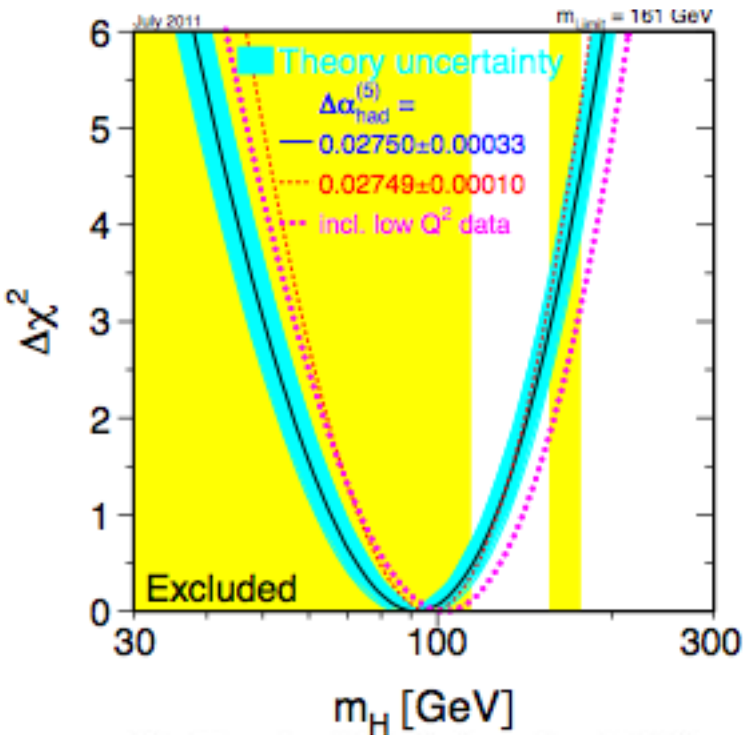
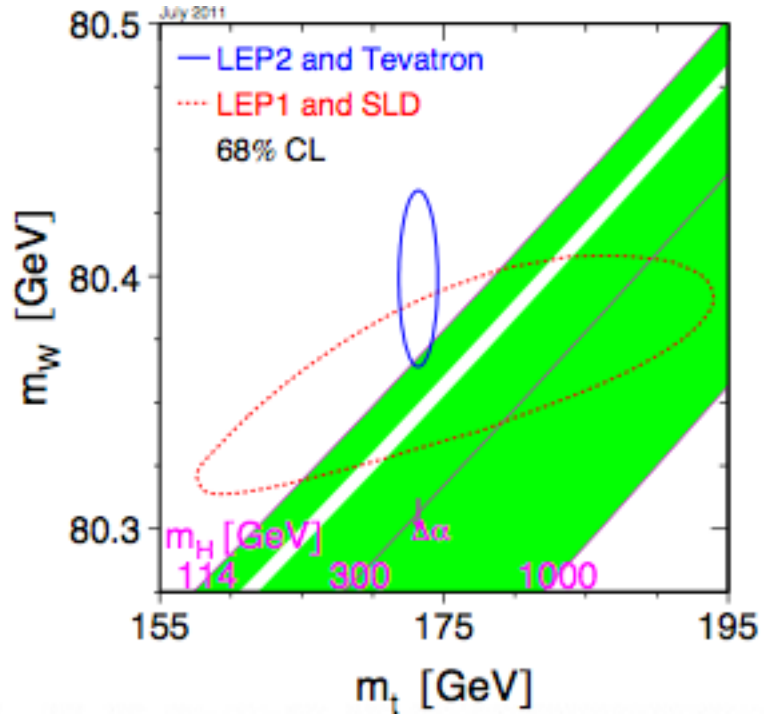
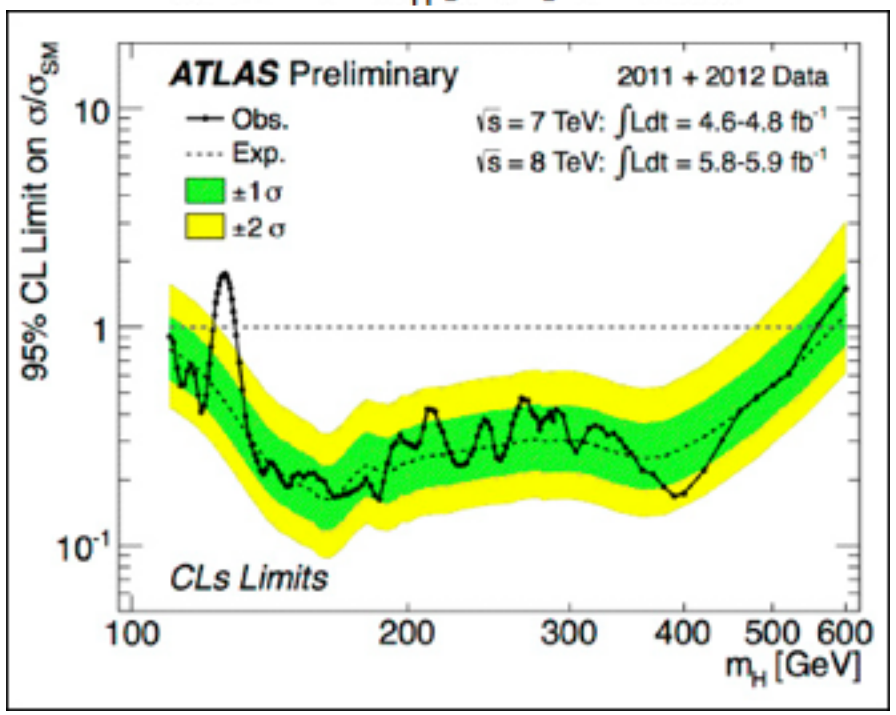
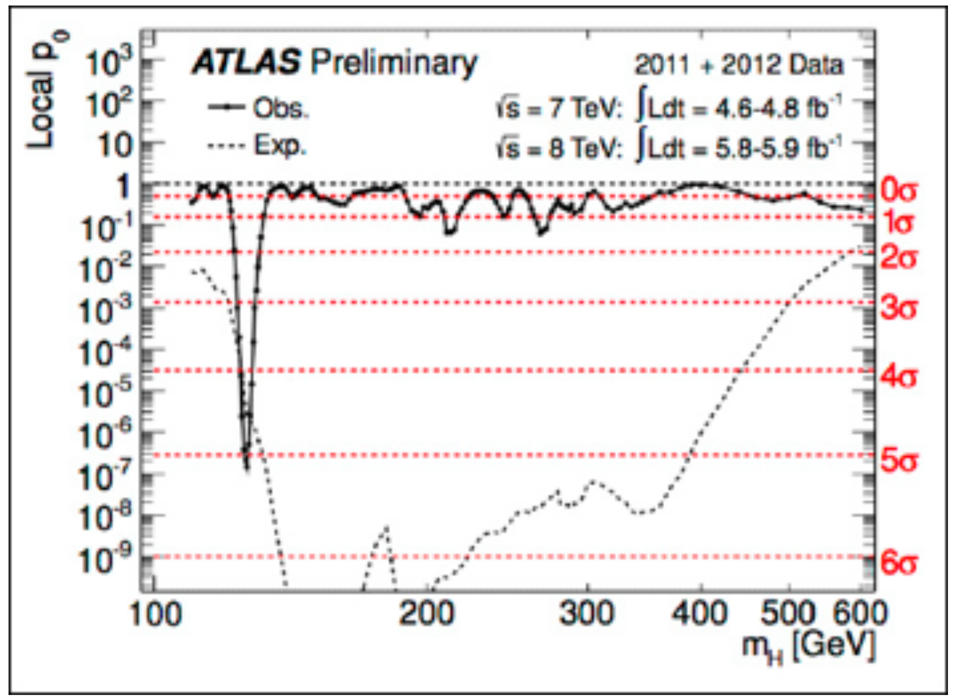








Basic statistics for HEP analyses

Youngjoon Kwon (Yonsei U.)

Aug. 11, 2014 Lecture at Sungkyunkwan U.

apologies

- **freely taking from other people's lecture slides, w/o properly citing the references**
  - just a rough list (from which I composed this lecture) is given
- **not paying attention to any mathematical rigor at all**
- **It will be simply impossible to cover "everything" even with the extended time of 180 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.**

References (*very rough*)

- Glen Cowan @ CERN lectures, July 2011

  http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

- Tom Junk @ TRIUMF, July 2009
- S. T'Jampens @ FAPPS '09, Oct. 2009
- mini-reviews on Probability & Statistics in RPP (PDG)

  http://pdg.lbl.gov/2013/reviews/rpp2013-rev-statistics.pdf

- ...

Outline

- **Basic elements**
    - some vocabulary
    - Probability axioms
    - some probability distributions
- **Two approaches: Freq. vs. Bayesian**
- **Hypothesis testing**
- **Parameter estimation**
- **Other subjects — “nuisance”, “spurious”, “elsewhere”…**

Basic elements

some vocabulary

- random variables, PDF, CDF
- expectation values
- mean, median, mode
- standard deviation, variance, covariance matrix
- correlation coefficients
- ...

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

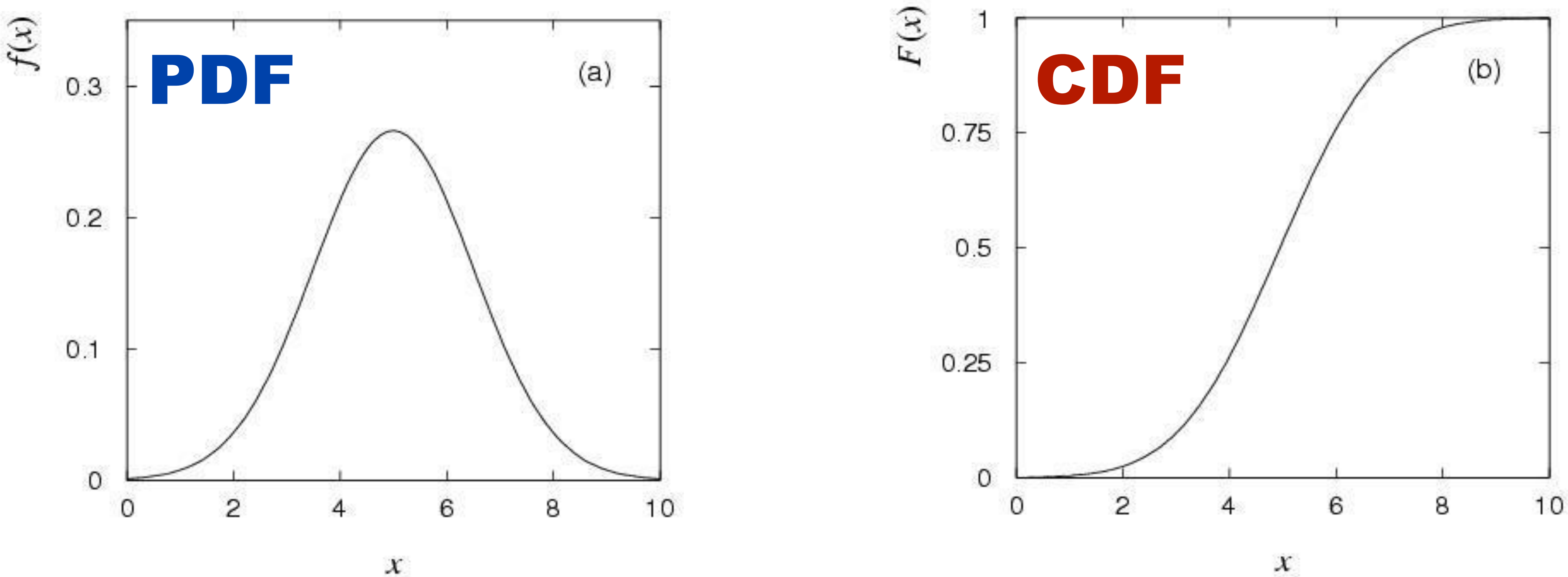
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

- 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

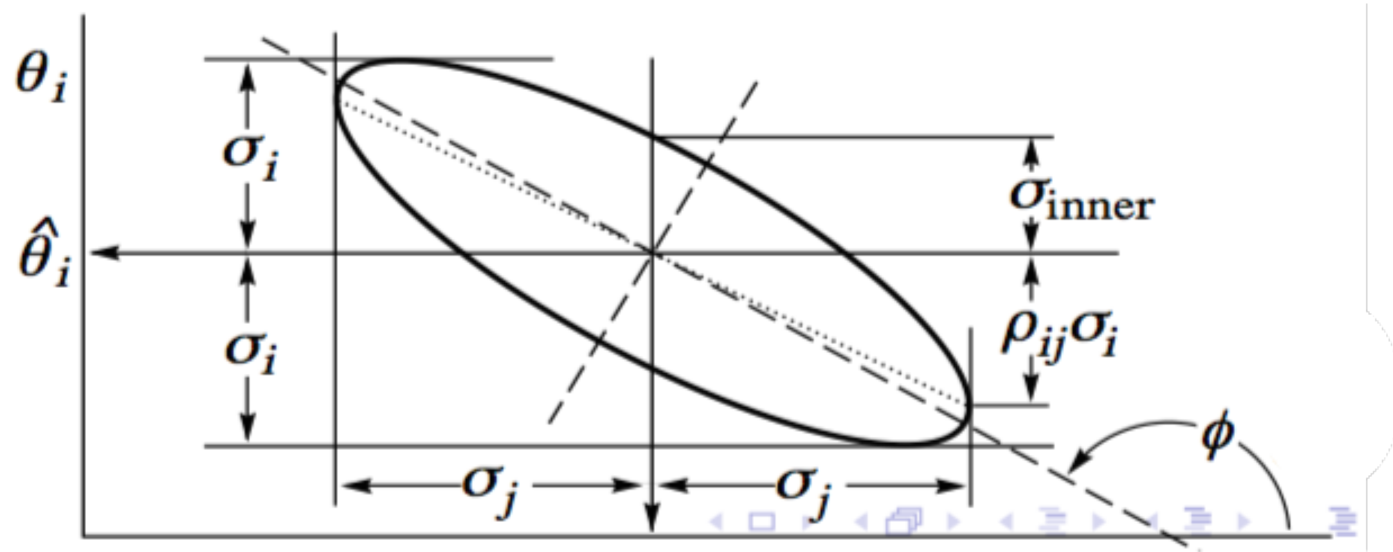
$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j}$$

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)

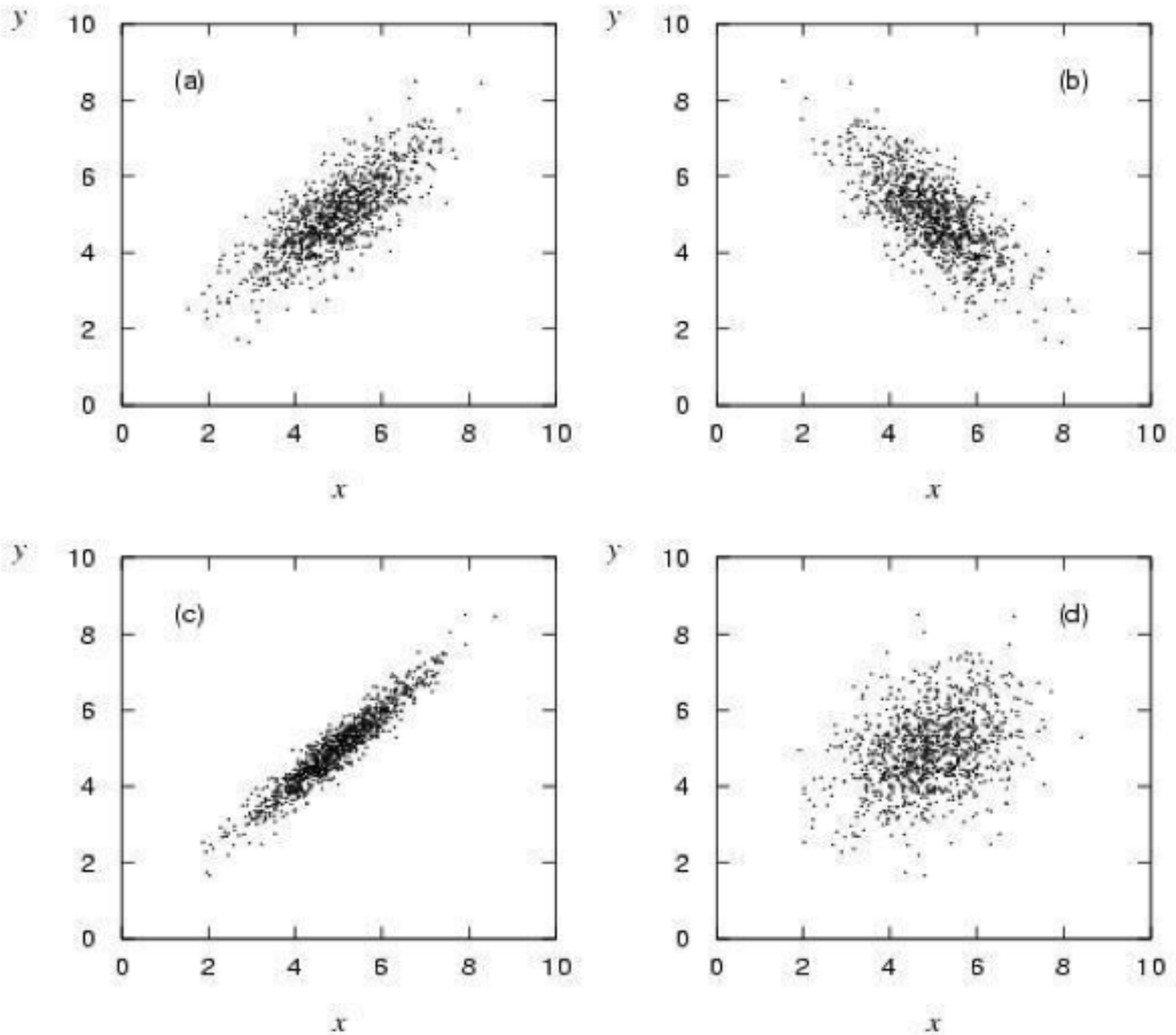
$$\tan 2\phi = \frac{2\,\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j)}{\sigma_j^2 - \sigma_i^2} = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}$$



Correlations - 2D examples

$\rho = 0.75$

$\rho = -0.75$



$\rho = 0.95$

$\rho = 0.25$

Error propagation on *f(x,y)*

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

(Q) What if x and y are independent?

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

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 about our data and need to incorporate all of what we find.

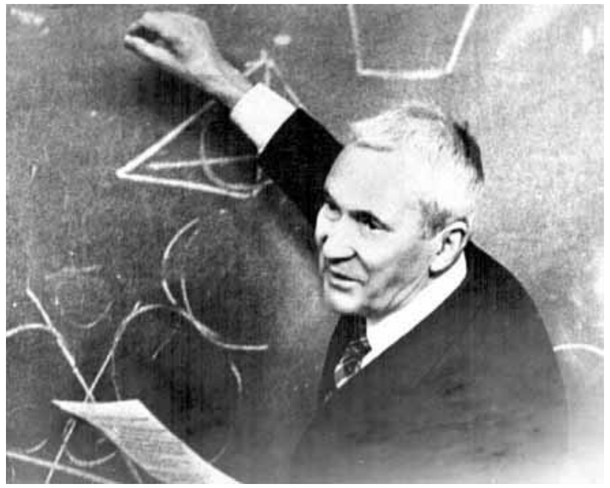
Probability Axioms

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

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

$$P(S) = 1$$

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

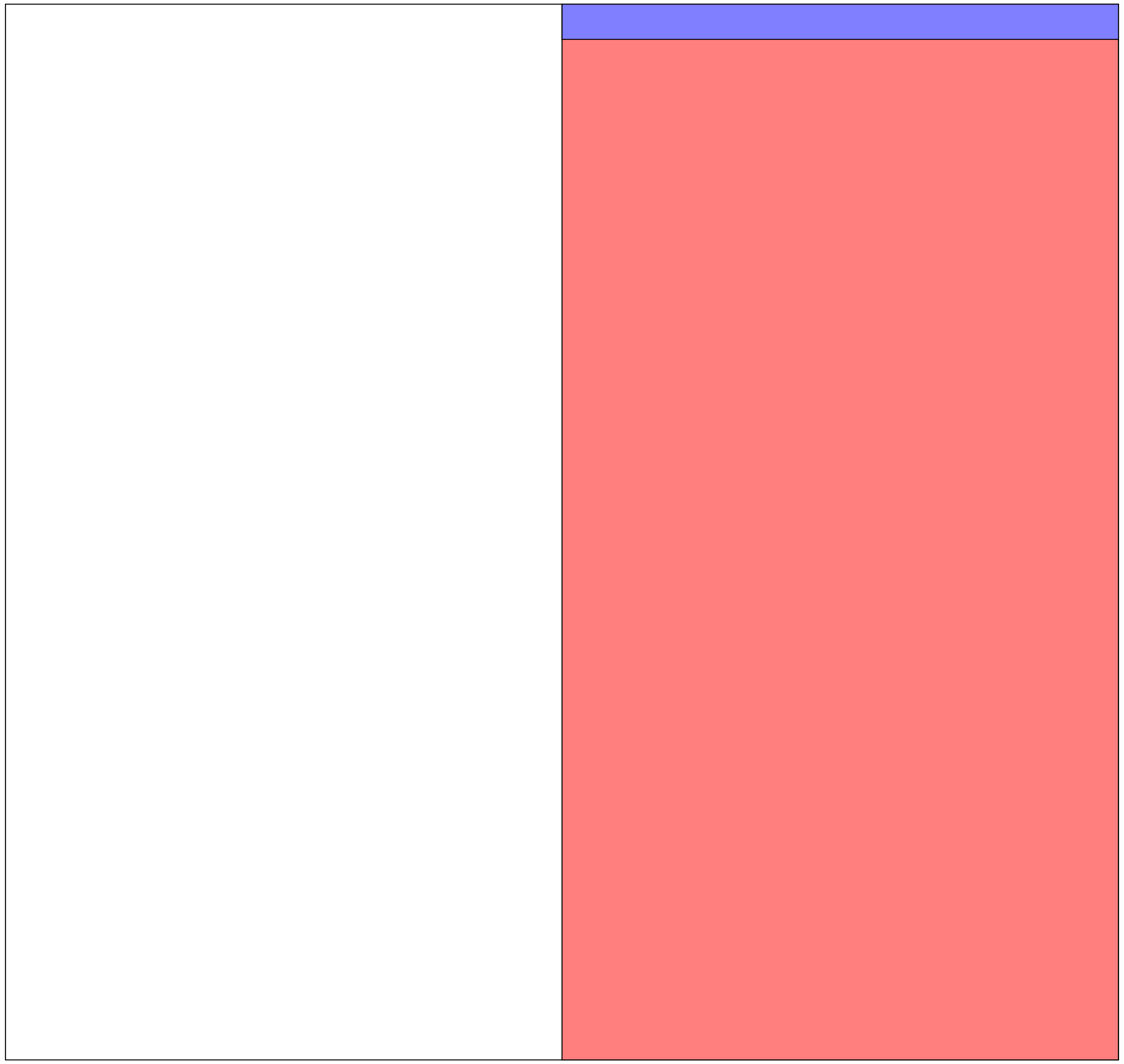


Kolmogorov (1933)

Also define conditional probability:

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

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



An extreme (and personal) case:

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

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

An extreme (and personal) case:

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

Indeed

$$P(w|p) = \frac{P(p|w) \cdot P(w)}{P(p)}$$

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

P(data|theory) ≠ P(theory|data)

Two interpretations of Probability

- **Relative frequency** **Frequentist**

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

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

- **Subjective probability** **Bayesian**

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

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

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), ...

