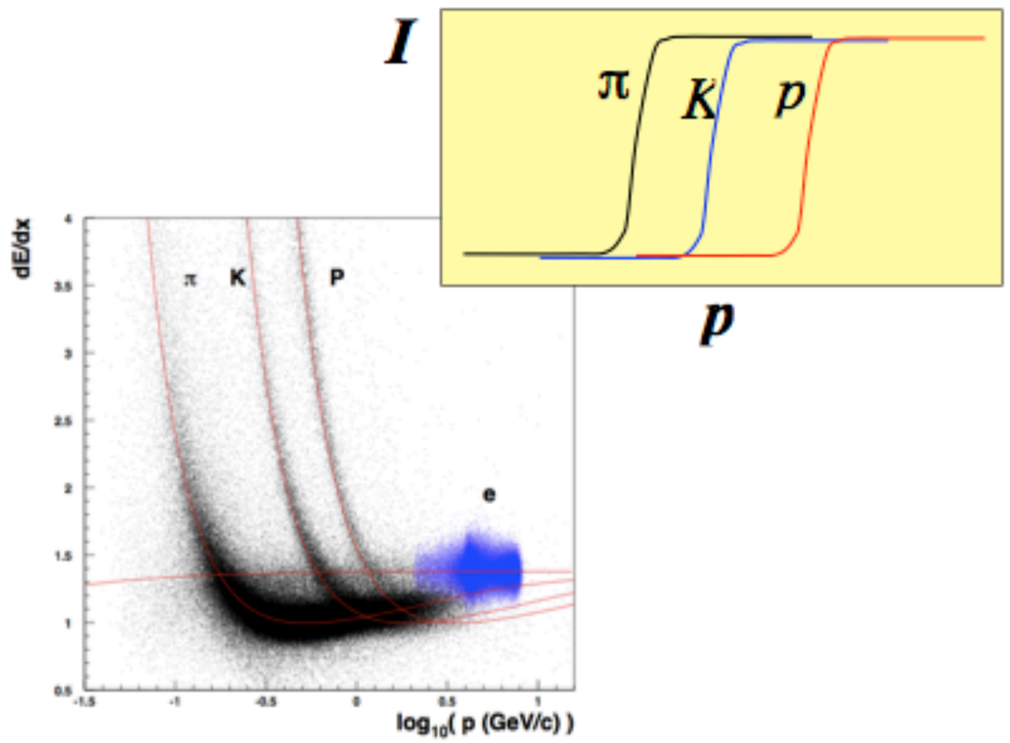




the Neyman-Pearson Lemma

Application in Belle particle ID





$$P_i \equiv P_i^{dE/dx} \times P_i^{\mathrm{TOF}} \times P_i^{\mathrm{Ch}} \quad \text{e.g. } (i = \pi \text{ or } K)$$

For optimal statistic, construct the likelihood ratio
$R_{K/\pi} = P_K / P_\pi$ (or any ftn. that is monotonic to it)

more on this lemma, later

the Wilk's theorem

- *We will encounter it later when we discuss the "likelihood ratio" ...*

THE LARGE-SAMPLE DISTRIBUTION OF THE LIKELIHOOD RATIO FOR TESTING COMPOSITE HYPOTHESES[1]

BY S. S. WILKS

By applying the principle of maximum likelihood, J. Neyman and E. S. Pearson[2] have suggested a method for obtaining functions of observations for testing what are called *composite statistical hypotheses*, or simply *composite*

...

[1] Presented to the American Mathmatical Society, March 26, 1937.

...

We can summarize in the

Theorem: If a population with a variate x is distributed according to the probability function $f(x, \theta_1, \theta_2 \cdots \theta_h)$, such that optimum estimates $\bar{\theta}_i$ of the θ_i exist which are distributed in large samples according to (3), *then when the hypothesis H is true that $\theta_i = \theta_{0i}$, $i = m + 1, m + 2, \cdots h$, the distribution of $-2 \log \lambda$, where λ is given by* (2) *is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h - m$ degrees of freedom.*

the Wilk's theorem

http://wwwusers.ts.infn.it/~milotti/Didattica/StatisticaAvanzata/Cowan_2013.pdf

Suppose we model the data $\vec{X}$ with a likelihood $L(\vec{\mu})$ that depends on a set of N parameters $\vec{\mu} = (\mu_1, \cdots, \mu_N)$. (For simplicity, let's just consider a single parameter μ.)