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?**

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

where k is the number of success trials

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

Poisson distribution

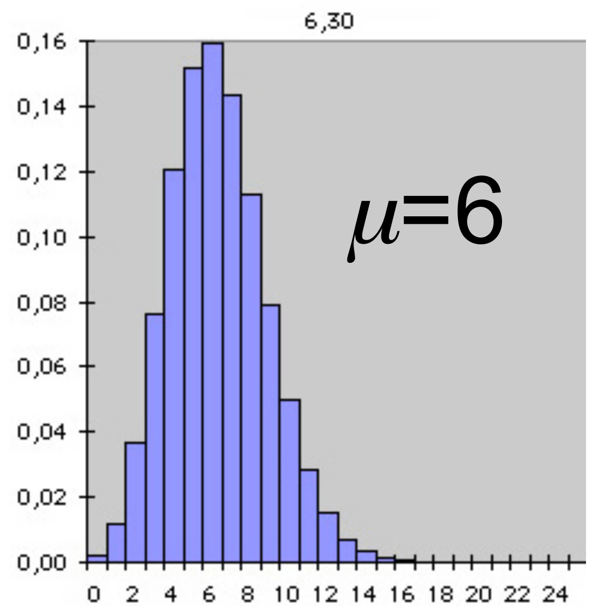
- Limit of Binomial when $N \to \infty$ and $p \to 0$ with $Np = \mu$ being finite and fixed $\Rightarrow$ Poisson distribution

$$\mathrm{Poiss}(k \mid \mu) = \frac{e^{-\mu}\mu^k}{k!} \qquad \sigma(k) = \sqrt{\mu}$$

Normalized to unit area in two different senses

$$\sum_{k=0}^{\infty} \mathrm{Poiss}(k \mid \mu) = 1, \quad \forall \mu$$

$$\int_0^{\infty} \mathrm{Poiss}(k \mid \mu)\, d\mu = 1 \quad \forall k$$

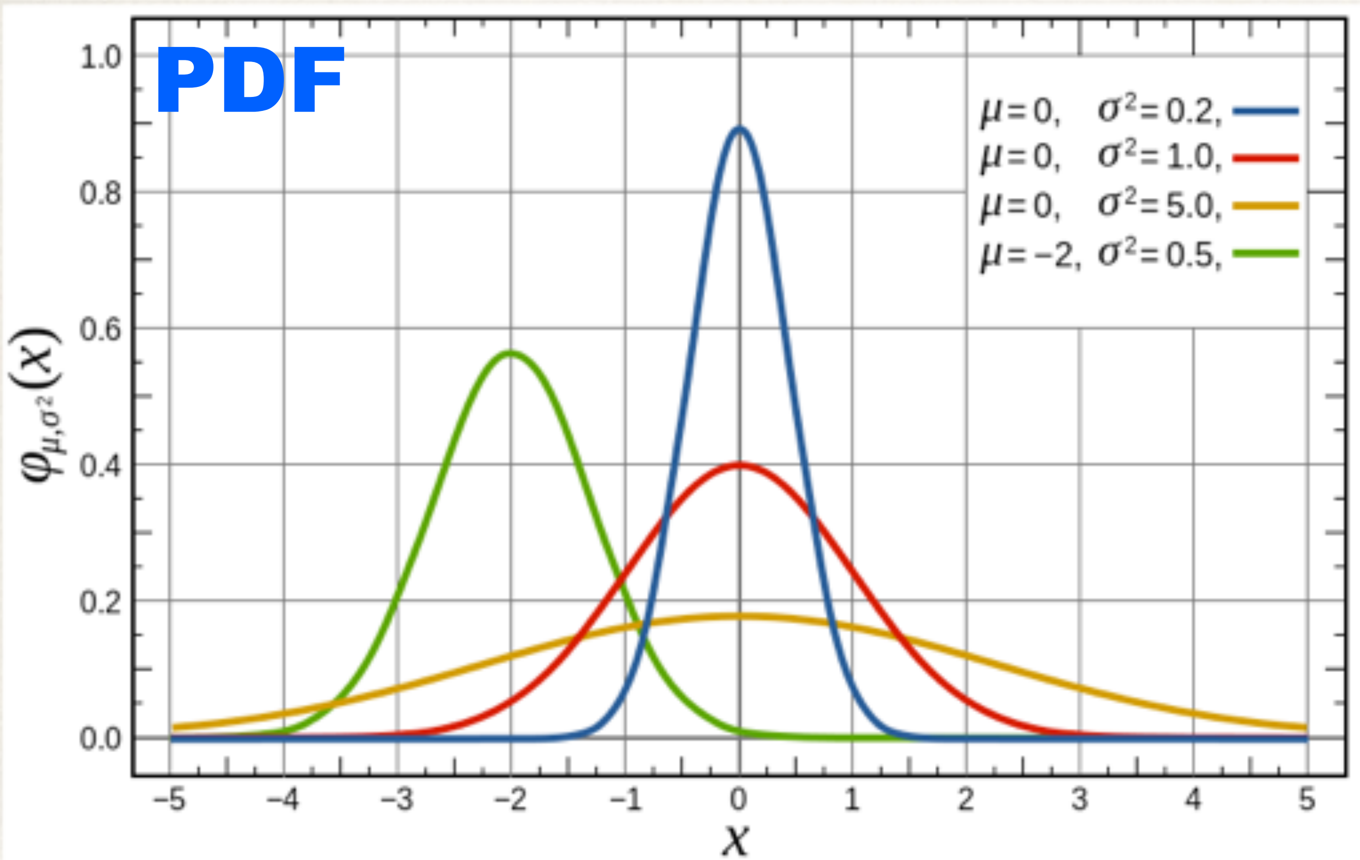


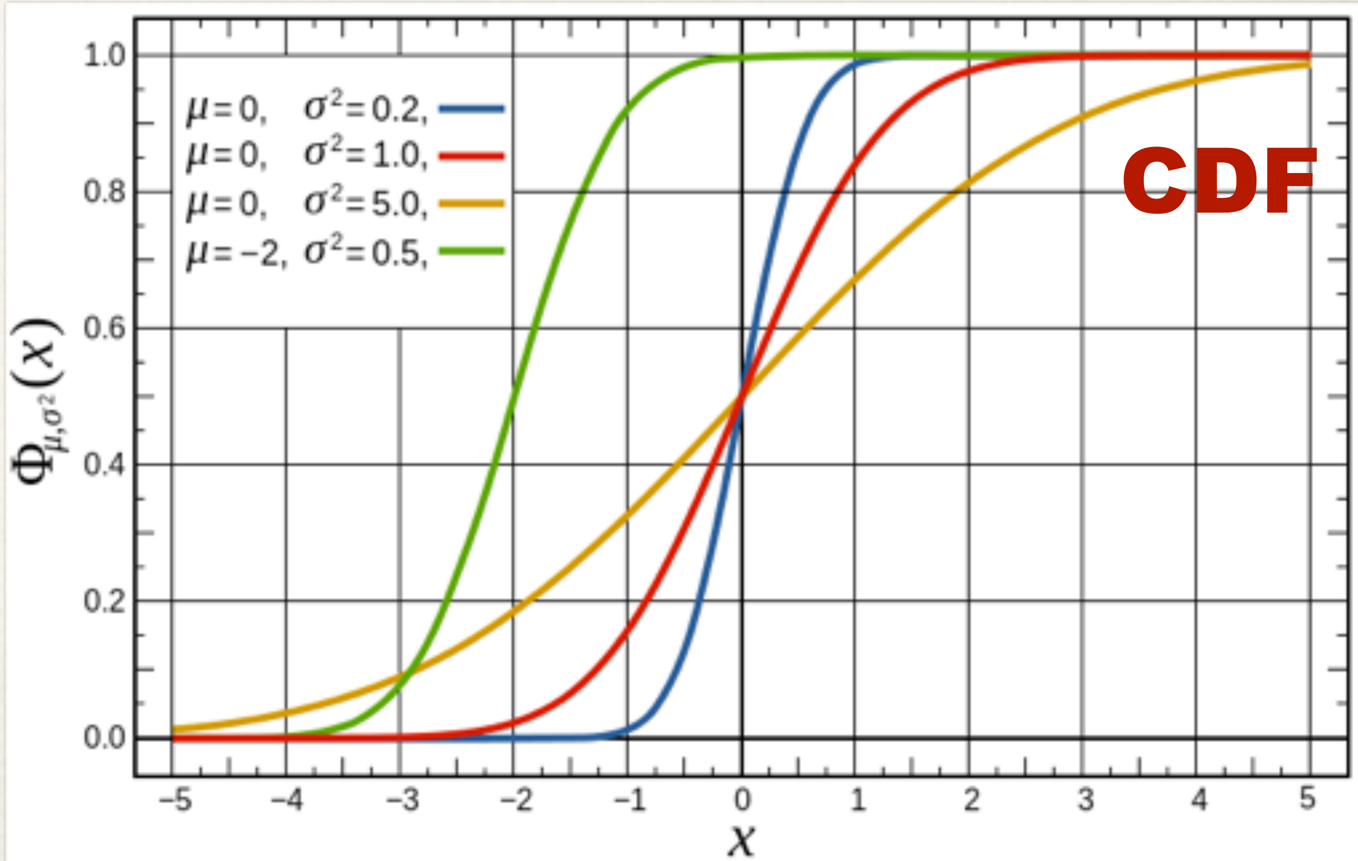
All counting results in HEP are assumed to be Poisson-distributed

Gaussian (Normal) distribution

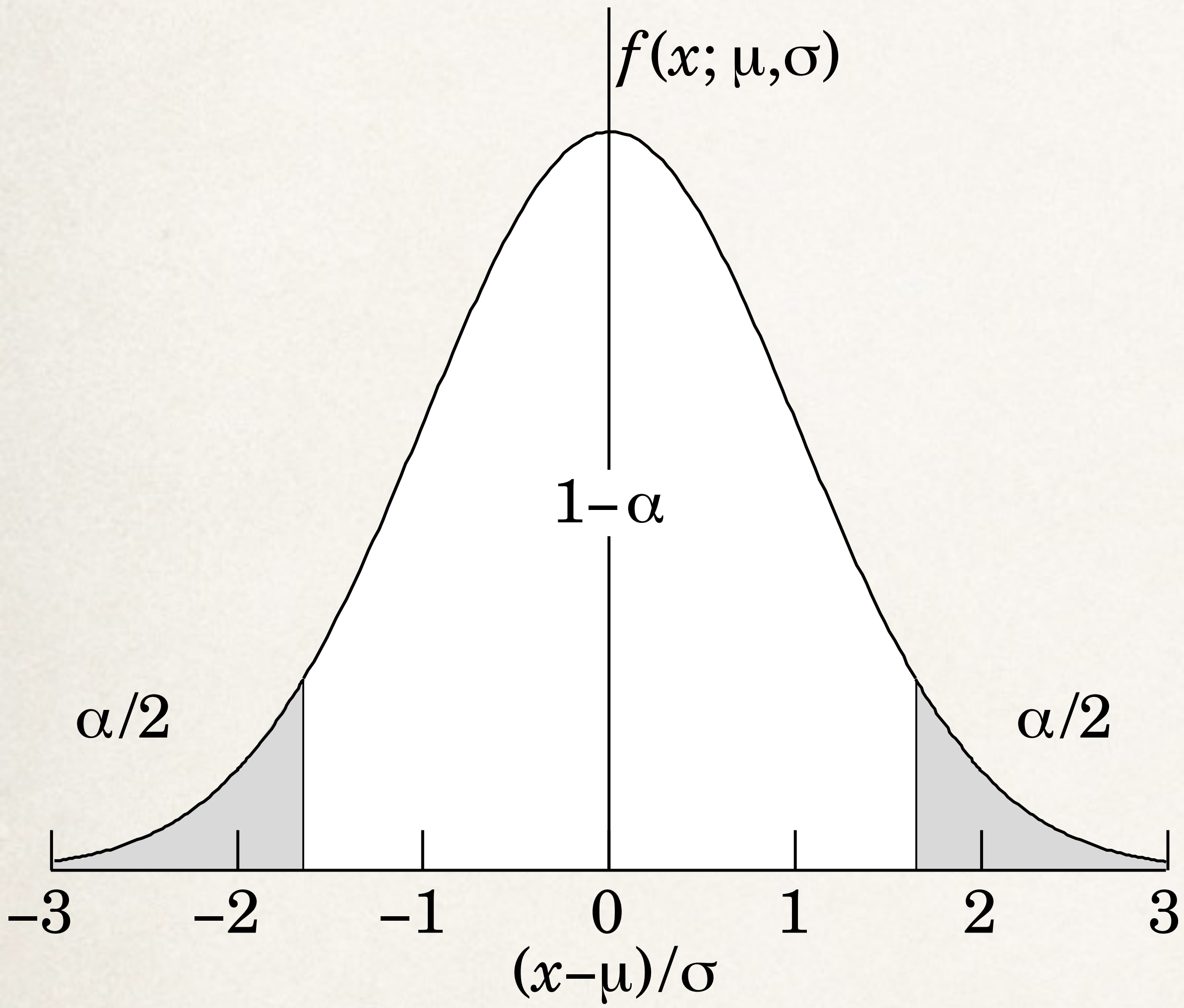
$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)$$

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





Gaussian (Normal) distribution

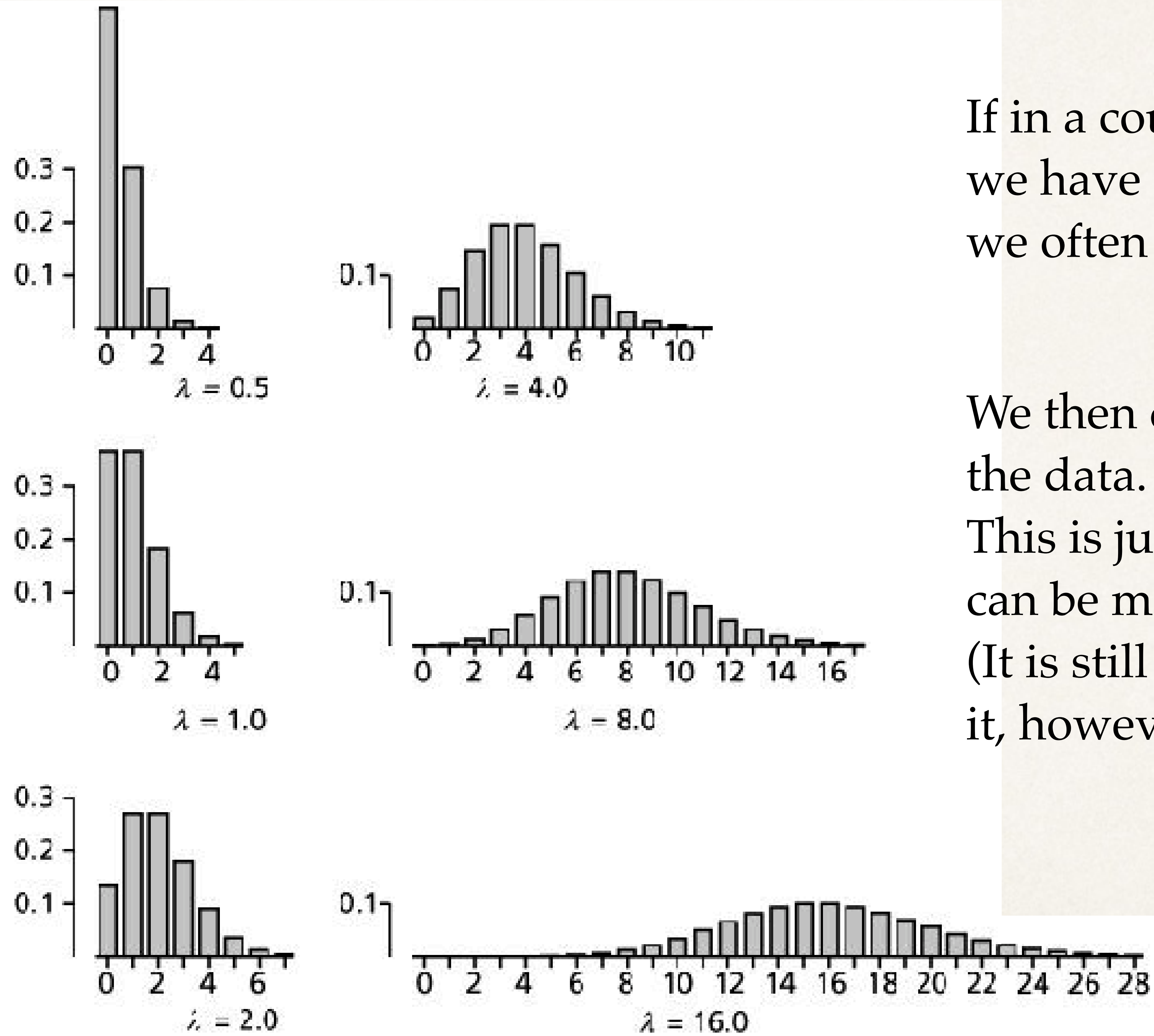


`TMath::Prob(δ²,1)`

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 $\sigma = \sqrt{\mu}$



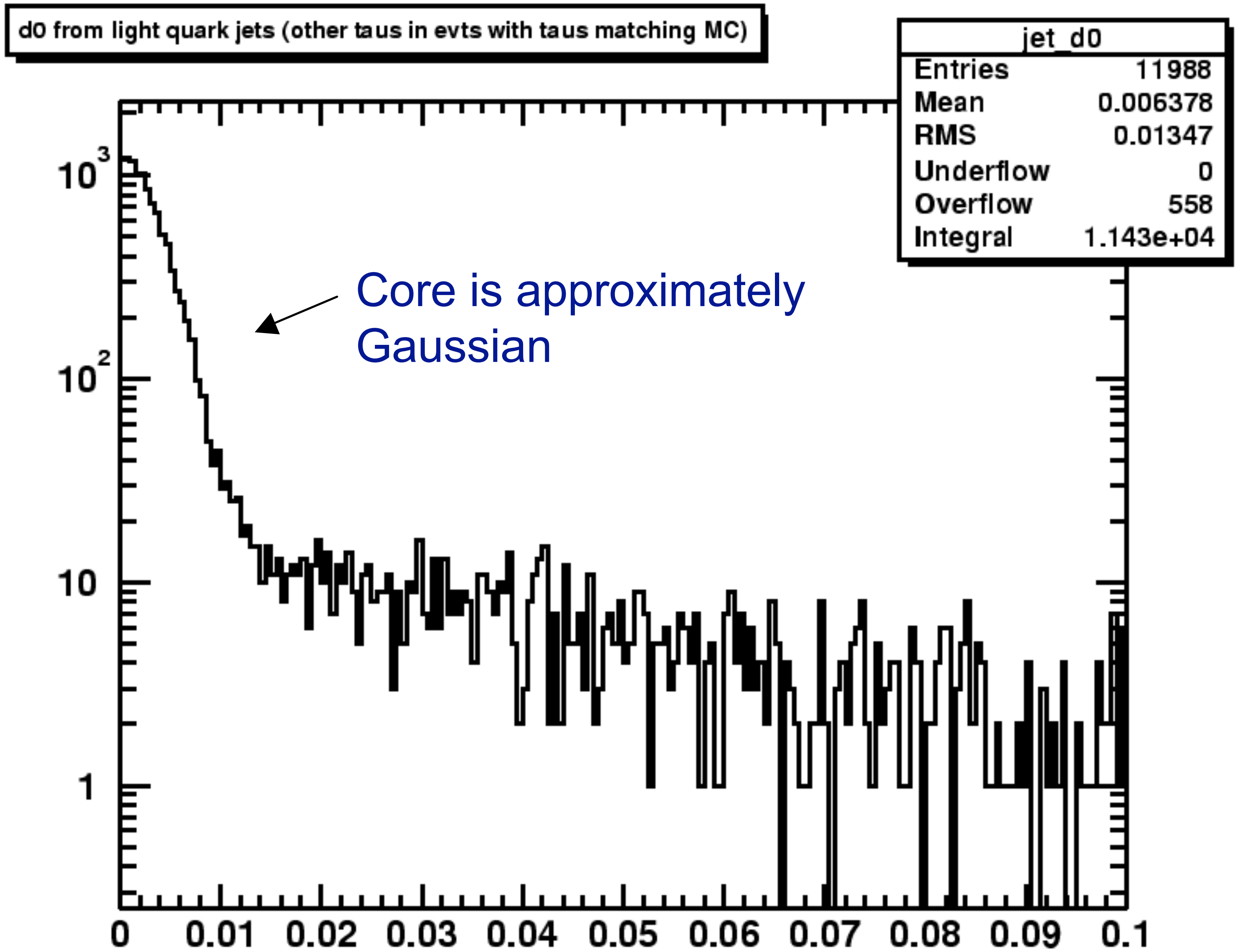
If in a counting experiment all we have is a measurement *n*, we often use this to estimate μ.

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

Track impact parameter distribution for example

Multiple scattering -- core: Gaussian; rare large scatters; heavy flavor, nuclear interactions, decays (taus in this example)

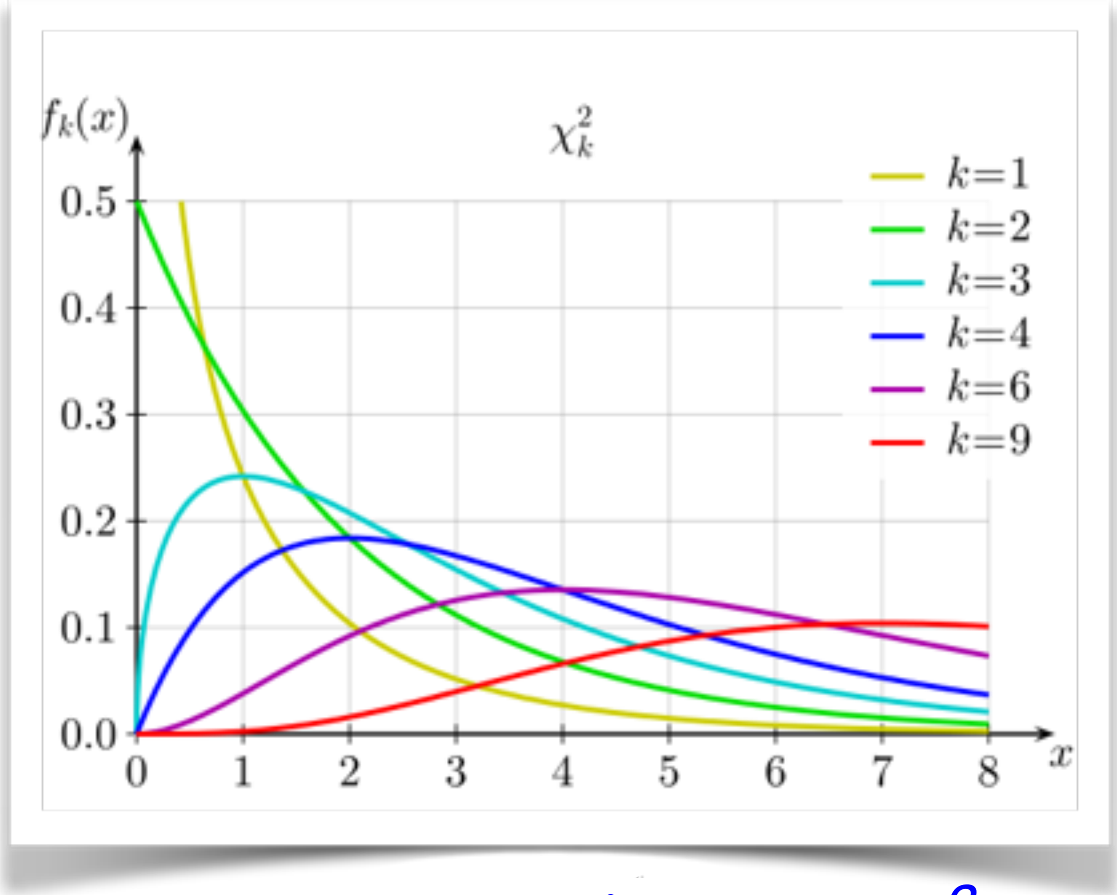


"All models are false. Some models are useful."

Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

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



n = 1, 2, ... = number of 'degrees of freedom' (dof)

$$E[z] = n\,, \qquad V[z] = 2n\,.$$

For independent Gaussian x_i, i = 1, ..., n, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$ follows χ^2 pdf with n dof.

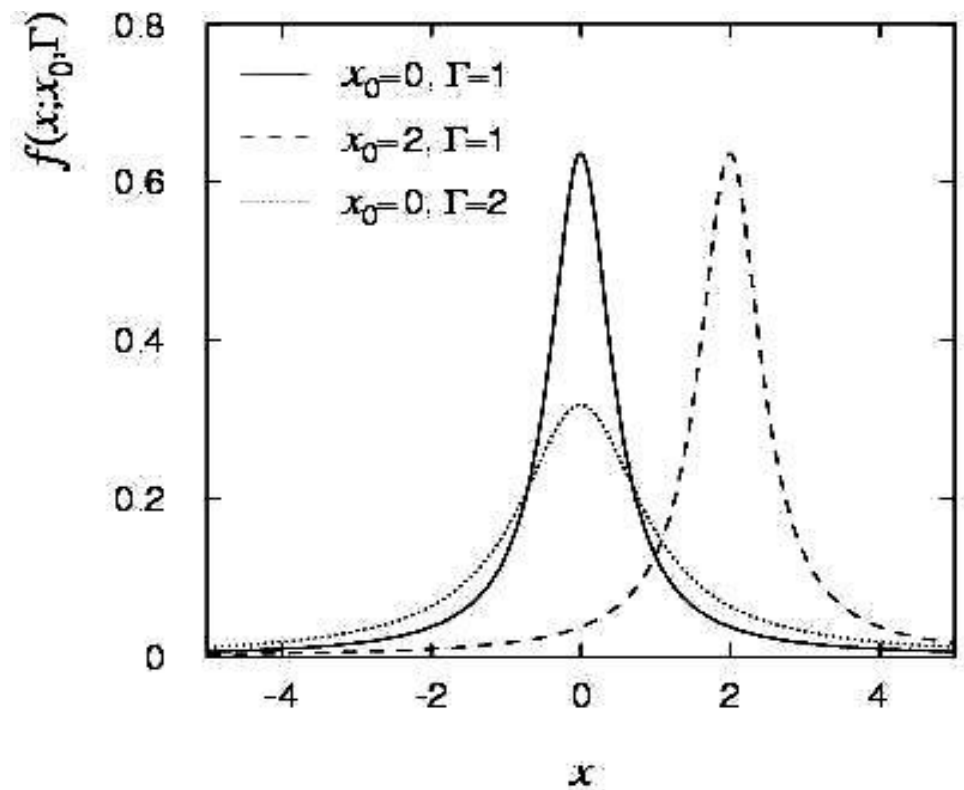
Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. x is defined by

$$f(x;\Gamma,x_0) = \frac{1}{\pi}\frac{\Gamma/2}{\Gamma^2/4+(x-x_0)^2}$$

($\Gamma = 2$, $x_0 = 0$ is the Cauchy pdf.)



$E[x]$ not well defined, $V[x] \to \infty$.

x_0 = mode (most probable value)

Γ = full width at half maximum

Example: mass of resonance particle, e.g. ρ, K^*, ϕ^0, ...

Γ = decay rate (inverse of mean lifetime)

Landau distribution

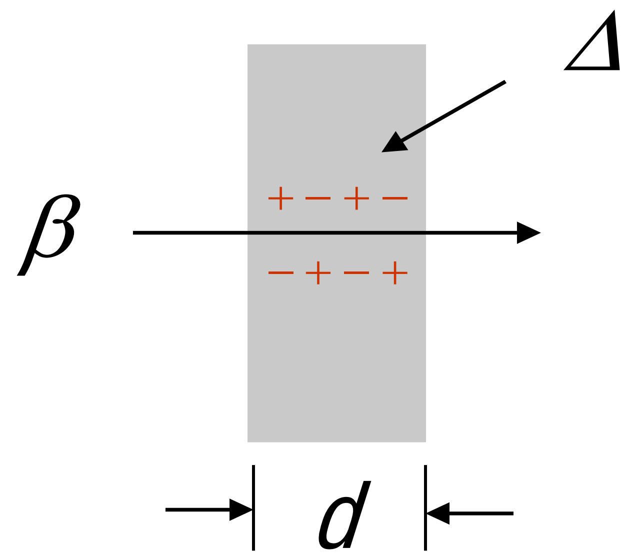
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau pdf:

$$f(\Delta;\beta) = \frac{1}{\xi}\phi(\lambda) \ ,$$

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp(-u\ln u - \lambda u)\sin \pi u \, du \ ,$$

$$\lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_{\mathrm{E}}\right)\right] \ ,$$

$$\xi = \frac{2\pi N_{\mathrm{A}} e^4 z^2 \rho \sum Z}{m_{\mathrm{e}} c^2 \sum A}\frac{d}{\beta^2} \ , \qquad \epsilon' = \frac{I^2 \exp \beta^2}{2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2} \ .$$



L. Landau, J. Phys. USSR 8 (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. 30 (1980) 253.

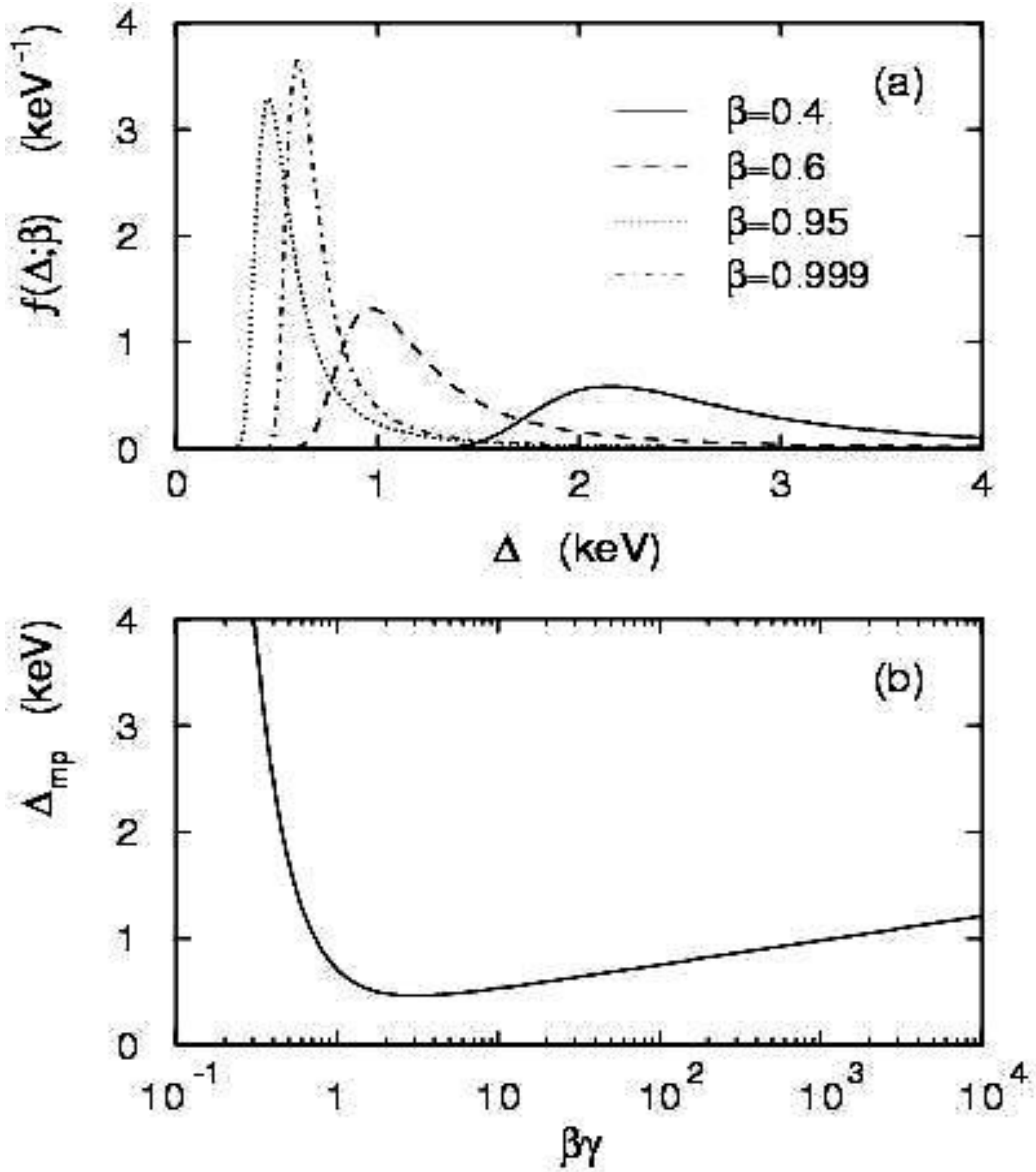
Landau distribution (2)

Long 'Landau tail'

→ all moments ∞

Mode (most probable value) sensitive to β,

→ particle i.d.

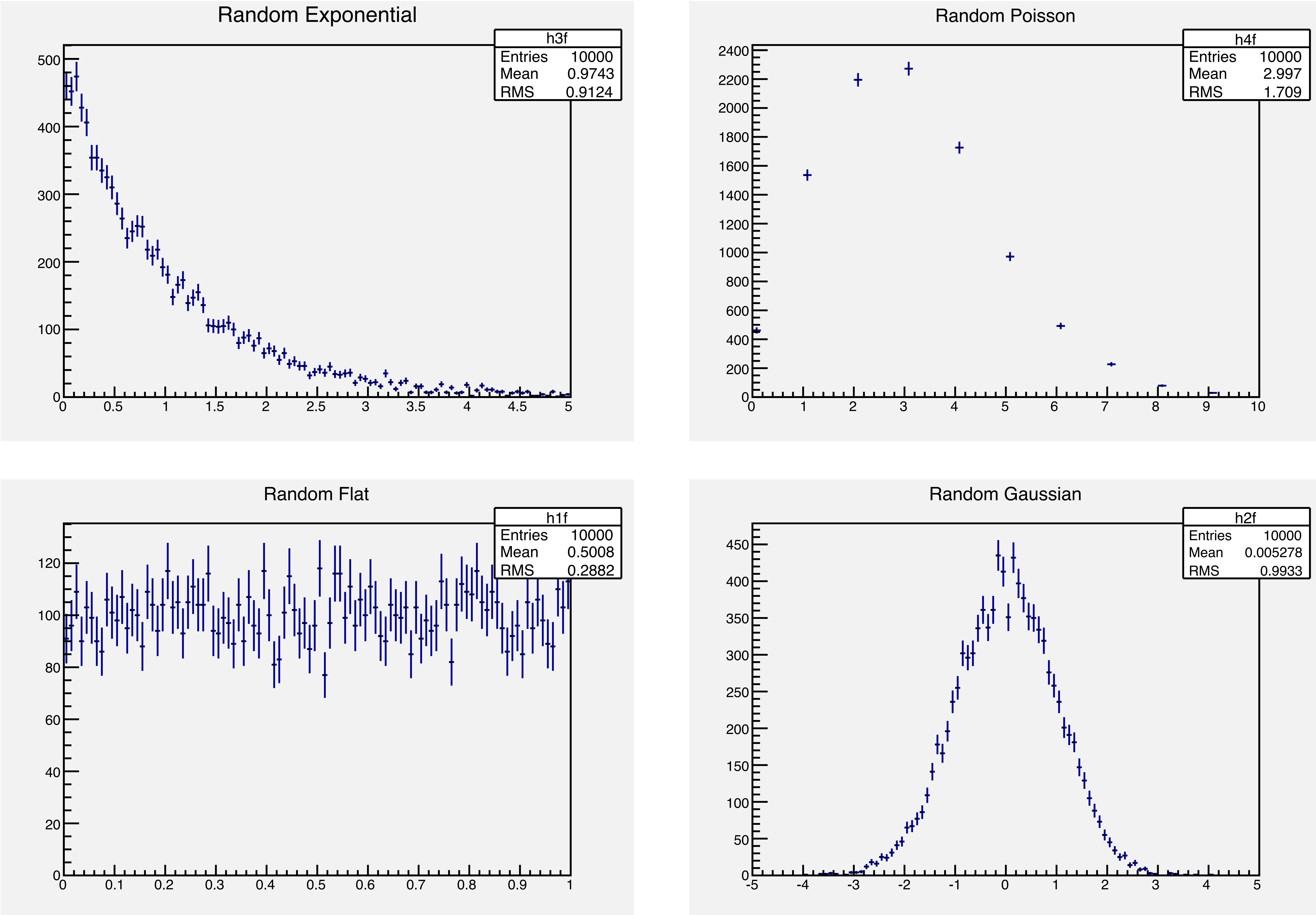


Why not make your own random variables?

- **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
0 100 200 300 400 500
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Random Poisson
h4f
Entries 10000
Mean 2.997
RMS 1.709
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400
0 1 2 3 4 5 6 7 8 9 10
Random Flat
h1f
Entries 10000
Mean 0.5008
RMS 0.2882
0 20 40 60 80 100 120
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Random Gaussian
h2f
Entries 10000
Mean 0.005278
RMS 0.9933
0 50 100 150 200 250 300 350 400 450
-5 -4 -3 -2 -1 0 1 2 3 4 5

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

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

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

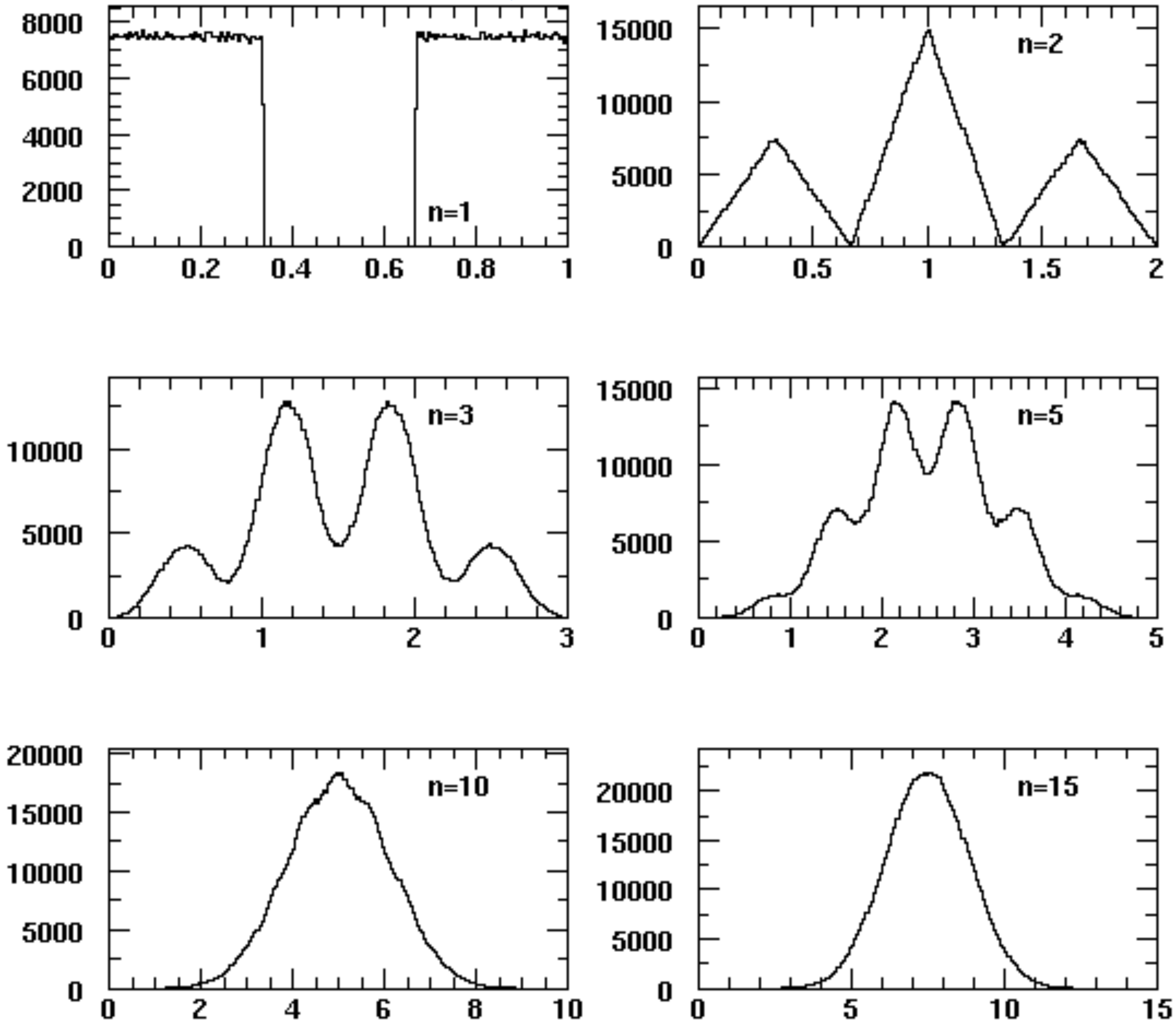
- Suppose you have a sequence of indep't random variables x_i
  - with means μ_i and variances σ_i^2
  - but otherwise distributed "however"
  - and under certain conditions on the variances

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

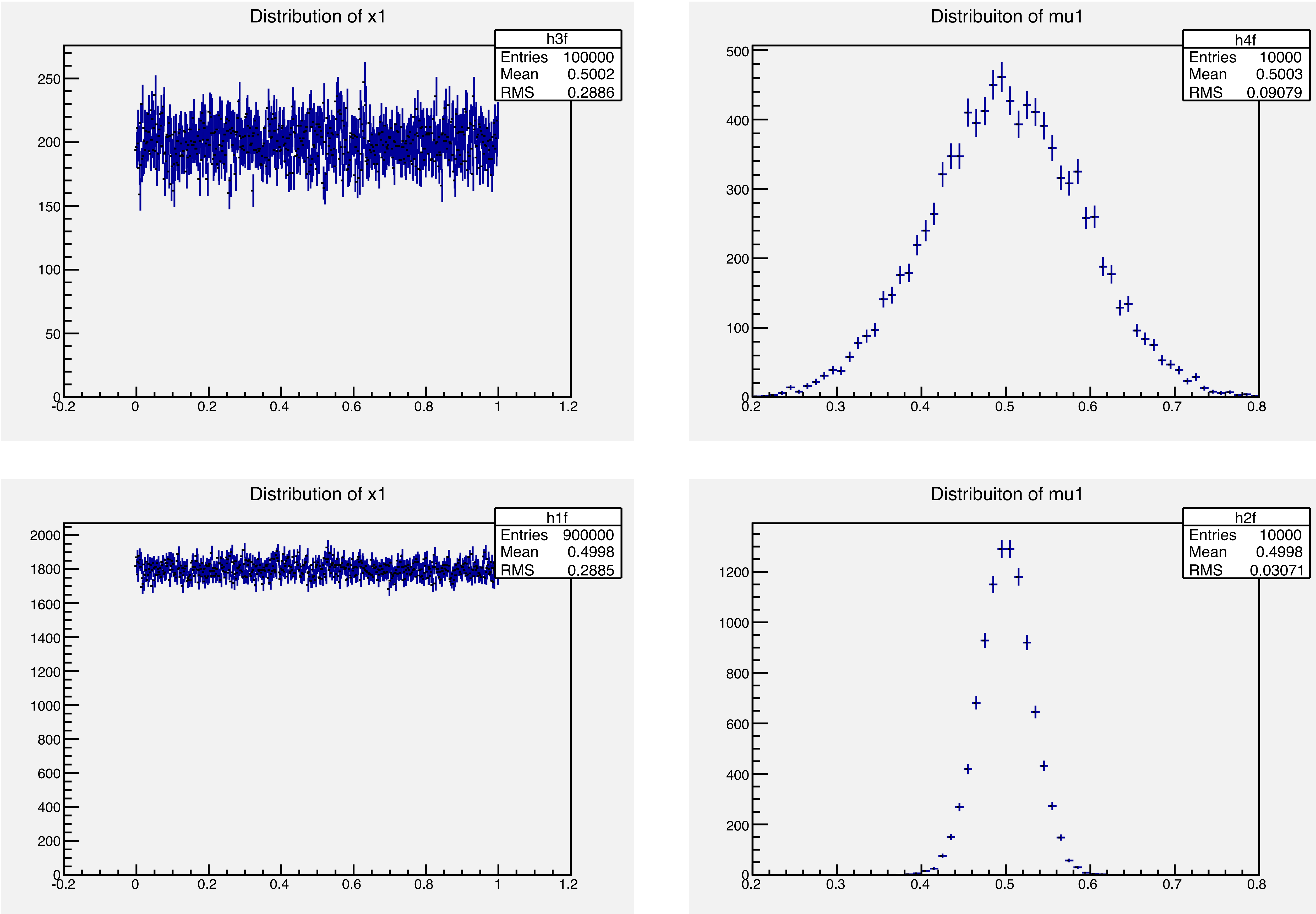
$$\lim_{N\to\infty} \frac{S - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \to \mathcal{N}(0,1) \tag{2}$$

- (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



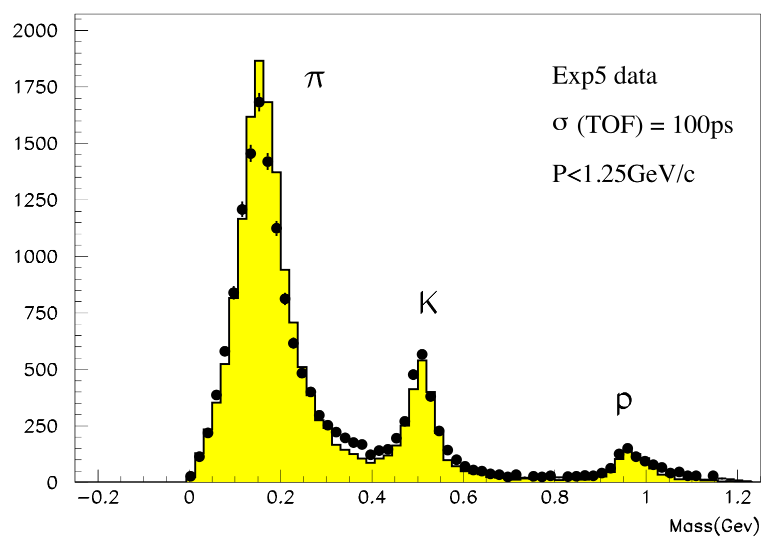
the Neyman-Pearson Lemma

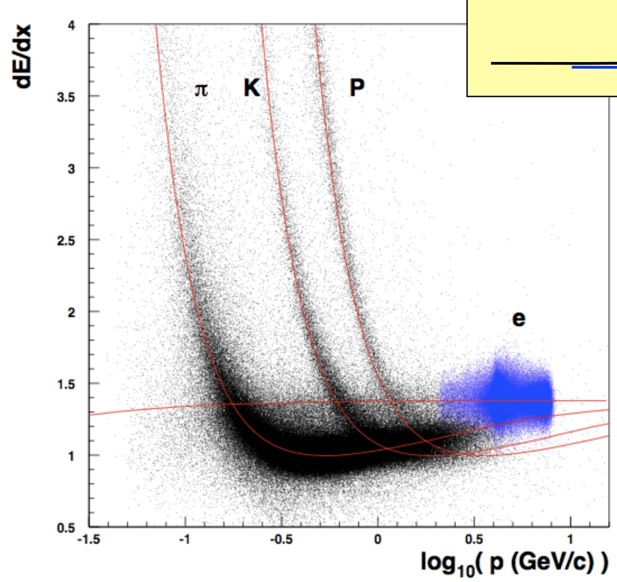
- ***We will explain it later when we discuss the "critical region" ...***

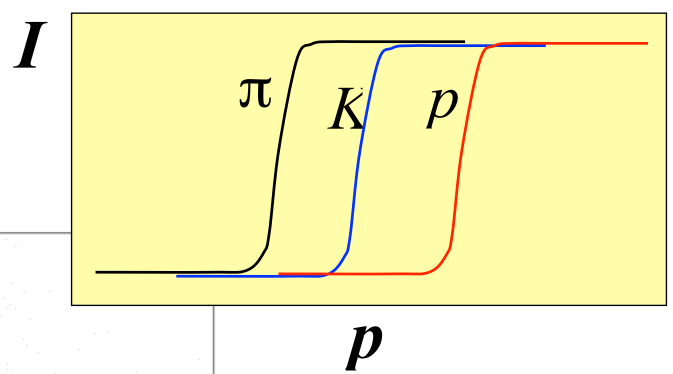
Particle identification with the `atc_pid` class is based on the likelihood of the detector response being due to an hypothesized signal particle species, compared to the likelihood for an assumed background particle species. This is expressed as a likelihood ratio

$$Prob(i:j) = \frac{P_i}{P_i + P_j} \qquad P_i = P_i^{dE/dx} \times P_i^{TOF} \times P_i^{ACC}$$

where P_i is the particle-ID likelihood calculated for the signal particle species and P_j for the background particle species; i and j can be any of five particle species, e, μ, π, K and p. Clearly $Prob(i:j)$ is distributed on the interval $[0,1]$, and we usually think of it as







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.*

Hypothesis Testing

Remember?