- Define the statistic $t_\mu = -2\ln[L(\mu)/L(\hat{\mu})]$, where $\hat{\mu}$ is the ML estimator.
- The value of t_μ is a measure of how well the hypothesized parameter μ stand in agreement with the observed data.
- Larger values of t_μ indicate increasing incompatibility between the data and the hypothesized μ.
- According to Wilk's theorem, if the parameter value μ is true, then in the asymptotic limit of a large data sample, the PDF of t_μ is a χ^2 distribution for N d.o.f.

$$f(t_\mu|\mu) \sim \chi^2_N$$

A TALE OF TWO STATISTICS …

Frequentist vs. Bayesian

"*Bayes and Frequentism: a particle physicist's perspective*"
by Louis Lyons, arXiv:1301.1273

Two approaches

Relative frequency

A, B, ... are outcomes of a repeatable experiment **Frequentist**

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A}{n}$$

Subjective probability

A, B, ... are hypotheses (statements that are true or false) **Bayesian**

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Frequentist approach is, in general, easy to understand, but some HEP phenomena are best expressed by subjective prob., e.g. systematic uncertainties, prob(Higgs boson exists), ...

Bayes' theorem

- **From the definition of conditional prob., we have**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

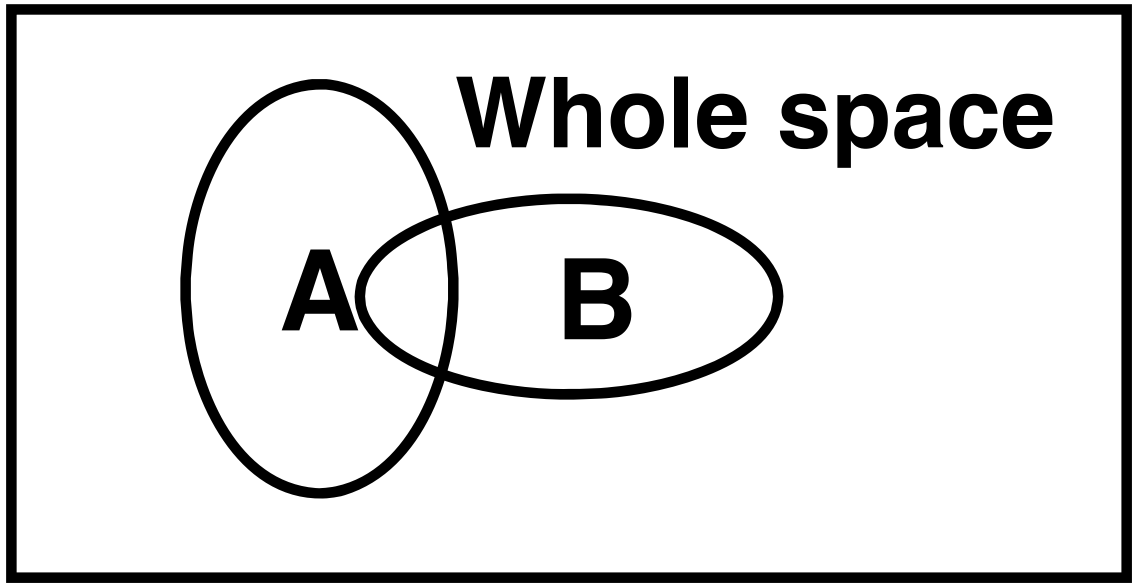
- but $P(A \cap B) = P(B \cap A)$
- therefore,