Two approaches

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

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

$$P(S) = 1$$

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

- **Relative frequency**

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

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

Frequentist

- **Subjective probability**

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

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

Bayesian

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**

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

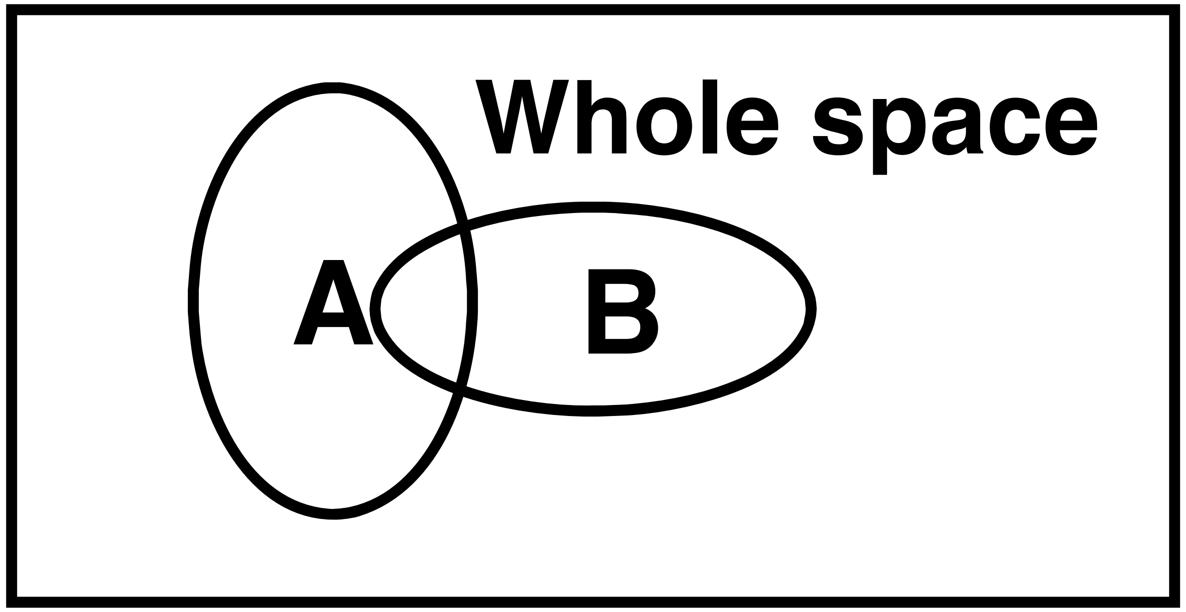
- but $P(A \cap B) = P(B \cap A)$
- therefore,

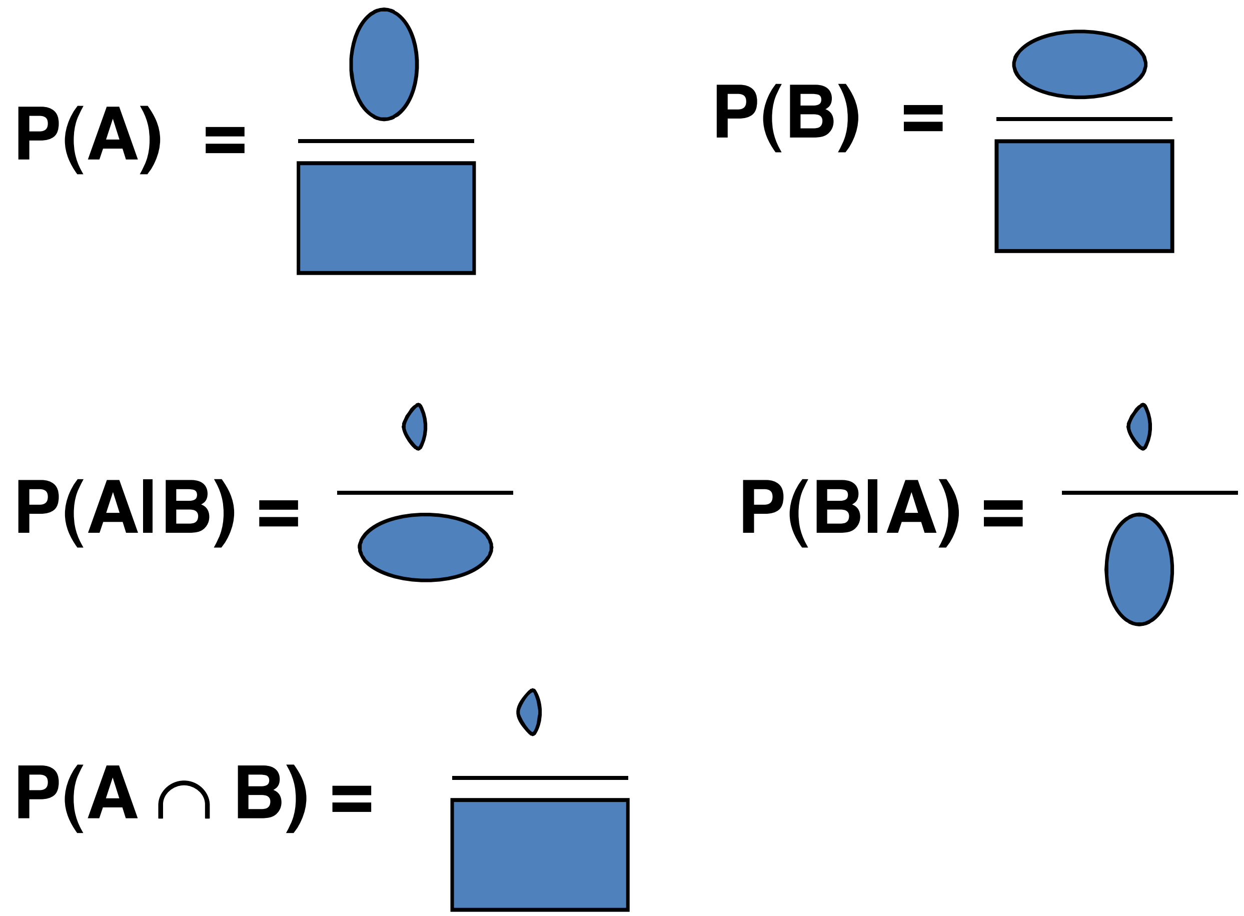
$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

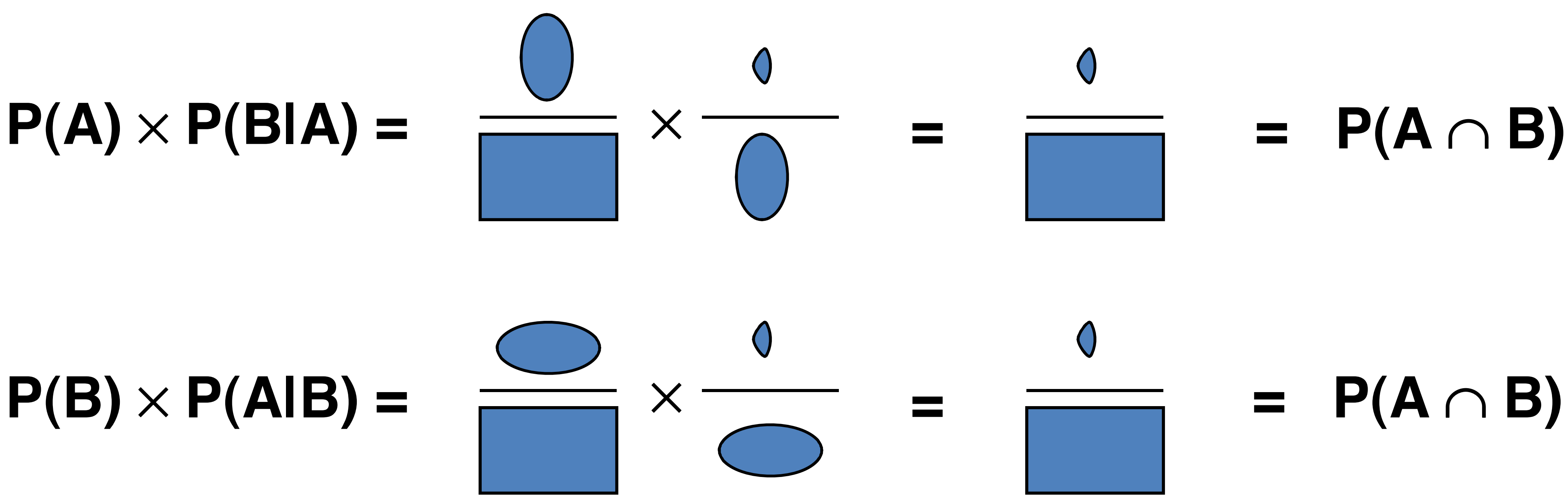
- First published (posthumous) by Rev. Thomas Bayes (1702-1761)

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







$\Rightarrow$ $P(B|A) = P(A|B) \times P(B) / P(A)$

Bob Cousins, CMS, 2008

Bayesian probability: tossing a coin

- suppose I stand to win or lose money in a game of chance
- my companion gives me a coin to use in the game
- do I trust the coin?
- what is P(fair coin)?
- frequentist answer:
  - toss the coin n times
  - $P(\text{heads}) = \lim_{n\to\infty} n_H/n$
  - make a complicated statement about the results, which is *only indirectly* about whether the coin is fair (see Lec.2 . . .)
- but I can only test the coin with five throws:
  - I get 4H, 1T
  - do I trust the coin?
- frequentist answer based on these 5 trials: not much info
- Bayesian answer depends on your *prior belief* . . .
- assume for illustration that a bad coin has $P(\text{heads}) = 0.75$
- a proper analysis would involve integrating over priors, *etc.*

Bayesian probability: interpreting the coin tosses

Likelihoods:

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

Priors:

$$P(\mathrm{fair} \mid \mathbf{GG}) = 0.95$$
$$P(\mathrm{bad} \mid \mathbf{GG}) = 0.05$$

Posterior:

$$\begin{aligned} P(\mathrm{fair} \mid (4H,1T), \mathbf{GG}) &= \frac{P((\mathrm{4H,1T}) \mid \mathrm{fair}) \cdot P(\mathrm{fair} \mid \mathbf{GG})}{\sum_i P((\mathrm{4H,1T}) \mid i) \cdot P(i \mid \mathbf{GG})} \\ &= \frac{0.1563 \cdot 0.95}{0.1563 \cdot 0.95 + 0.3955 \cdot 0.05} \\ &= 0.882 \end{aligned}$$

Bayesian probability: interpreting the coin tosses

Likelihoods:

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

Priors:

$$P(\text{fair} \mid \textbf{BG}) = 0.50$$
$$P(\text{bad} \mid \textbf{BG}) = 0.50$$

Posterior:

$$\begin{aligned} P(\text{fair} \mid (4H, 1T), \text{BG}) &= \frac{P((4\text{H},1\text{T}) \mid \text{fair}) \cdot P(\text{fair} \mid \textbf{BG})}{\sum_i P((4\text{H},1\text{T}) \mid i) \cdot P(i \mid \textbf{BG})} \\ &= \frac{0.1563 \cdot 0.50}{0.1563 \cdot 0.50 + 0.3955 \cdot 0.50} \\ &= 0.283 \end{aligned}$$

Frequentist statistics – general philosophy

- In frequentist statistics, probabilities such as

  $P(\text{Higgs boson exists})$

  $P(0.117 < \alpha_s < 0.121)$

  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)

Hypothesis testing

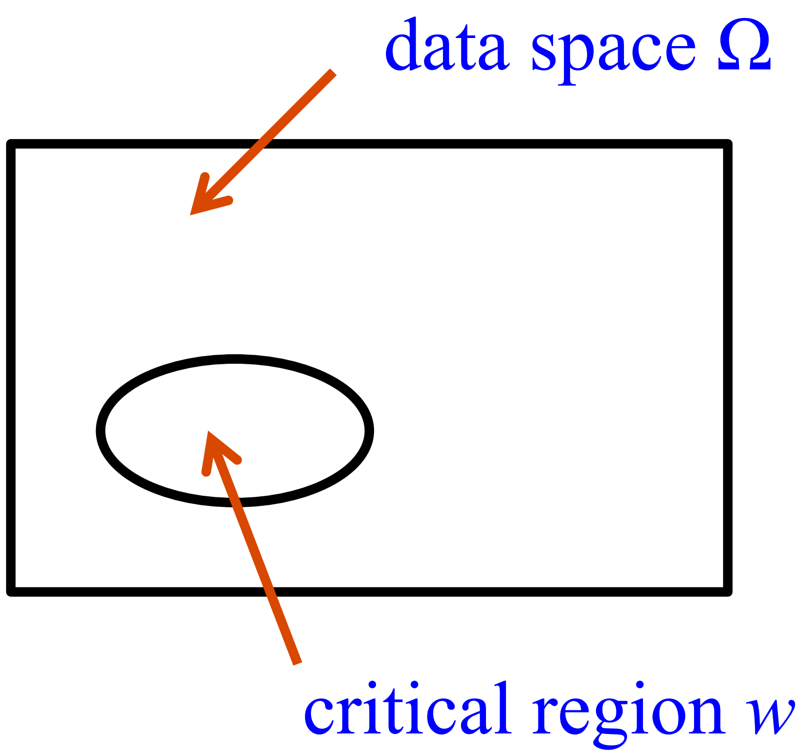
- A hypothesis H specifies the probability for the data
  (*shown symbolically as* $\vec{x}$ *here*),
  often expressed as a function $f(\vec{x}|H)$
- The measured data $\vec{x}$ could be anything:
  - observation of a single particle, a single event, or an entire experiment
  - uni-/multi-variate, continuous or discrete
- the two kinds:
  - simple (or "point") hypothesis – $f(\vec{x}|H)$ is completely specified
  - composite hypothesis – H contains unspecified parameter(s)
- The probability for $\vec{x}$ given H is also called the **likelihood** of the hypothesis, written as $L(\vec{x}|H)$

Hypothesis test

- Consider e.g. a simple hypothesis H_0 and an alternative H_1
- A (frequentist) test of H_0:

  Specify a critical region w of the data space Ω such that, assuming H_0 is correct, there is no more than some (small) probability α to observe data in w

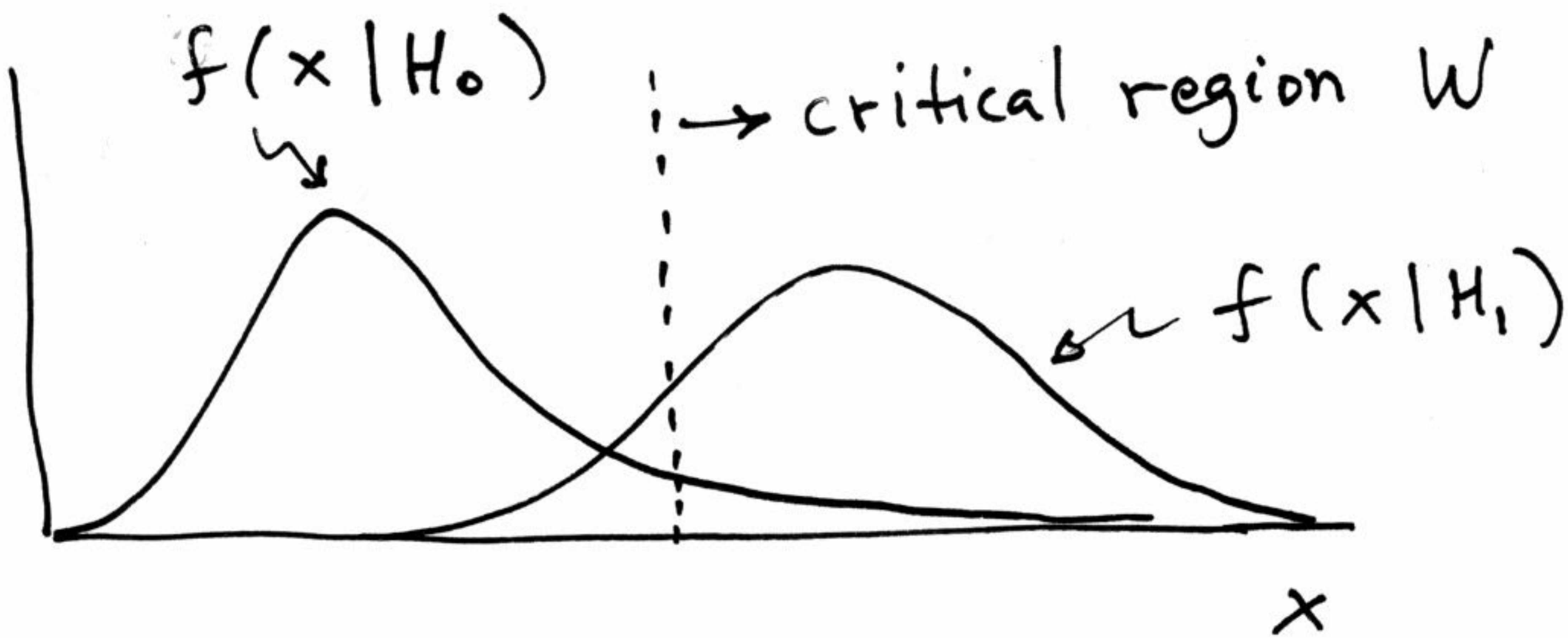
$$P(\vec{x} \in w | H_0) \leq \alpha$$



- α: "size" or "significance level" of the test
- If $\vec{x}$ is observed within w, we reject H_0 with a confidence level $1 - \alpha$

Hypothesis test

- In general, $\exists$ an ∞ number of possible critical regions that give the same significance level α
- Usually, we place the critical region where there is a low probability α for $\vec{x} \in w$ if H_0 is true, but high if the alternative (H_1) is true



Test statistic

- The boundary surface of the critical region for an n-dim. data space can be defined by an equation of the form:

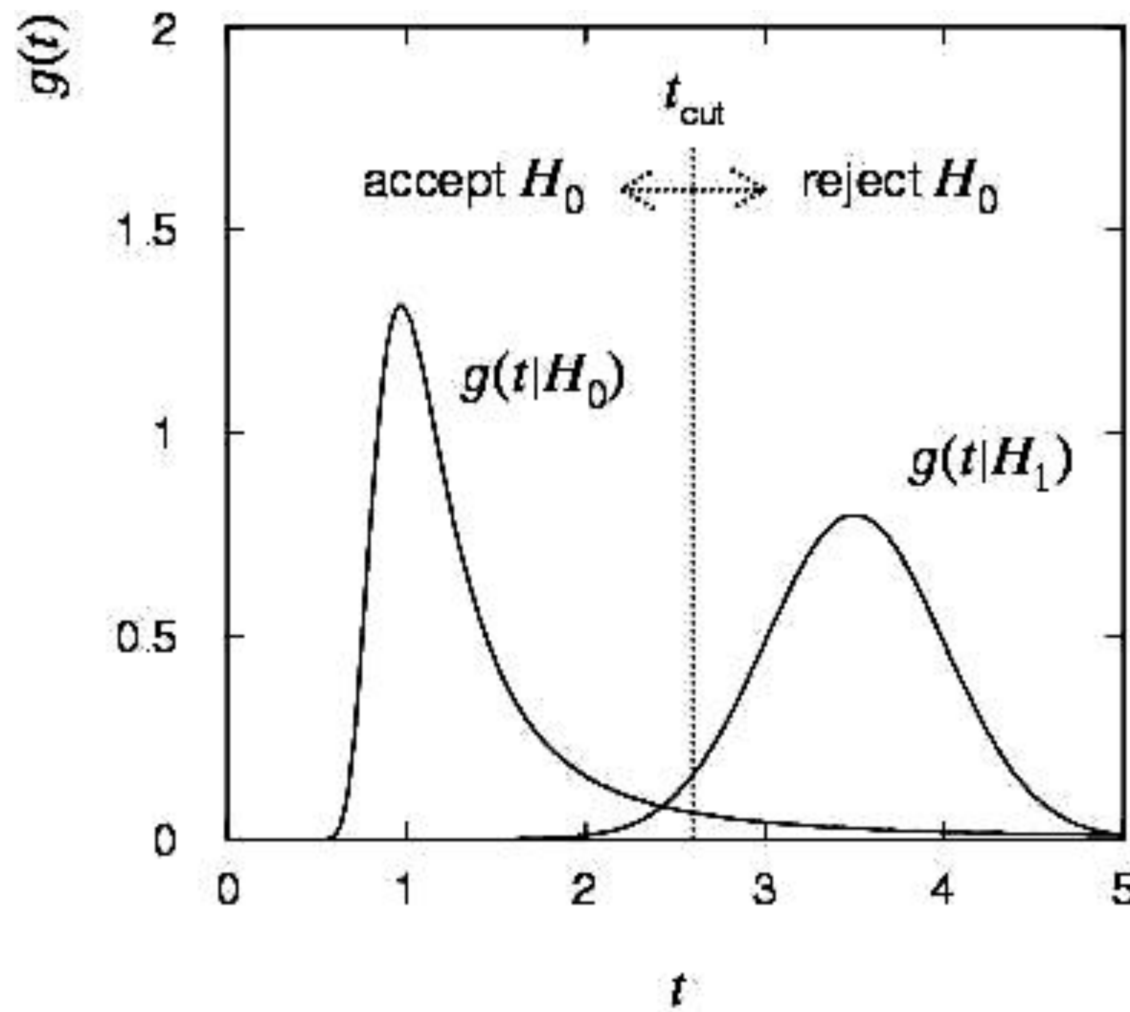
$$t(x_1, \cdots, x_n) = t_c$$

  where $t(x_1, \cdots, x_n)$ is a scalar **test statistic**.

- For the test statistic t, we can work out the PDFs $g(t|H_0)$, $g(t|H_1)$, etc.

- Decision boundary is now given by a signle 'cut' on t, thus defining the critical region

  $\Rightarrow$ for an n-dim. data space, the problem is reduced to a 1-dim. problem



Type-I, Type-II errors

- Rejecting H_0 when it is true is called the **Type-I error**

  (Q) Given the significance α of the test, what is the maximum probability of Type-I error?

- We might also accept H_0 when it is indeed false, and an alternative H_1 is true. This is called the **Type-II error**

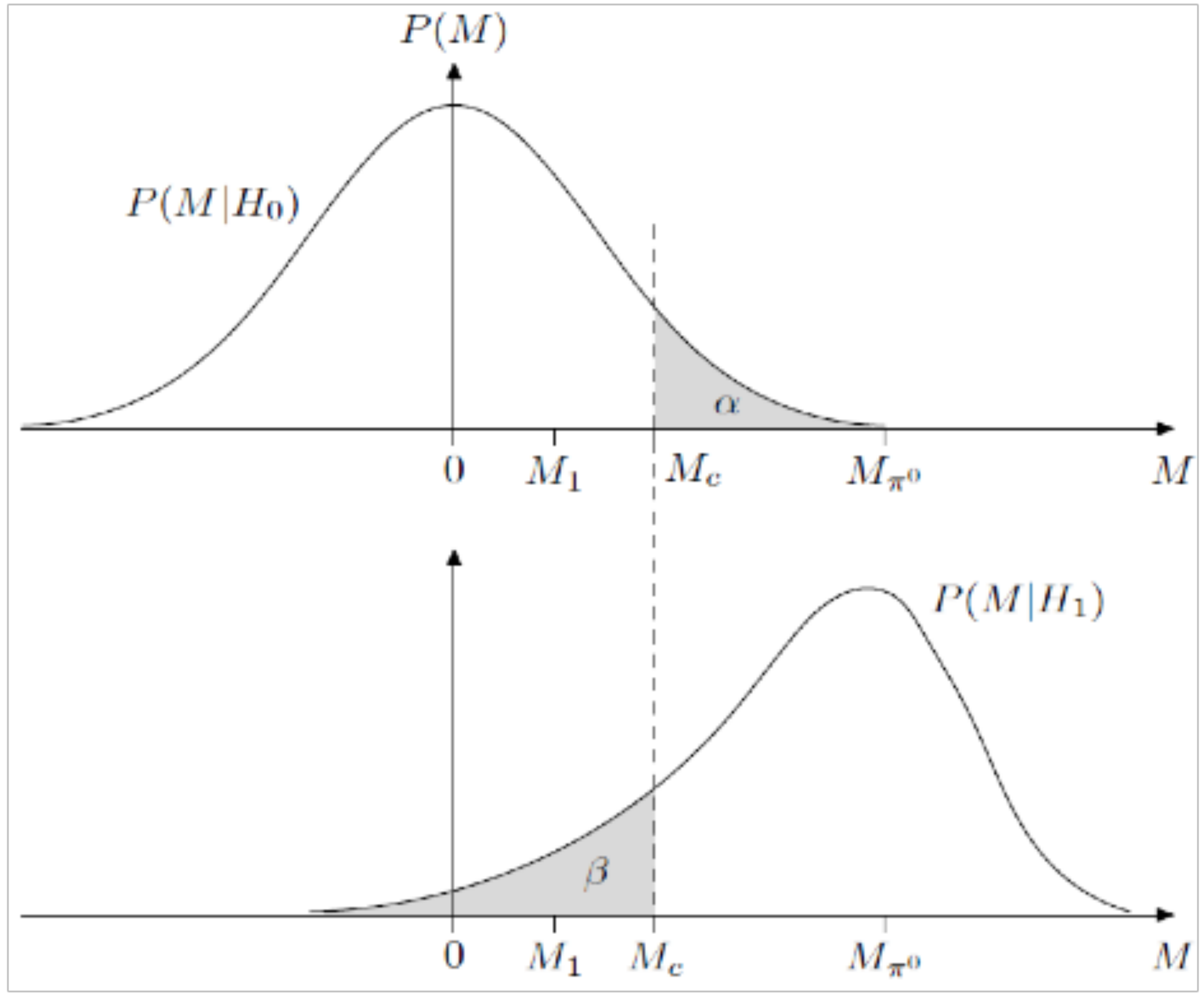
  The probability β of Type-II error:

$$P(\vec{x} \in \Omega - w | H_1) = \beta$$

  $1 - \beta$ is called the **power** of the test with respect to H_1

Two possible errors

	H_0 chosen	H_1 chosen
H_0 true	Correct decision, Prob = 1-α	Type I error, Prob = α
H_1 true	Type II error, Prob = β	Correct decision, Prob = 1-β



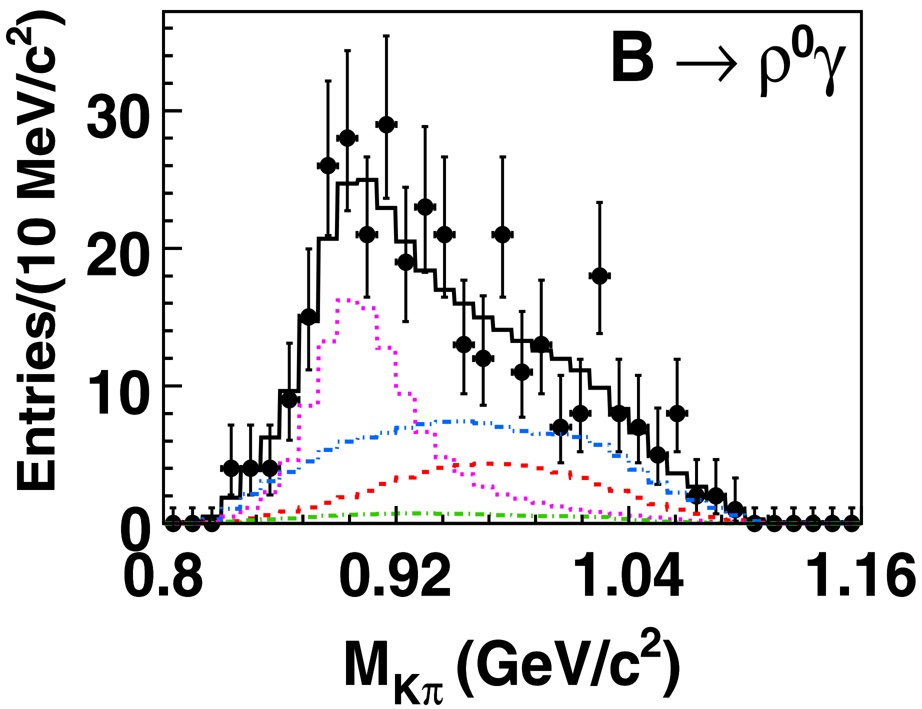
Optimal decision: minimize β for given α

- The size of the test is $\Pr_0(Y \in R_\alpha)=\alpha$.
- The power of the test is $\Pr_1(Y \in R_\alpha)=1-\beta$.

from an FAPPS09 Lecture by S. T'Jampens

exercise on Type-I, II errors

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. ***What are H_0 & H_1, for each case?***

- $f_{\pi^+ \to K^+}$ = 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^+

Probability $P(H|\vec{x})$

- In the frequentist approach, we do not, in general, assign probability of a hypothesis itself.

  Rather, we compute the probability to accept/reject a hypothesis assuming that it (or some alternative) is true.

- In Bayesian, on the other hand, probability of any given hypothesis (*degree of belief*) could be obtained by using the Bayes' theorem:

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

  which depends on the prior probability $\pi(H)$

How to choose an *optimal* test statistic

- Use **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.

- Equivalently, the optimal scalar test statistic is

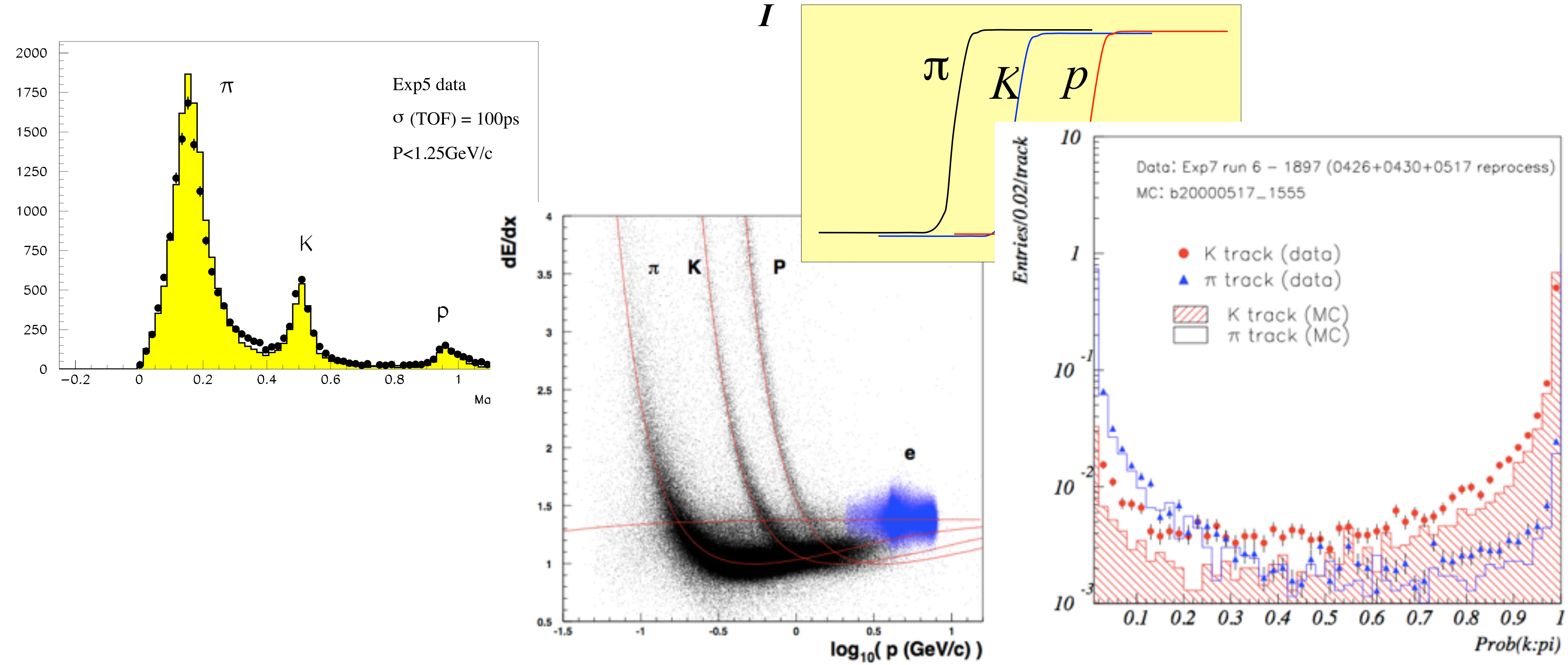
$$t(\vec{x}) = P(\vec{x}|H_1)/P(\vec{x}|H_0)$$

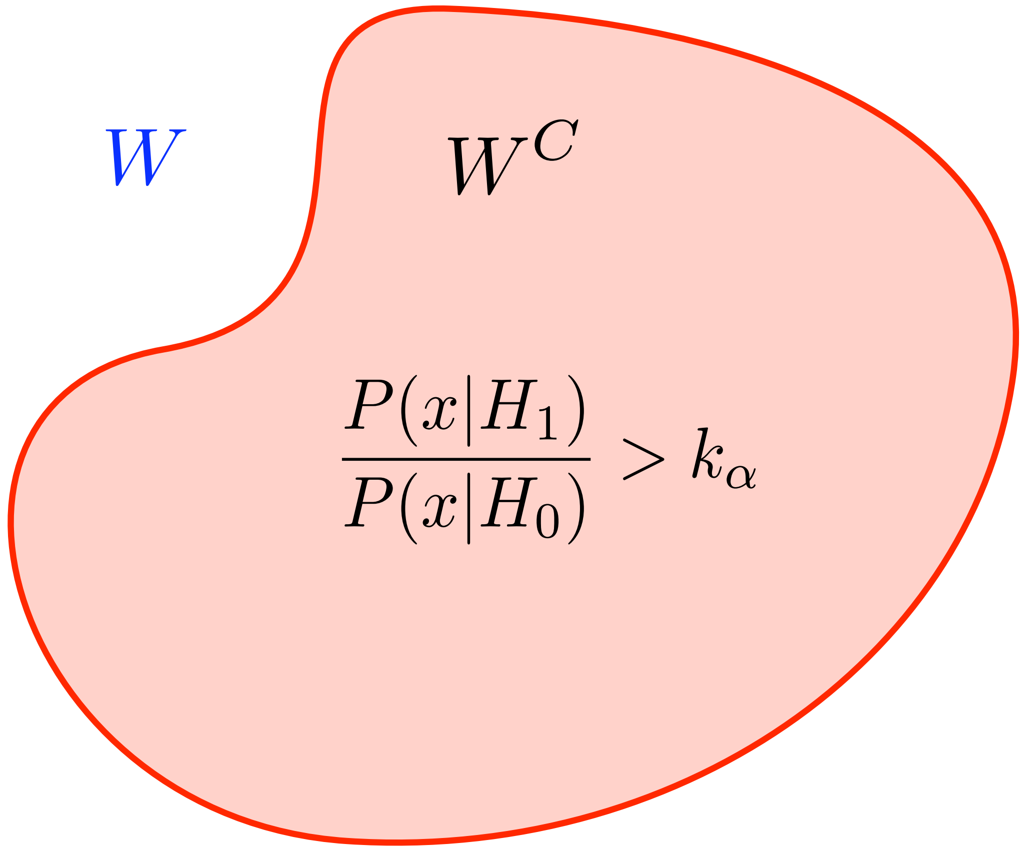
  (Note) Any monotonic function of this leads to the *same test*.

Particle identification with the atc_pid class is based on the likelihood of the detector response being due to an hypothesized signal particle species, compared to the likelihood for an assumed background particle species. This is expressed as a likelihood ratio

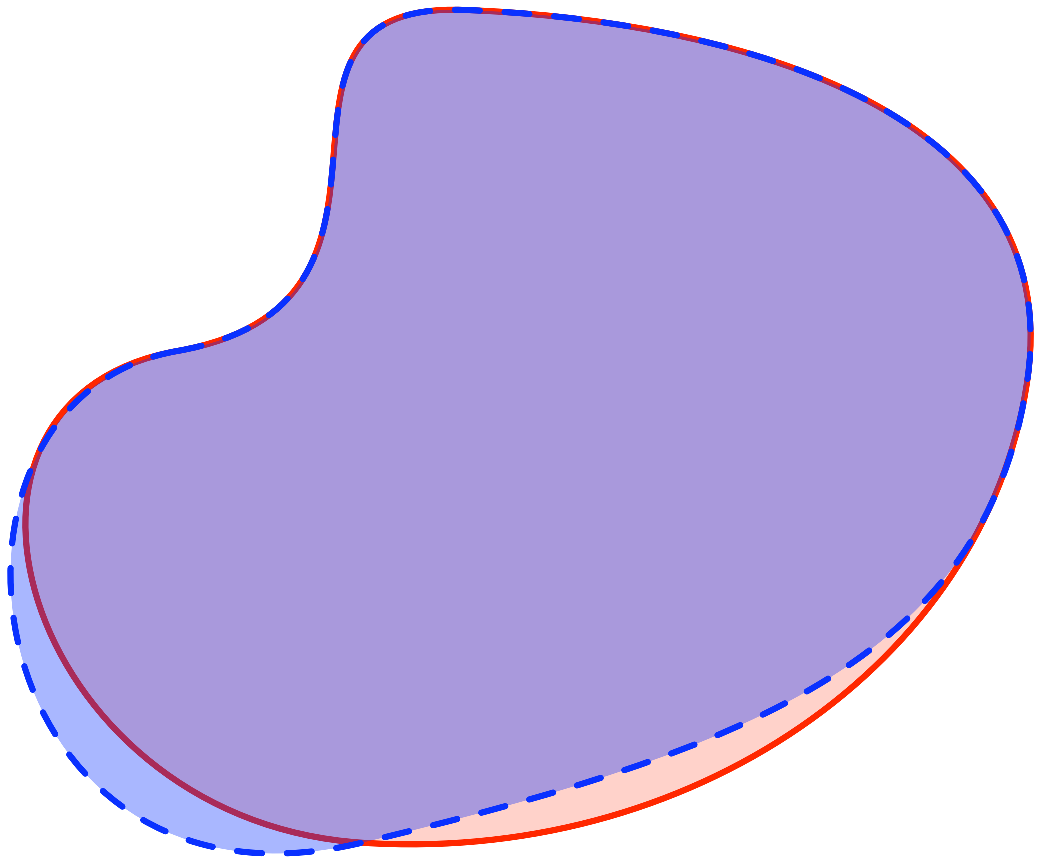
$$Prob(i:j) = \frac{P_i}{P_i + P_j} \qquad P_i = P_i^{dE/dx} \times P_i^{TOF} \times P_i^{ACC}$$

where P_i is the particle-ID likelihood calculated for the signal particle species and P_j for the background particle species; i and j can be any of five particle species, e, μ, π, K and p. Clearly $Prob(i:j)$ is distributed on the interval $[0,1]$, and we usually think of it as

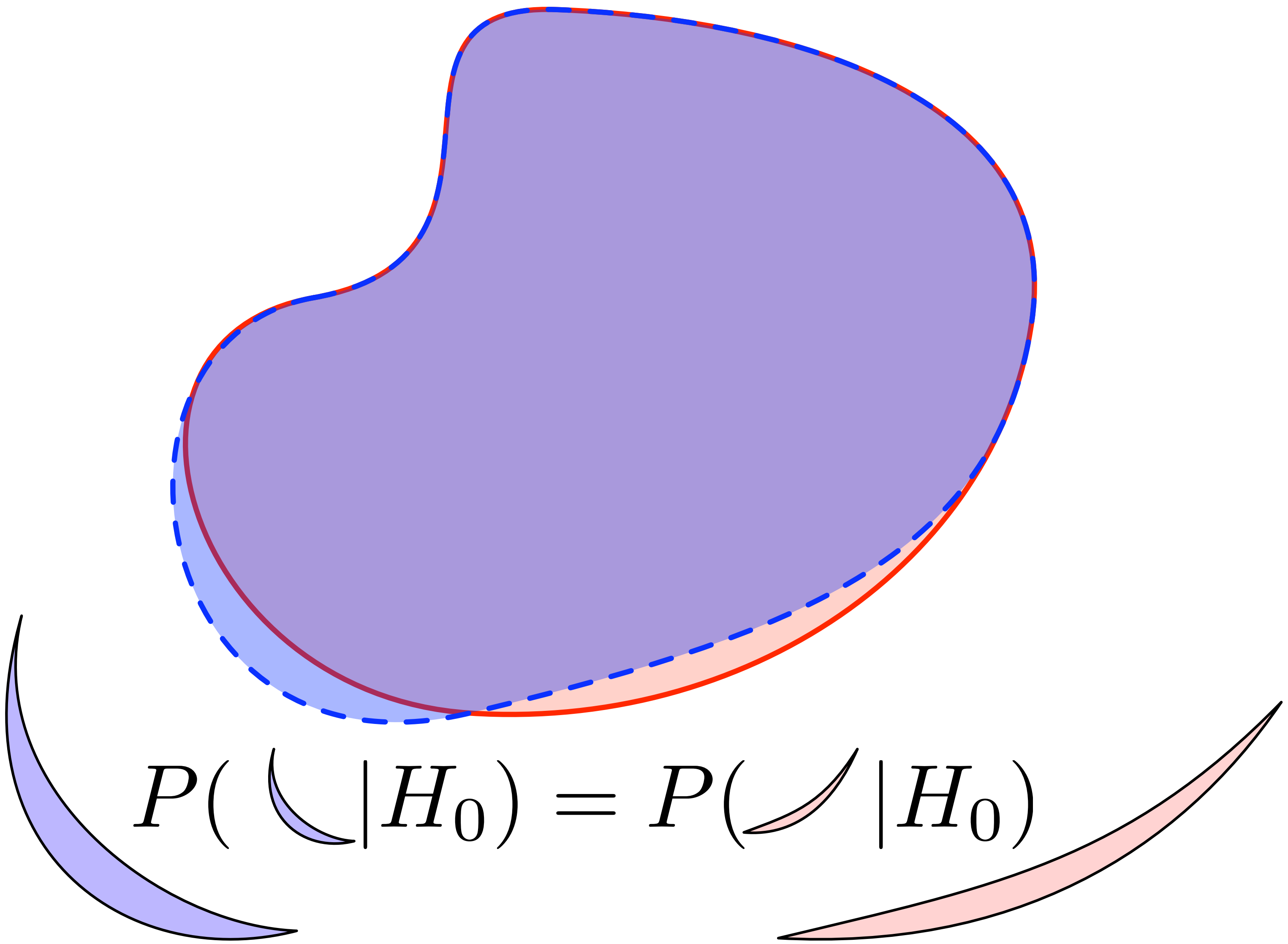




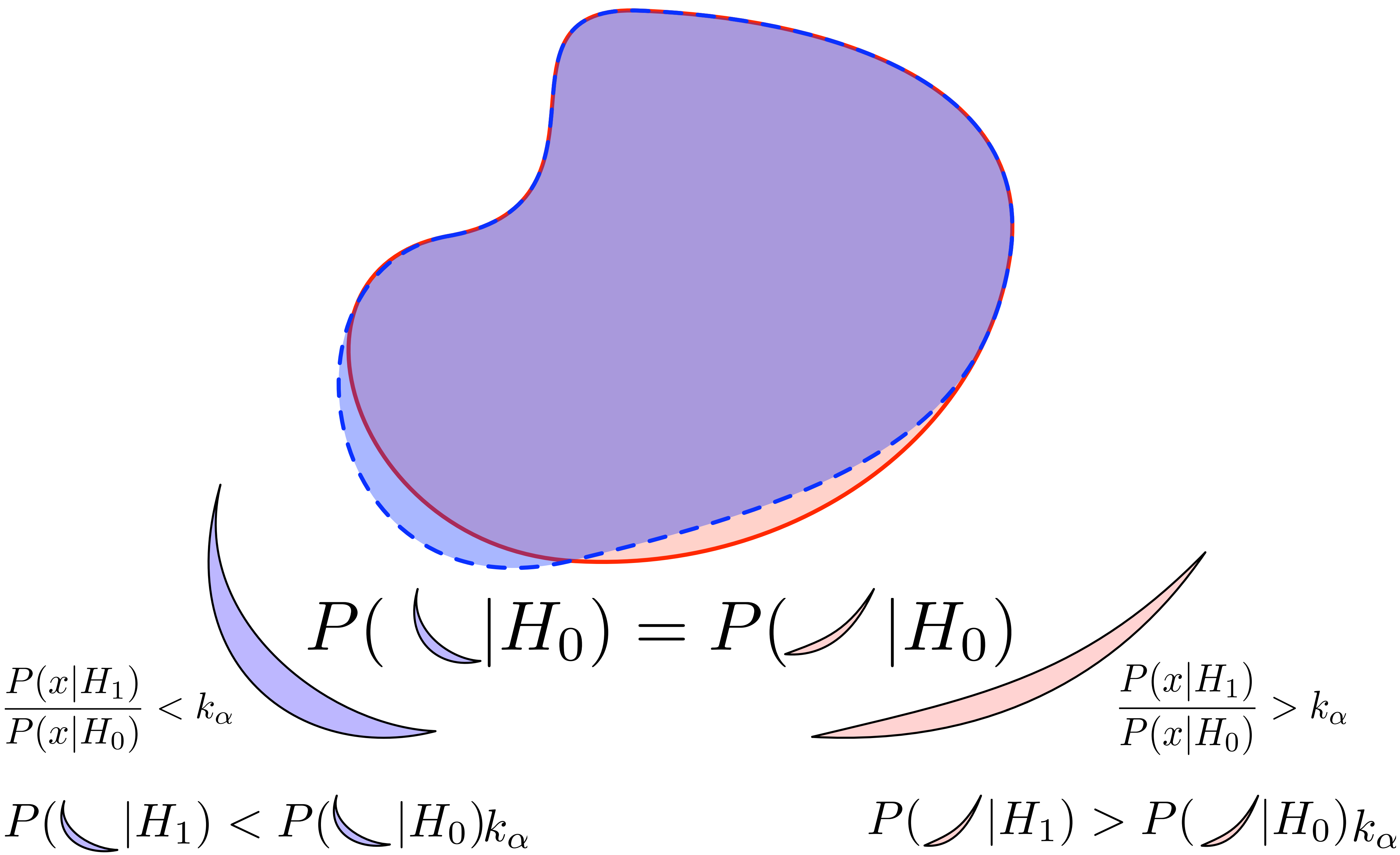
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is 1-α)



Now consider a variation on the contour that has the same size



Now consider a variation on the contour that has the same size (eg. same probability under H_0)



And for the region we lost, we also have an inequality

Together they give...

the p-value

- With p-value, we express the level of agreement b/w data and H

  $p =$ probabilty, under assumption of H, to observe data with equal or lesser compatibility with H, in comparison to the data we obtained

  $\neq$ *the probability that H is true*

- In frequentist statistics, we don't talk about $P(H)$.

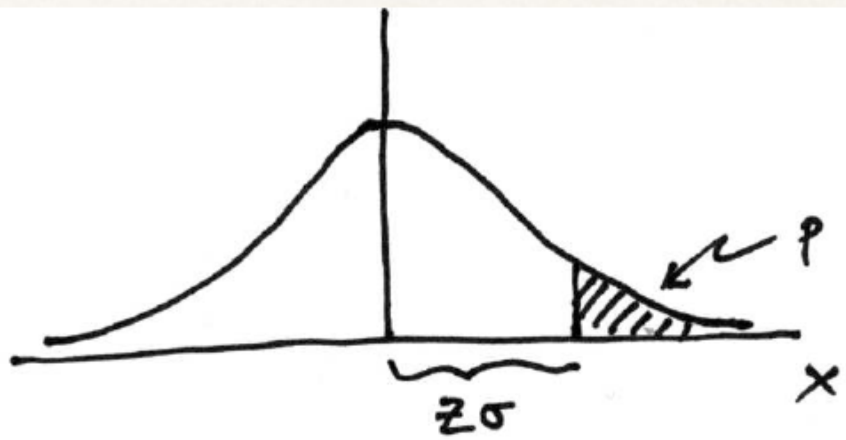
  In Bayesian, however, we determine $P(H|\vec{x})$ using the Bayes' theorem

  $\Leftarrow$ depending on the prior probabilty $\pi(H)$

- For now, we stick with the frequentist interpretation of the p-value

Significance from the *p*-value

- **Often we quote the significance Z, for a given p-value**
  - Z = the number of standard dev. that a Gaussian random variable would fluctuate in one direction to give the same *p*-value



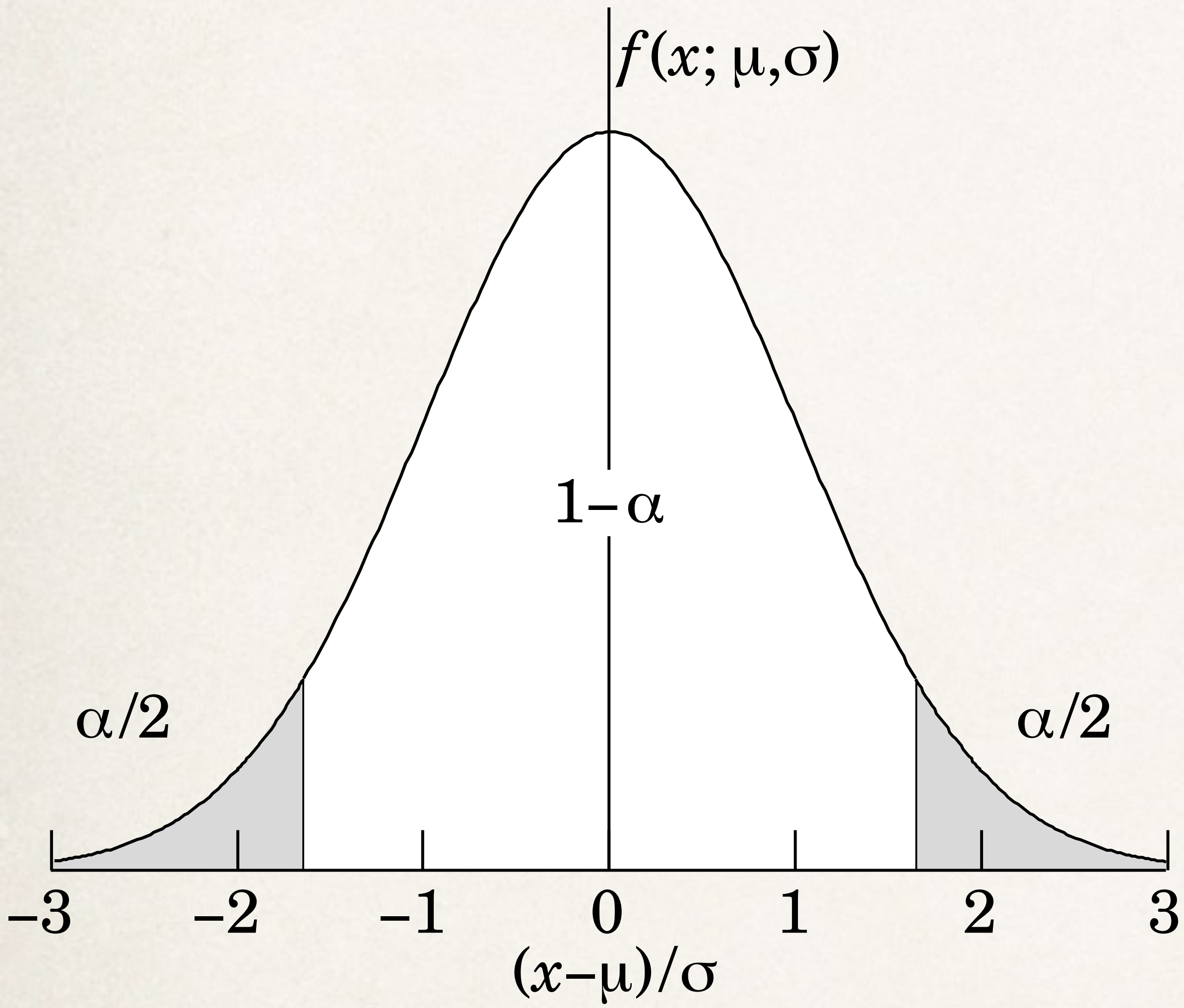
$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

(Ex) *Z* = 5 (a "5-sigma effect") ⇔ *p* = 2.9 x 10^{-7}

Remember?

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.