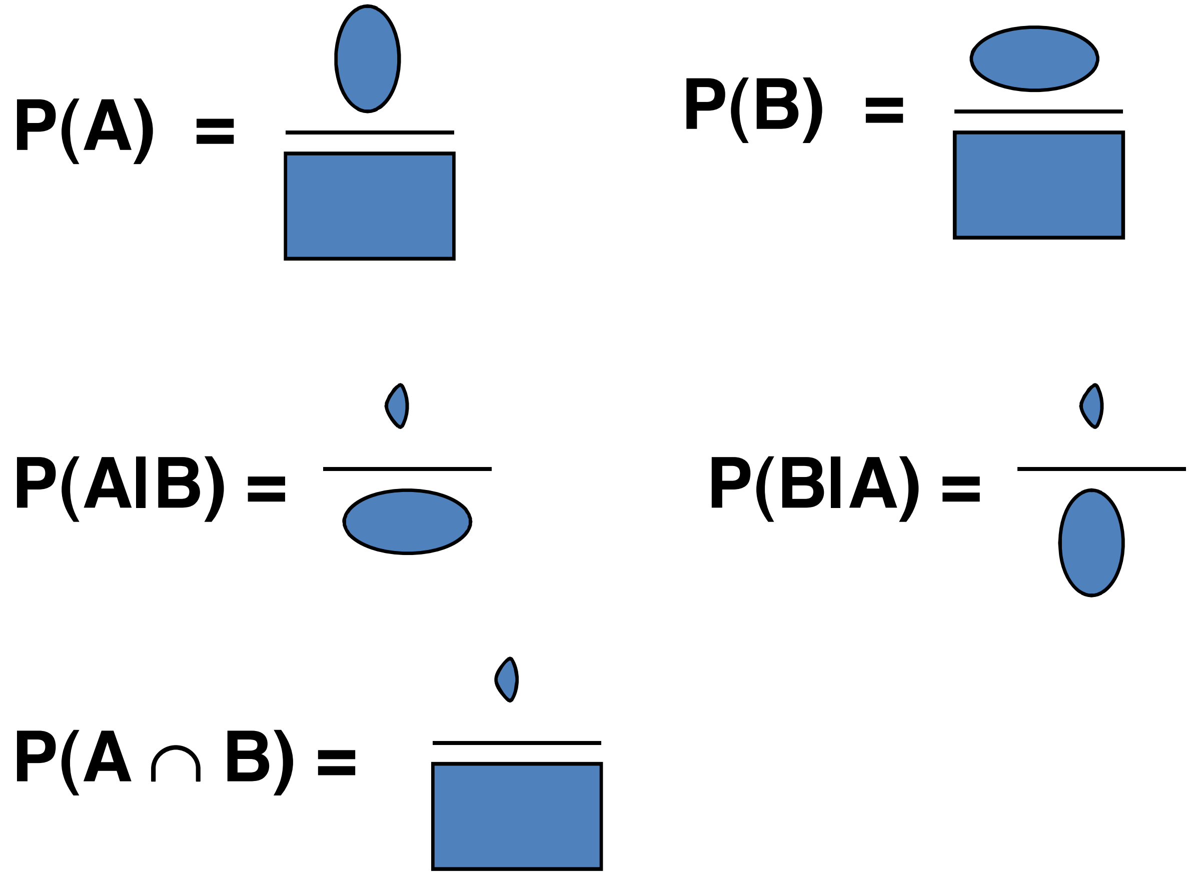
$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

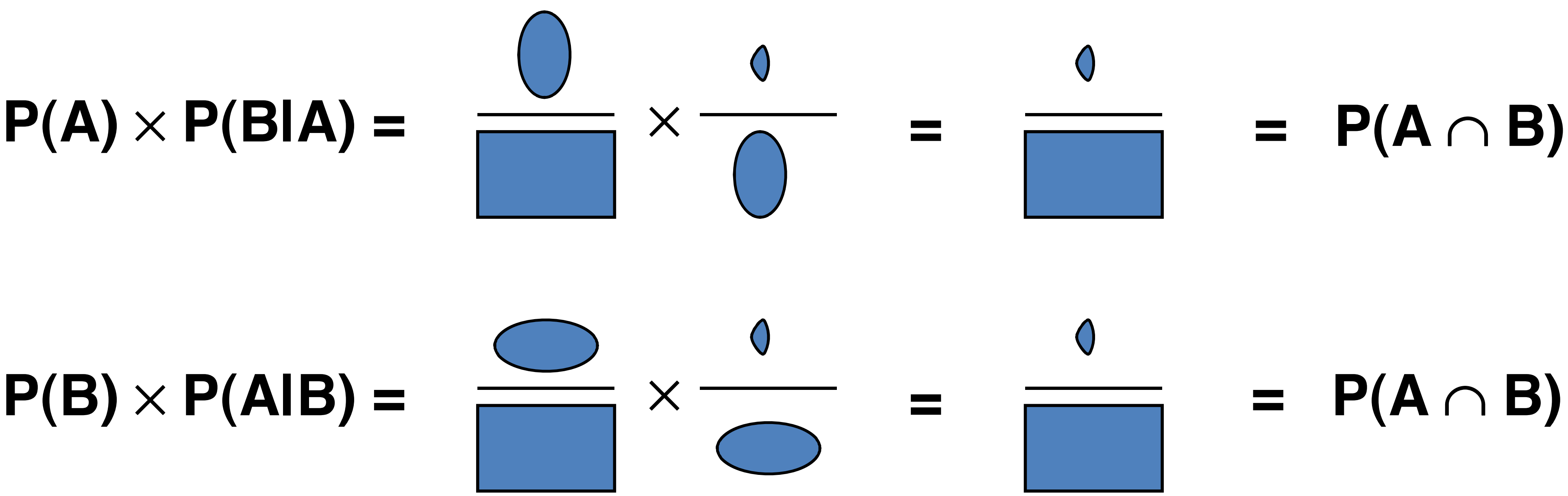
- First published (posthumous) by Rev. Thomas Bayes (1702-1761)