(Ex) $Z = 5$ (a "5-sigma effect") $\Leftrightarrow$ $p = 2.9 \times 10^{-7}$

p-value example: testing whether a coin is 'fair'

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \; .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_\mathrm{s} + n_\mathrm{b}$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Suppose $b = 0.5$, and we observe $n_\mathrm{obs} = 5$. Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_\mathrm{s} + n_\mathrm{b}$ is also Poisson, mean $= s + b$:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Suppose $b = 0.5$, and we observe $n_\mathrm{obs} = 5$. Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

Quiz

1983 프로야구 챔피언 해태 타이거즈 선발 타순

			타율	출루율	장타율	홈런	타점	도루
1	김일권	CF	.275	.345	.364	6	26	48
2	서정환	SS	.257	.320	.339	3	34	13
3	김성한	1B	.327	.401	.448	7	40	13
4	김봉연	DH	.280	.371	.552	22	59	2
5	김종모	LF	.350	.404	.524	11	44	7
6	김준환	RF	.248	.308	.362	10	43	11
7	김무종	C	.262	.313	.453	12	60	2
8	양승호	3B	.236	.292	.309	2	11	3
9	차영화	2B	.266	.308	.323	1	23	16

- (observation) Six out of 9 starting hitters have family name ‘Kim’.
- (fact) According to census, ~20% of all Koreans have family name ‘Kim’.
- (Hypothesis to test) The manager of 1983 Tigers (himself a ‘Kim’) has a bias toward players with family name ‘Kim’.

Model-independent test?

- In general, we cannot find a single critical region that gives the maximum power for all possible alternatives (no “uniformly most powerful” test)
- In HEP, we often try to construct a test of the Standard Model as H_0 (or sometimes called “background only”)

  such that we have a well specified *false discovery rate* α (=prob. to reject H_0 when it is true),

  and high power w.r.t. some interesting alternative H_1, e.g. SUSY, Z', etc.
- But, there is no such thing as a *model-independent* test.

  Any statistical test will inevitably have high power w.r.t. some alternatives and less for others

Intervals

Measurement with errors

- **Let's say we are doing a single measurement**

$$x = a \pm b$$

- **Frequentist interpretation**
  - Repeating the measurement many times under identical conditions ("ensemble"), in 68.3% of those results, the true value of x will lie between *a* - *b* and *a* + *b*
- **Result of each measurement is a sampling from a Gaussian distribution with mean μ and width σ**
  - We may not know μ
  - We have some idea about σ -- experimental sensitivity

when $\mu \pm \sigma$ is not enough…

If the PDF of the estimator is not Gaussian, or
if there are physical boundaries on the possible values of the parameter,
one usually quotes an interval given a confidence level.

Confidence interval from inversion of a test

- Suppose a model contains a parameter μ

  Which values are consistent with data and which disfavored?

- Carry out a test of size α for all values of μ.

  $\Rightarrow$ The values that are *not rejected* constitutes a **confidence interval** for μ at confidence level $\mathrm{CL} = 1 - \alpha$.

- Probability of rejecting true value of α is $\leq \alpha$

  $\therefore$ by construction the confidence interval will contain the true value of μ with probability $\geq 1 - \alpha$.

  * The interval depends on the choice of the test (critical region).
  * If the test is formulated in terms of a p-value, p_μ, then the confidence interval represents those values of μ for which $p_\mu > \alpha$.
  * To find the end points of the interval, set $p_\mu = \alpha$ and solve for μ.

a Bayesian procedure for intervals

$$1 - \alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta$$

If the physical value is non-negative, one may choose a prior:

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases}$$

Likelihood for s, given b, is

$$P(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

a Bayesian procedure for intervals

$$1 - \alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta$$

If the physical value is non-negative, one may choose a prior:

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases}$$

Likelihood for s, given b, is

$$P(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

If what we seek is of a very low (or no) signal, interval ➔ UL

Then,

$$1 - \alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} P(n|s)\, \pi(s)\, ds}{\int_{-\infty}^{\infty} P(n|s)\, \pi(s)\, ds}$$

$F^{-1}_{\chi^2}$: inverse of the CDF

➔ $$s_{\rm up} = \frac{1}{2} F^{-1}_{\chi^2} \left[p, 2(n+1)\right] - b$$

(Ex) UL on Poisson parameter

- Consider again the case of observing $n \sim \text{Poisson}(s+b)$. Suppose $b = 4.5$ and $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.
- Relevant alternative is $s = 0$, resulting in critical region at low n.
- The p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$.

  Therefore, the upper limit s_{up} at $\text{CL} = 1 - \alpha$ is obtained from

$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(0.95; 2(5+1)) - 4.5 = 6.0$$

Frequentist "confidence intervals"

- on repeated measurements

  Remember frequentist approach is always about repeated measurements!

- "confidence interval"

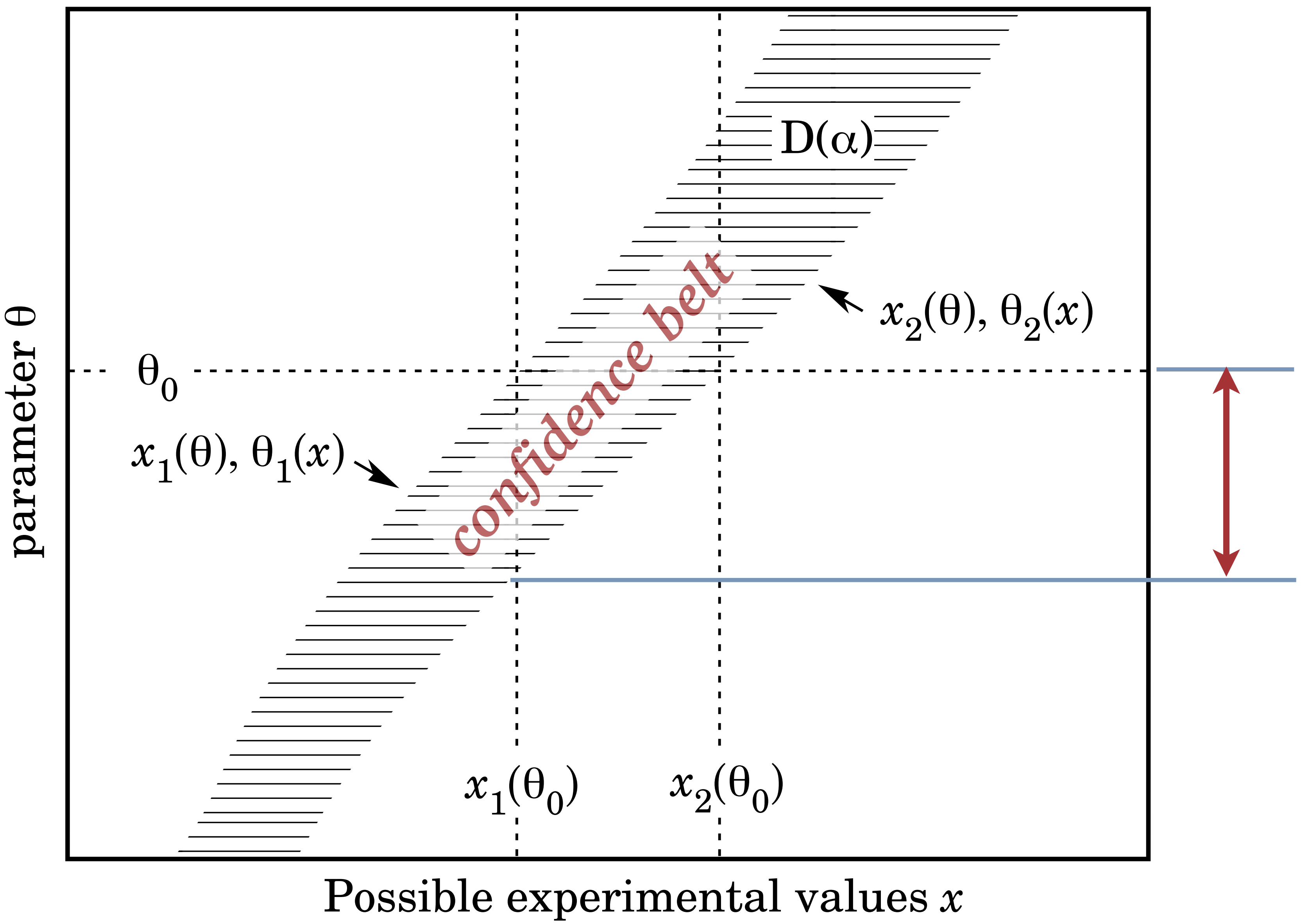
  = intervals constructed to include the true value of the parameter with a probability ≥ (*a specified value*)

Frequentist "confidence intervals"

- **Consider a pdf $f(x;\theta)$**

$$P(x_1 < x < x_2; \theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x;\theta)\, dx$$

  - x : outcome of an experiment
  - θ : unknown parameter for which we set the interval



the confidence interval, given a measured outcome of $x_0 = x_1(\theta_0)$

Coincidence of frequentist and Bayesian intervals

- If the expected background is zero, the Bayesian upper limit (for a Poisson RV) becomes equal to the limit determined by frequentist approach.

$$s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}\left[p, 2(n+1)\right] - b$$
$$= \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2(n+1))$$

For more details, you may read e.g. a statistics review in PDG.

http://pdg.lbl.gov/2013/reviews/rpp2013-rev-statistics.pdf

Parameter Estimation

Basics of parameter estimation

- The parameters of a PDF are constants characterizing its shape, e.g.

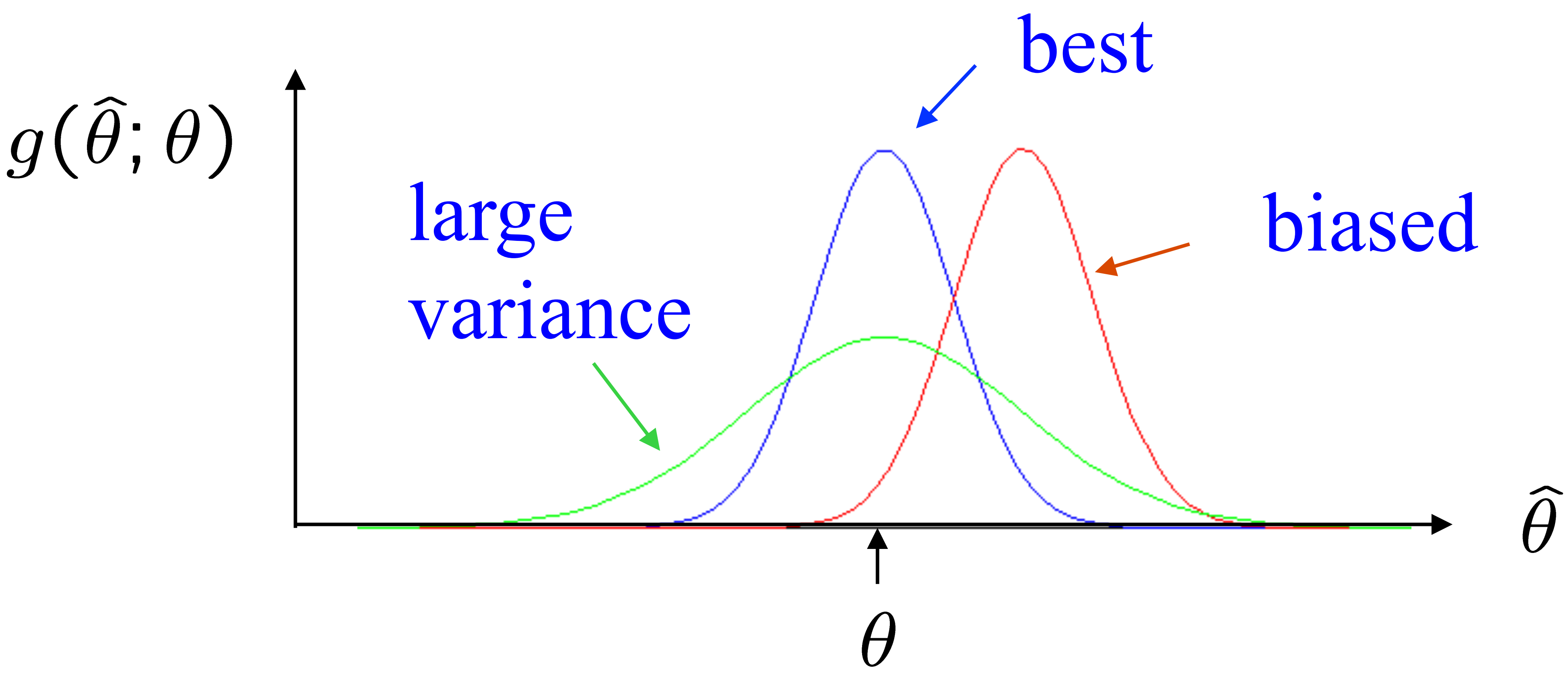
$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

  where θ is the parameter, while x is the random variable.

- Suppose we have a sample of observed values, $\vec{x}$.
  We want to find some function of the data to *estimate* the parameter(s): $\hat{\theta}(\vec{x})$.
  Often $\hat{\theta}$ is called an **estimator**.

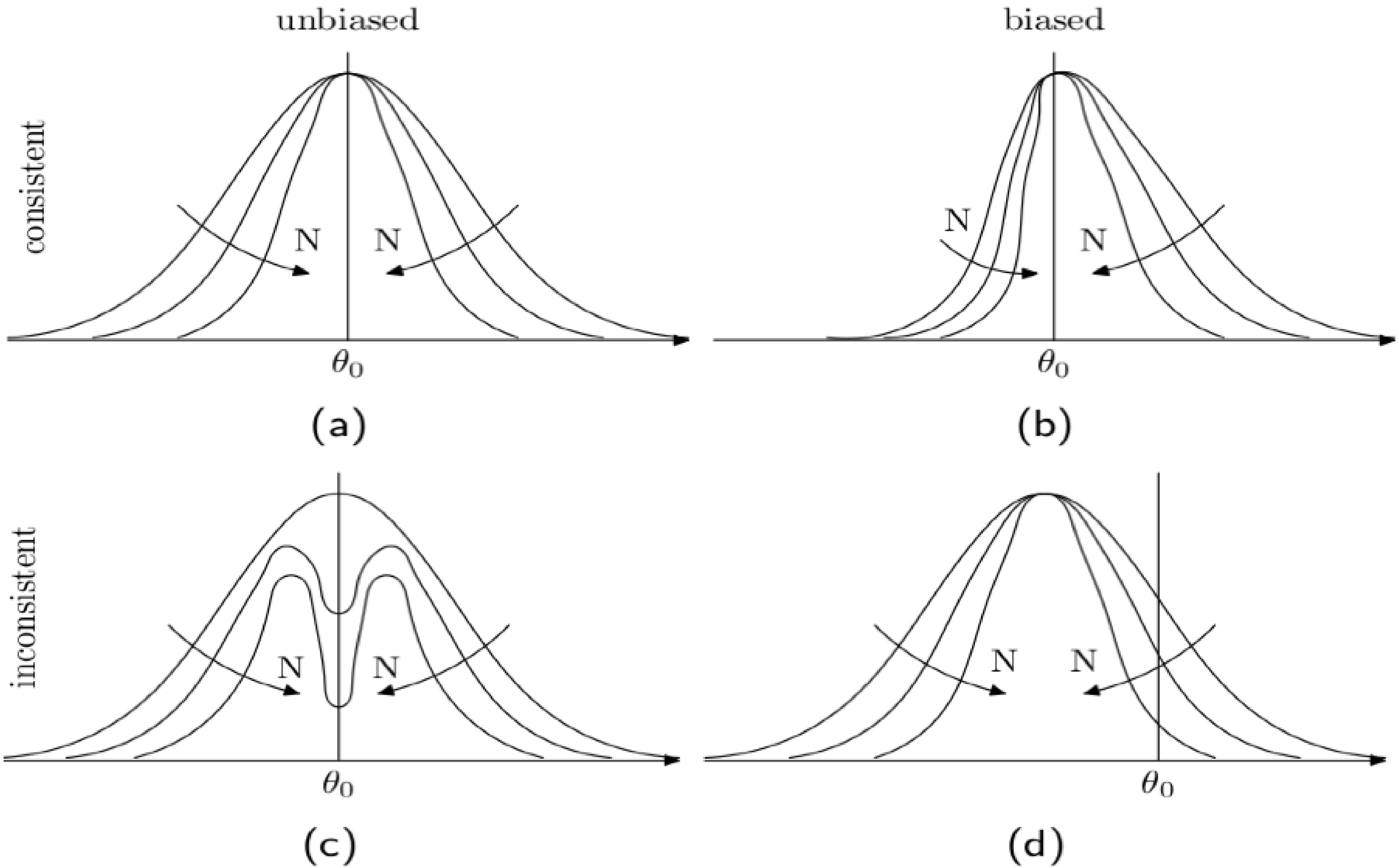
Properties of estimators

- If we were to repeat the entire measurement, the set of estimates would follow a PDF:



  - We want small (or zero) bias ($\Rightarrow$ syst. error): $b = E[\hat{\theta}] - \theta$
  - and we want a small variance ($\Rightarrow$ stat. error): $V[\hat{\theta}]$

Bias vs. Consistency



The likelihood function

- Suppose the entire result of an experiment (*set of measurements*) is a collection of numbers $\vec{x}$, and suppose the joint PDF for the data $\vec{x}$ is a function depending on a set of parameters $\vec{\theta}$: $f(\vec{x};\vec{\theta})$
- Evaluate this function with the measured data $\vec{x}$, regarding this as a function of $\vec{\theta}$ only. This is the **likelihood function**.

$$L(\vec{\theta}) = f(\vec{x};\vec{\theta}) \ (\vec{x}, \text{fixed})$$

The likelihood function for i.i.d. data

i.i.d. = *independent and identically distributed*

- Consider n independent observations of x: $x_1, \cdots, x_n$, where x follows $f(x, \theta)$. The joint PDF for the whole data sample is:

$$f(x_1, \cdots, x_n; \vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

- In this case, the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \quad (x_i \text{ constant})$$

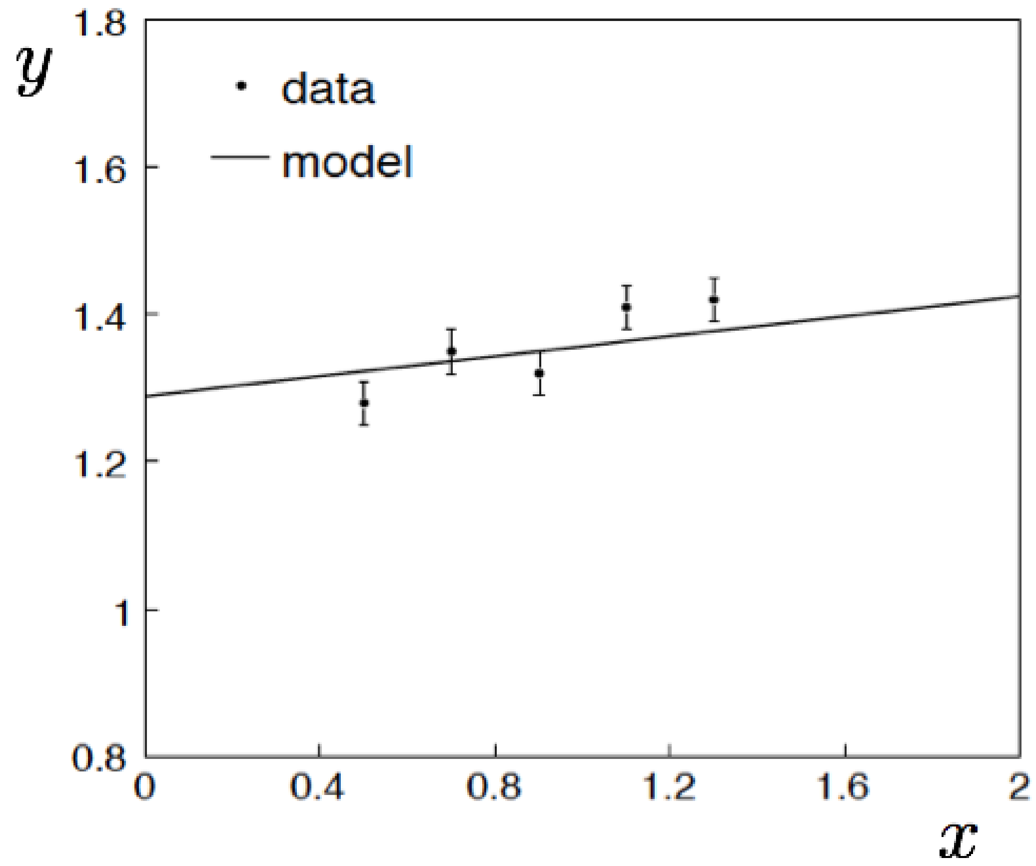
So we define the max. likelihood (ML) estimator(s) to be the parameter value(s) for which the L becomes maximum.

ML estimator example: fitting to a straight line

- Suppose we have a set of data:
  $(x_i, y_i, \sigma_i),\ i = 1, \cdots, n.$
- Modeling: y_i are independent and follow $y_i \sim G(\mu(x_i), \sigma_i)$ (G: Gaussian) where $\mu(x_i)$ are modelled as $\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x$

  Assume x_i and σ_i are known.
- Goal: to estimate θ_0

  Here, let's suppose we don't care about θ_1 (an example of a *nuisance parameter*)



ML fit with Gaussian data

- In this example, the y_i are assumed independent, so that likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right]$$

- Then maximizing L is equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2\ln L(\theta_0, \theta_1) + C = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}$$

i.e., for Gaussian data, ML fitting is the same as the method of least squares

Wilk's theorem

ML fit or Least-square fit?

- **Consider we have a random variable $x \in [0, 3]$, and a distribution $f(x)$.**
- **In a series of measurements, we obtained**
  - 9 events in [0,1), 10 events in [1,2), and 8 events in [2,3]
  - We have a model of uniform $f(x)$, and would like to estimate the mean value of $\int f(x)\,\mathrm{d}x$ for each histogram bin.
- **Run a thought-experiment, comparing**
  - maximum likelihood method, and least-square method
  - *Do they give the same result?*

Bayesian likelihood function

- Suppose our L-function contains two parameters θ_0 and θ_1, where we have some knoweldege about the prior probability on θ_1 from previous measurements:

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1)$$
$$\pi_0(\theta_0) = \text{const.}$$
$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\theta_1-\theta_p)^2/2\sigma_p^2}$$

- Putting this into the Bayes' theorem gives the posterior probability:

$$p(\theta_0, \theta_1|\vec{x}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i-\mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_p} e^{-(\theta_1-\theta_p)^2/2\sigma_p^2}$$

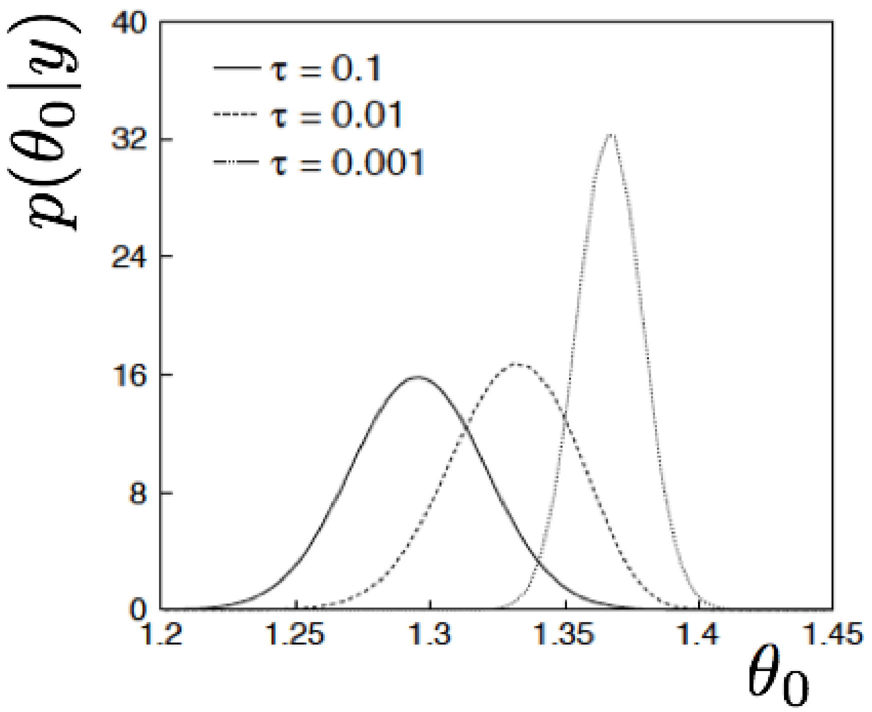
- Then, $p(\theta_0|\vec{x}) = \int p(\theta_0, \theta_1|\vec{x})\, d\theta_1$

with alternative priors

- Suppose we don't have a previous measurement of θ_1 but rather a theorist saying that θ_1 should be > 0 and not too much greater than, say, 0.1 or so. In that case, we may try modeling the prior for θ_1 as something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \; \theta_1 \geq 0, \; \tau = 0.1$$

- From this we obtain (numerically) the posterior PDF for θ_0



- This plot summarizes all knowledge about θ_0.

other advanced topics

- nuisance parameters & systematic uncertainties
- spurious exclusion ➔ the CL_s procedure
- look-elsewhere effect

Systematic uncertainties?

In statistics, they call it the "nuisance parameter"

All | Dictionary | Thesaurus | Apple | Wikipedia

nui•sance |ˈn(y)oōsəns|

noun

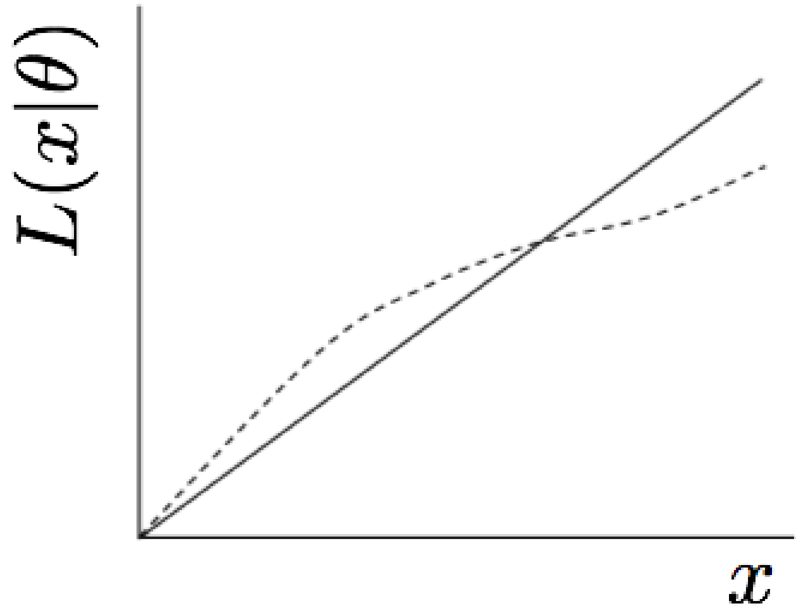
a person, thing, or circumstance causing inconvenience or annoyance : *an unreasonable landlord could become a nuisance | I hope you're not going to make a nuisance of yourself.*

- (also **private nuisance**) Law an unlawful interference with the use and enjoyment of a person's land.
- Law see **PUBLIC NUISANCE** .

ORIGIN late Middle English (in the sense [injury, hurt]): from Old French, 'hurt,' from the verb ***nuire***, from Latin ***nocere*** ***'to harm.'***

Nuisance parameters

- In general our model of the data is *not perfect*



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

- can improve model by including additional adjustable parameters: $L(x|\theta) \rightarrow L(x|\theta, \nu)$
- Nuisance parameter $\leftrightarrow$ systematic uncertainty

  Some point in the parameter space of the enlarged model must be "true"
- Presence of nuisance parameter(s) decreases sensitivity of analysis to the parameter of interest (e.g. larger variance of estimate).

p-values with nuisance parameters

- Suppose we have a statistic q to test a hypothesized value of a parameter θ, such that the p-value of θ is

$$p_\theta = \int_{q_\theta,\mathrm{obs}}^{\infty} f(q_\theta|\theta,\nu)\, dq_\theta$$

- But what value of ν should we use for $f(q_\theta|\theta,\nu)$?
- In the large-sample limit, $f(q_\theta|\theta,\nu)$ becomes independent of the nuisance parameters – a feature of statistics based on the profile likelihood ratio
- But in general for finite sample this is not true.
- One may therefore be unable to reject some θ values if all values of ν shall be considered. (Interval for θ "overcovers").

The profile likelihood ratio

- Base significance test on the profile likelihood ratio

profile likelihood

maximizes L for Specified μ

$$\lambda(\mu) = \frac{L_\mathrm{p}(\mu)}{L_\mathrm{max}} = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}$$

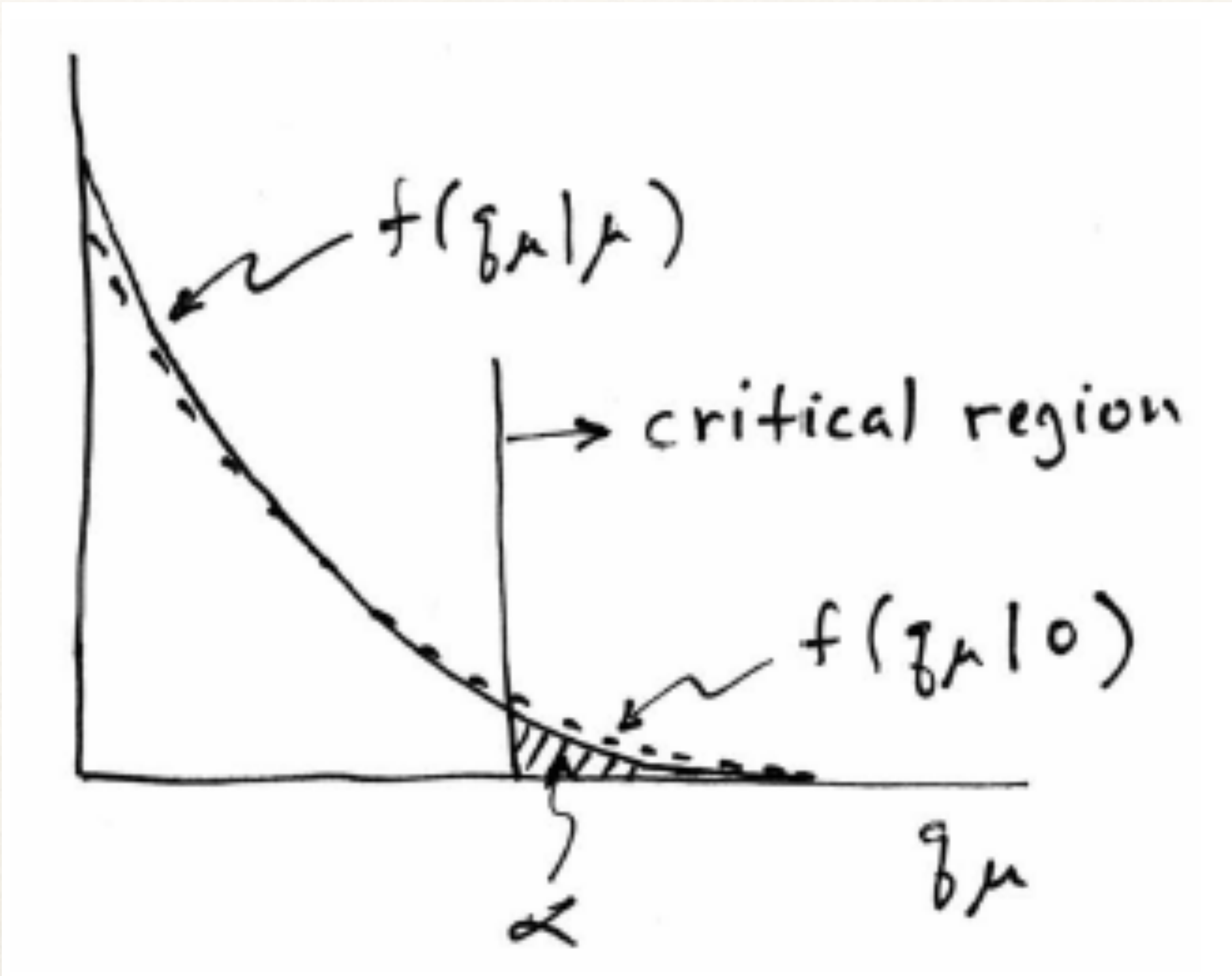
maximize L

  - the likelihood ratio of point hypotheses gives optimal test (by Neyman-Pearson lemma)
  - the statistic above is nearly optimal

- Advantage of $\lambda(\mu)$ – in large sample limit, $f(-2\ln\lambda(\mu)\,|\,\mu)$ approaches a χ^2 pdf for $n = 1$ (by *Wilk's theorem*)

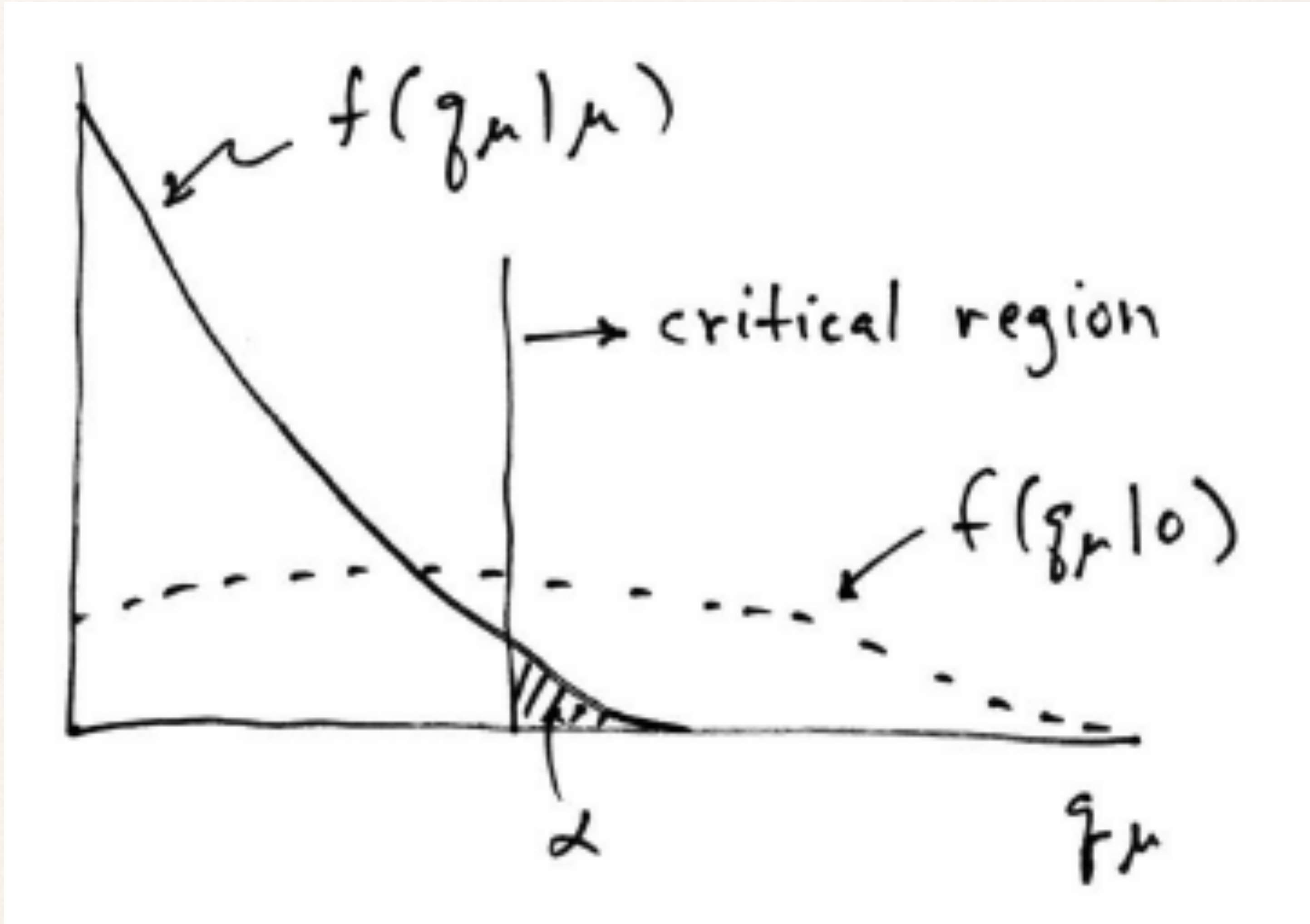
low sensitivity & spurious exclusion

- **Sometimes, the effect of a given hypothesized μ is very small relative to the null (μ =0) prediction**
  - This means that the distributions $f(q_\mu \mid \mu)$ and $f(q_\mu \mid 0)$ will be almost the same.



low sensitivity & spurious exclusion

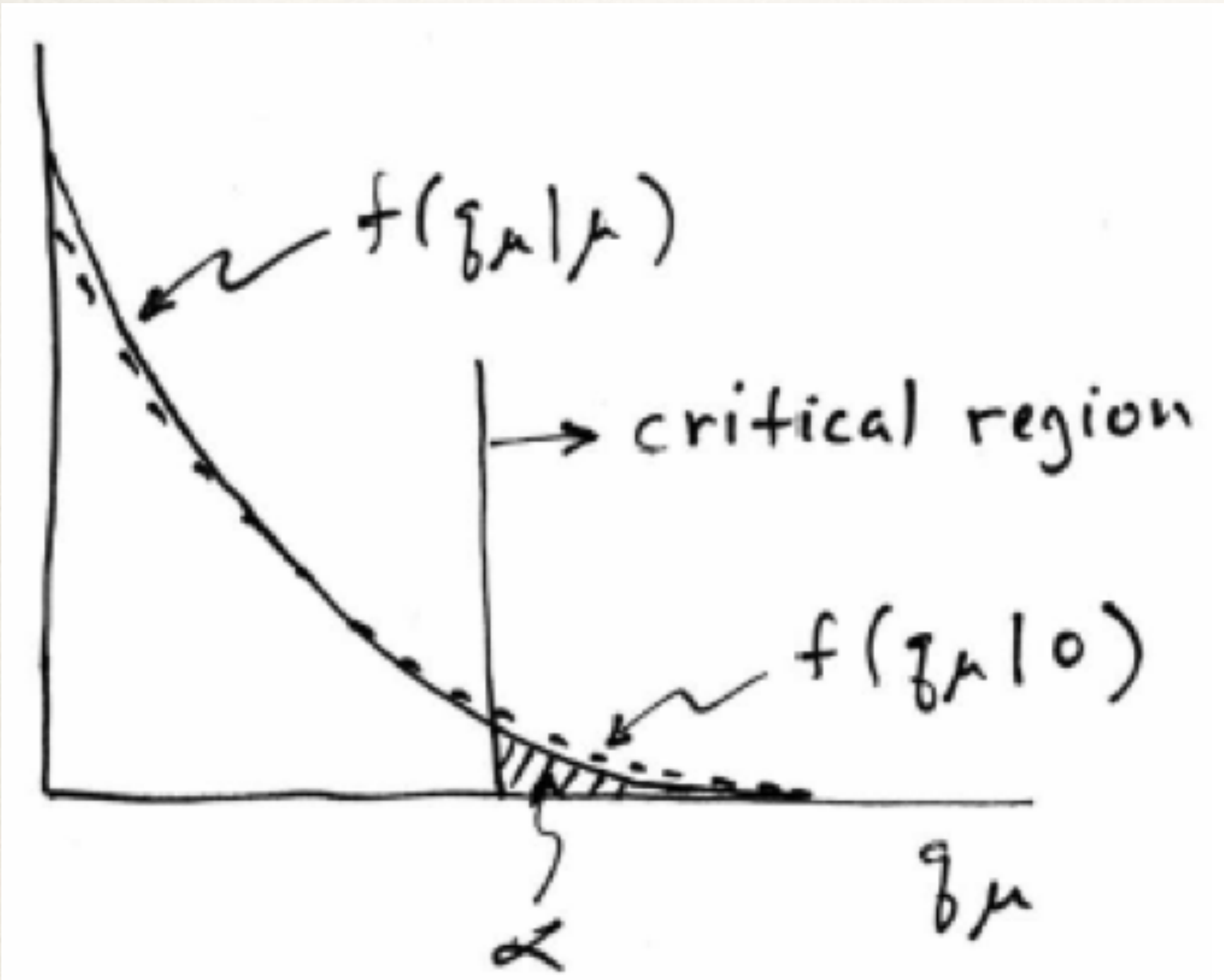
- **In contrast, for a high-sensitivity test, the two pdf's --** $f\,(q_\mu\,|\,\mu)$ and $f\,(q_\mu\,|\,0)$ **-- are well separated**



In this case, the power is substantially higher than 1–α.
Use this 'power' as a measure of the sensitivity.

low sensitivity & spurious exclusion

Consider again the case of low-sensitivity



- This means that one excludes hypotheses to which one has essentially no sensitivity (e.g. m_{H} = 1000 TeV)
- It is called the "**spurious exclusion**"

spurious = not being what it claims to be

low sensitivity & spurious exclusion

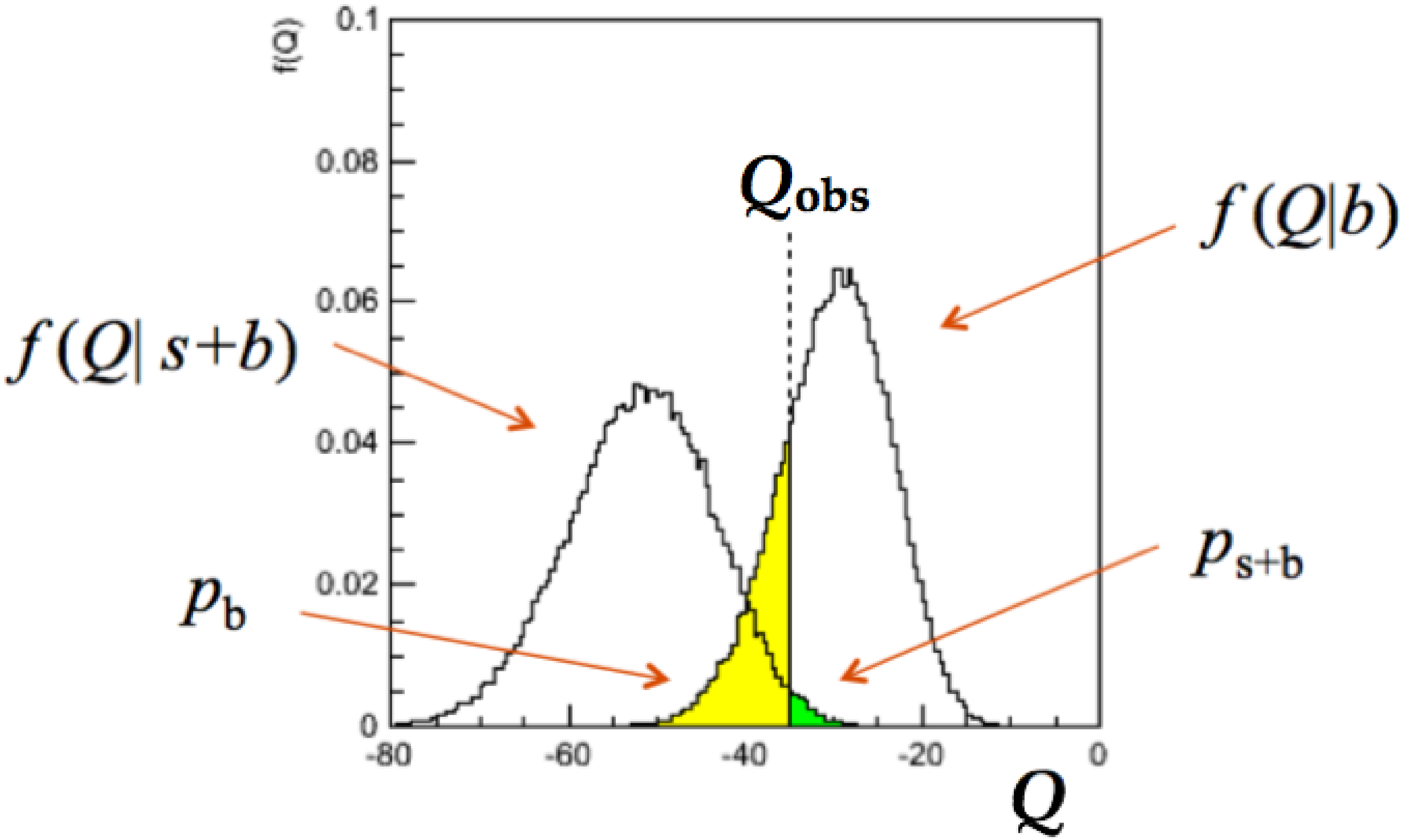
- The problem of excluding values to which one has no sensitivity is known for a long time
- In the 1990s this problem was re-examined for the LEP Higgs search, e.g.

  T. Junk, NIM A 434, 435 (1999); A.L. Read, J. Phys. G 28, 2693 (2002).

  and led to the "CL_s" procedure for upper limits

The CL$_s$ procedure

- In the CL$_s$ formulation, one tests both the $\mu = 0$ (b) and $\mu > 0$ $(s+b)$ hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$

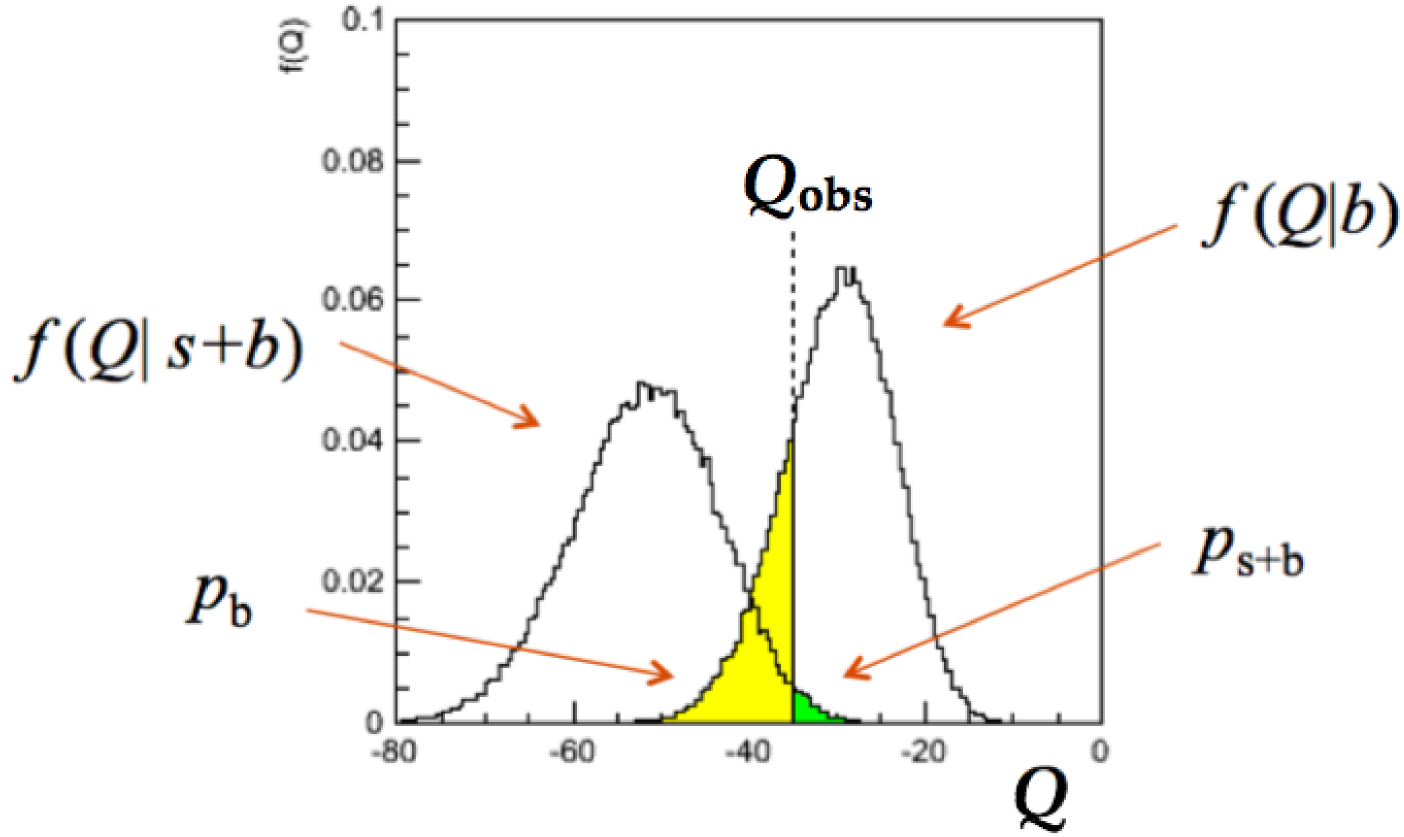


The CL$_s$ procedure

- The CL$_s$ prescription is to base the test on the usual p-value (CL_{s+b}), but rather to divide this by $\mathrm{CL}_b (= 1 - p_b)$

$$\mathrm{CL}_s \equiv \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} = \frac{p_{s+b}}{1 - p_b}$$