An essay towards solving a problem in the doctrine of chances,
Phil. Trans. R. Soc. 53 (1763) 370.

P, Conditional P, and Derivation of Bayes' Theorem in Pictures







Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Frequentist statistics – general philosophy

- In frequentist statistics, probabilities such as

  $P(\text{SUSY does exist})$

  $P(0.117 < \alpha_s < 0.121)$

  are either 0 or 1, but we don't have the answer

Bayesian statistics – general philosophy

- In Bayesian statistics, interpretation of probability is extended to the **degree of belief** (*i.e.* subjective).
- suitable for **hypothesis testing** (but no golden rule for priors)

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

- can also provide more natural handling of non-repeatable things: *e.g.* systematic uncertainties, P(Higgs boson exists)

(Ex) Bayesian answer for coin toss

Suppose I stand to win or lose money in a single coin-toss. My companion gives me a coin to use for the game.

- Do I trust the coin? What is $P(\text{faircoin})$?
- Frequentist answer:
  - toss the coin n times
  - $P(\text{heads}) = \lim_{n\to\infty}(n_{\mathrm{H}}/n)$
  - make a complicated statement about the results, which is *only indirectly* about whether the coin is fair ...
- But I can only test the coin with five throws:
  - What if I get 4H, 1T?
  - Do I trust the coin, or claim that the game is unfair?
- What about Bayesian answer?

(Ex) Bayesian answer for coin toss

Priors: a 'bad' coin has a 75% probability to show 'head'
for a 'fair' coin, it's 50%

$$P(\text{fair}\,|\,\text{BG}) = 0.50$$
$$P(\text{bad}\,|\,\text{BG}) = 0.50$$

Likelihoods:

$$P(4\text{H}, 1\text{T}|\text{fair}) = 0.1563$$
$$P(4\text{H}, 1\text{T}|\text{bad}) = 0.3955$$

Posterior:

$$P(\text{fair}|4\text{H}, 1\text{T}, \text{BG}) = \frac{P(4\text{H}, 1\text{T}|\text{fair}) \cdot P(\text{fair}|\text{BG})}{\sum_i P(4\text{H}, 1\text{T}|i) \cdot P(i|\text{BG})}$$
$$= \frac{0.1563 \cdot 0.50}{0.1563 \cdot 0.50 + 0.3955 \cdot 0.50} = 0.283$$

(Ex) Bayesian answer for coin toss

Priors: a 'bad' coin has a 75% probability to show 'head'
for a 'fair' coin, it's 50%

$$P(\text{fair}\,|\,\text{GG}) = 0.95$$
$$P(\text{bad}\,|\,\text{GG}) = 0.05$$

Likelihoods:

$$P(4\text{H}, 1\text{T}|\text{fair}) = 0.1563$$
$$P(4\text{H}, 1\text{T}|\text{bad}) = 0.3955$$

Posterior:

$$P(\text{fair}|4\text{H}, 1\text{T}, \text{GG}) = \frac{P(4\text{H}, 1\text{T}|\text{fair}) \cdot P(\text{fair}|\text{GG})}{\sum_i P(4\text{H}, 1\text{T}|i) \cdot P(i|\text{GG})}$$

$$= 0.88$$

Frequentist or Bayesian?

- While the classic or frequentist approach can lead to a well-defined probability for a given situation, it is not always usable.

  ➔ In such circumstances one is left with only one option: ***Bayesian***.

- When data are scarce ➔ these two approaches can give somewhat different predictions,

  but given sufficiently large data sample, they give pretty much the same conclusion. In that case the choice between the two may be regarded arbitrary.

- Perhaps, we may choose one for the main result, and try the other for a cross-check.