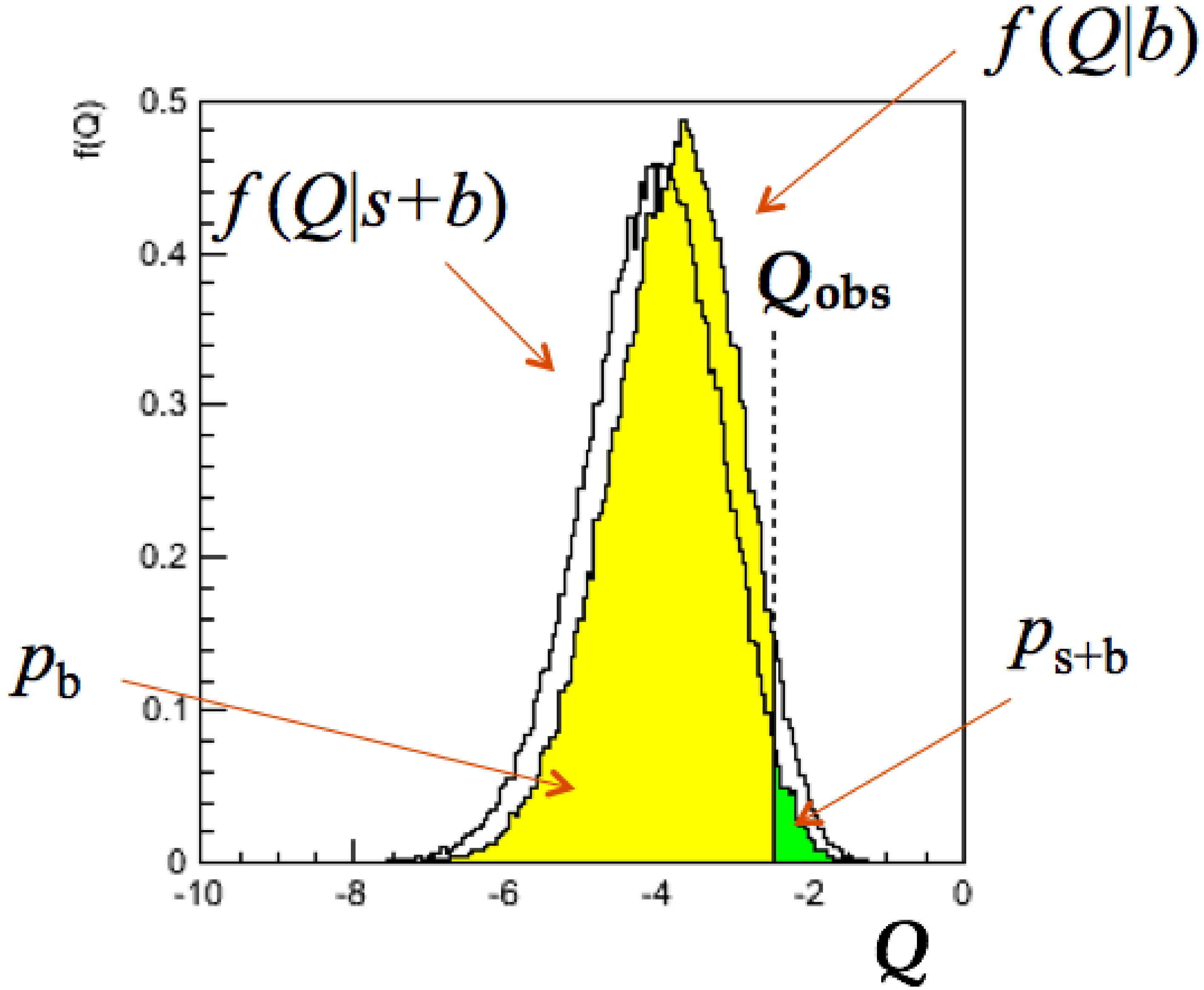
- Reject $s + b$ hypothesis if $\mathrm{CL}_s < \alpha$
- Reduces "effective" p-value when the two distributions become close, thus preventing exclusion if sensitivity is low



The CL_s procedure

$$\mathrm{CL}_s \equiv \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} = \frac{p_{s+b}}{1 - p_b}$$

- Reject $s + b$ hypothesis if $\mathrm{CL}_s < \alpha$
- Reduces “effective” p-value when the two distributions become close, thus preventing exclusion if sensitivity is low



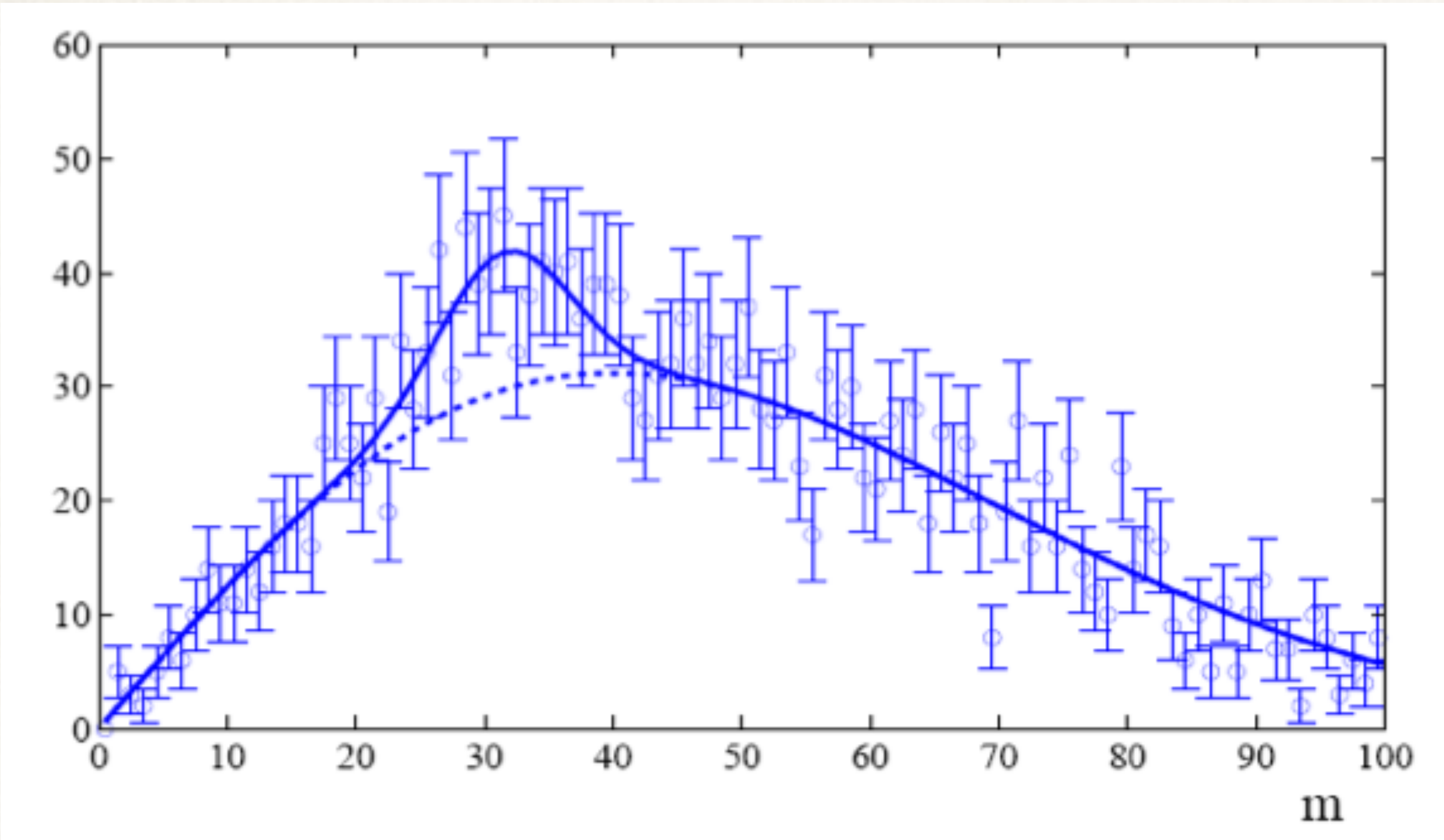
the Look Elsewhere Effect

consider…

- **Suppose you throw a coin 10 times, and you've got 10 heads, zero tails.**
  - It's very unusual.
  - Can you quantify how unusual this result is?
- **In particular, can you say the probability for this kind of peculiarity happening is $1/10^{24}$?**
  - No! Think why!
- **What must then be the correct answer?**

Look-Elsewhere Effect

- **Suppose a model for a mass distribution allows for a peak at a mass *m* with amplitude μ**
- **and the data show a bump at a mass m_0**



How consistent is this with the no-bump (μ =0) hypothesis?

Local p-value

- First, suppose that the mass peak value m_0 was known a priori.
- Test consistency of bump with the $\mu = 0$ hypothesis with e.g. L-ratio

$$t_{\text{fix}} = -2\ln\left(\frac{L(0, m_0)}{L(\mu, m_0)}\right)$$

where "fix" indicates that the mass peak value is fixed to m_0.

- The resulting p-value

$$p_{\text{local}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0)\, dt_{\text{fix}}$$

gives the probability to find a value of t_{fix} at least as great as the observed value at the specific mass m_0, and is called the local p-value.

Global p-value

- Now, suppose we did not know where to expect a peak. In other words, the signal can be found at every value of m.
- What we want is the probability to find a peak at least as significant as the one observed **anywhere** in the distribution
- For this, include the mass as an *adjustable parameter* in the fit, then test significance of peak using

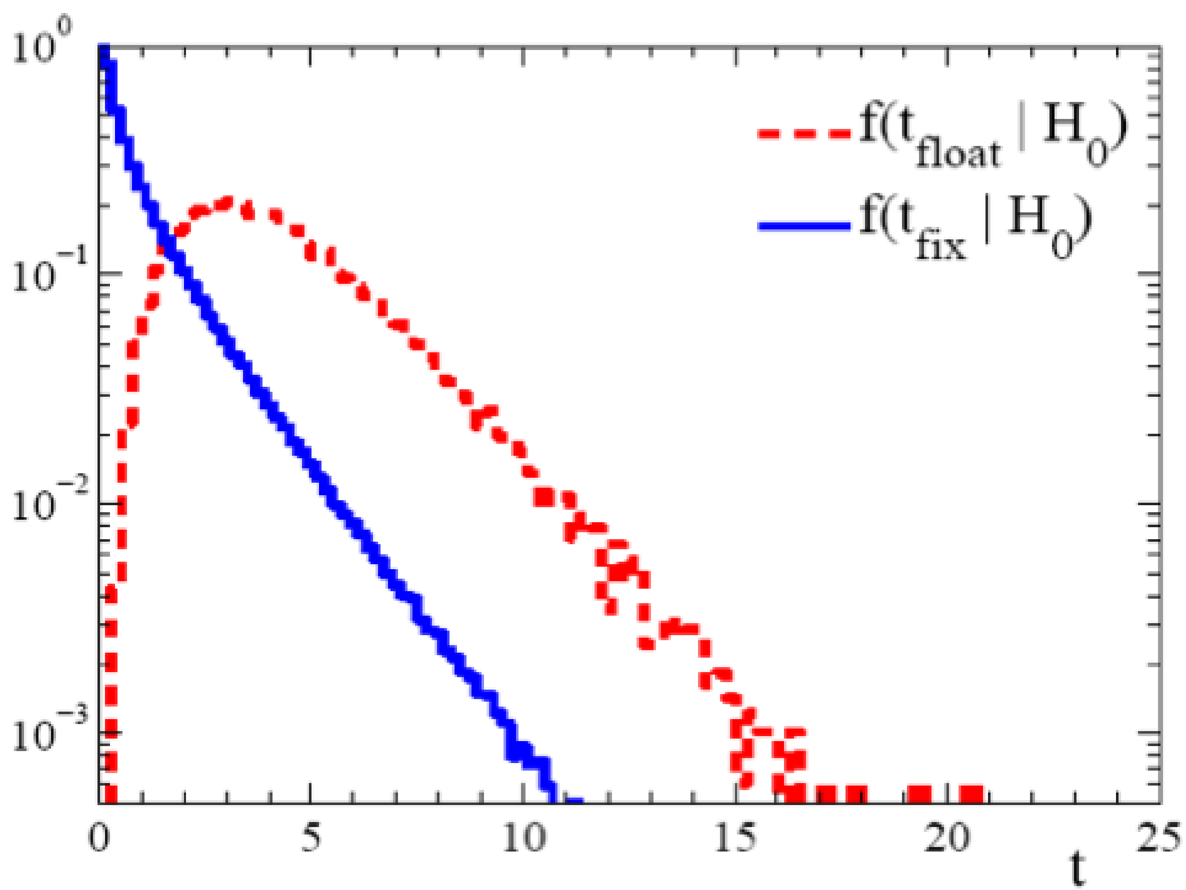
$$t_{\text{float}} = -2\ln\frac{L(0)}{L(\mu, m)}$$

Note: m does not appear in the μ=0 model

$$p_{\text{global}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0)\, dt_{\text{float}}$$

$t_{\rm fix}$ vs. $t_{\rm float}$

- For a sufficiently large data sample, $t_{\rm fix} \sim \chi^2$ for 1 deg. of freedom (*Wilk's theorem*)
- For $t_{\rm float}$ there are two adjustable parameters, μ and m, and naively Wilk's theorem says $t_{\rm float} \sim \chi^2$ for 2 d.o.f.



But, Wilk's theorem does not hold in the floating mass case because one of the parameters (m) is not defined in the $\mu = 0$ model.

$\therefore$ getting $t_{\rm float}$ distribution is more difficult.

Approximate correction for LEE

- Need to related the p-values for the fixed and floating-mass analyses (at least approximately)
- (Gross & Vitells) The p-values are approximately related by

$$p_{\text{global}} \approx p_{\text{local}} + \langle N(c) \rangle$$

where $\langle N(c) \rangle =$ mean # of *upcrossings* of $-2\ln L$ in the fit range based on a threshold

$$c = t_{\text{fix}} = Z_{\text{local}}^2$$

- We may carry out the full MC (time and CPU-consuming) or do fixed-m analysis and apply a correction factor (much faster!)

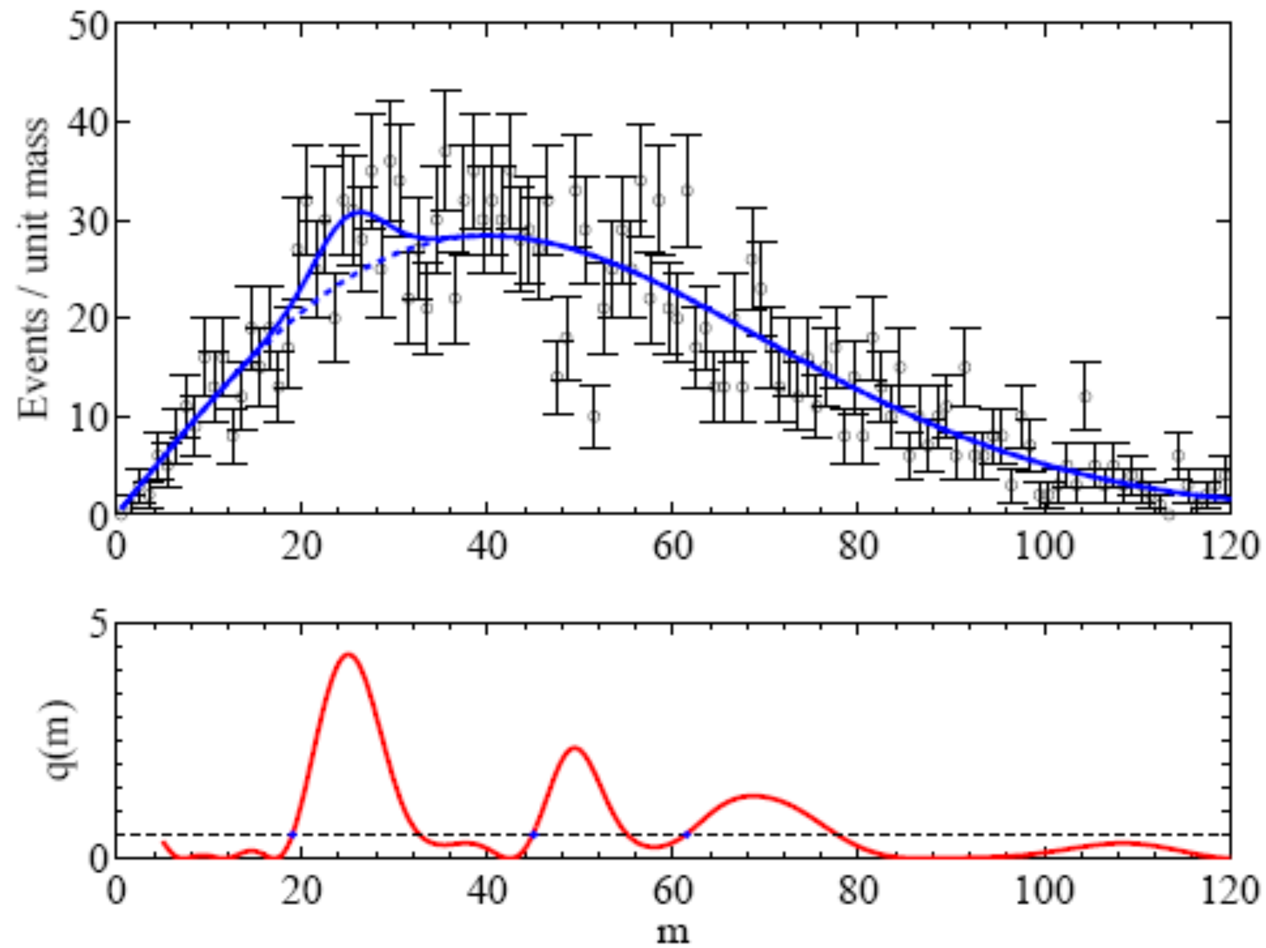
Up-crossings of $-2\ln L$

$p_{\text{global}} \approx p_{\text{local}} + \langle N(c) \rangle$ where $\langle N(c) \rangle =$ mean # of *upcrossings* of $-2\ln L$ in the fit range based on a threshold $c = t_{\text{fix}}$

- What is 'up-crossing'? How can we obtain this number?
- With high threshold c, you need a huge MC sample to estimate p_{global}.
- For an economic alternative, $\langle N(c) \rangle$ can be estimated from MC using a much lower threshold c_0:

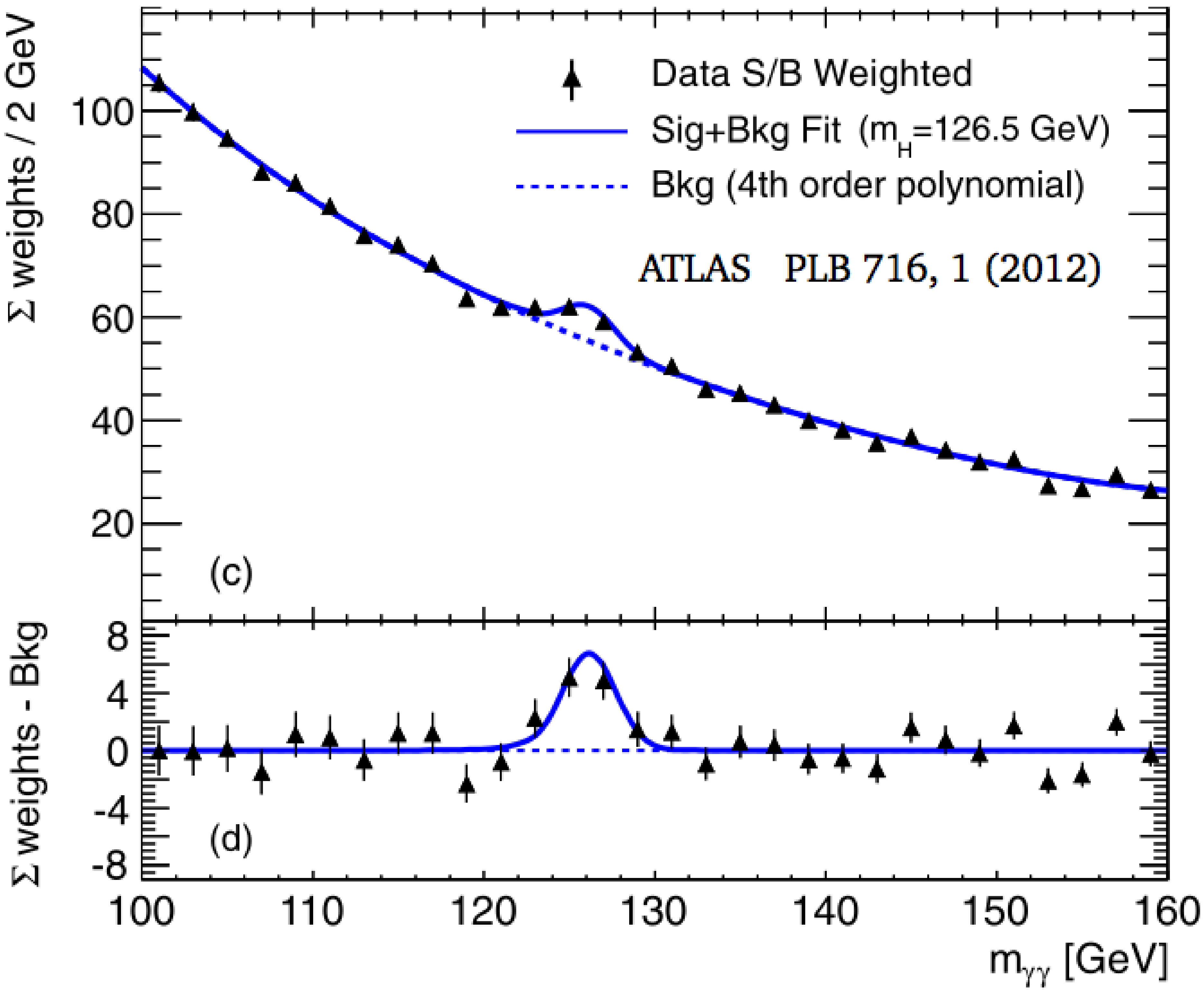
$$\langle N(c) \rangle \approx \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

so we don't need a huge computing resource



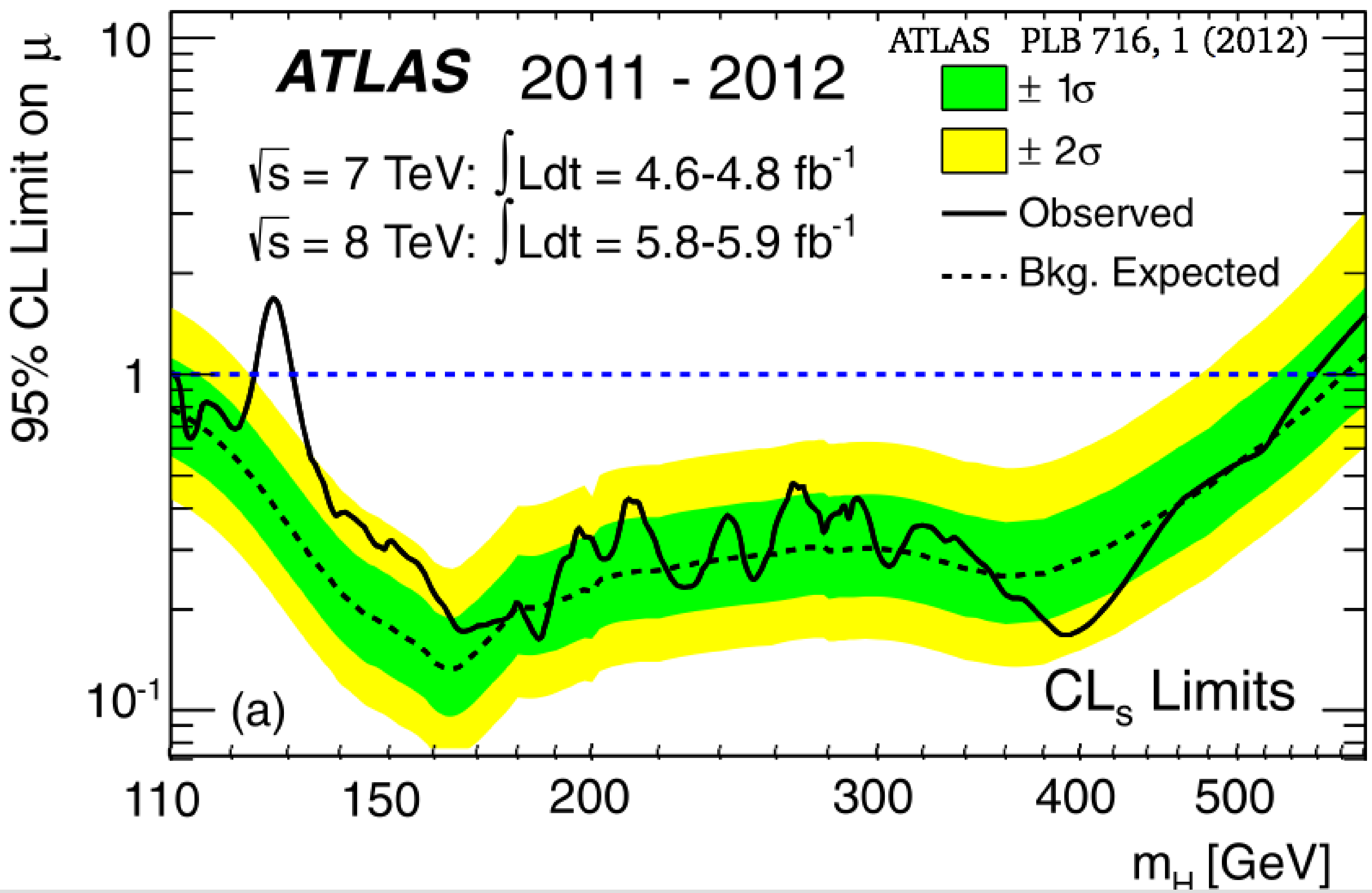
Examples to test what you've learned

what to make sense of m_H plots, statistically



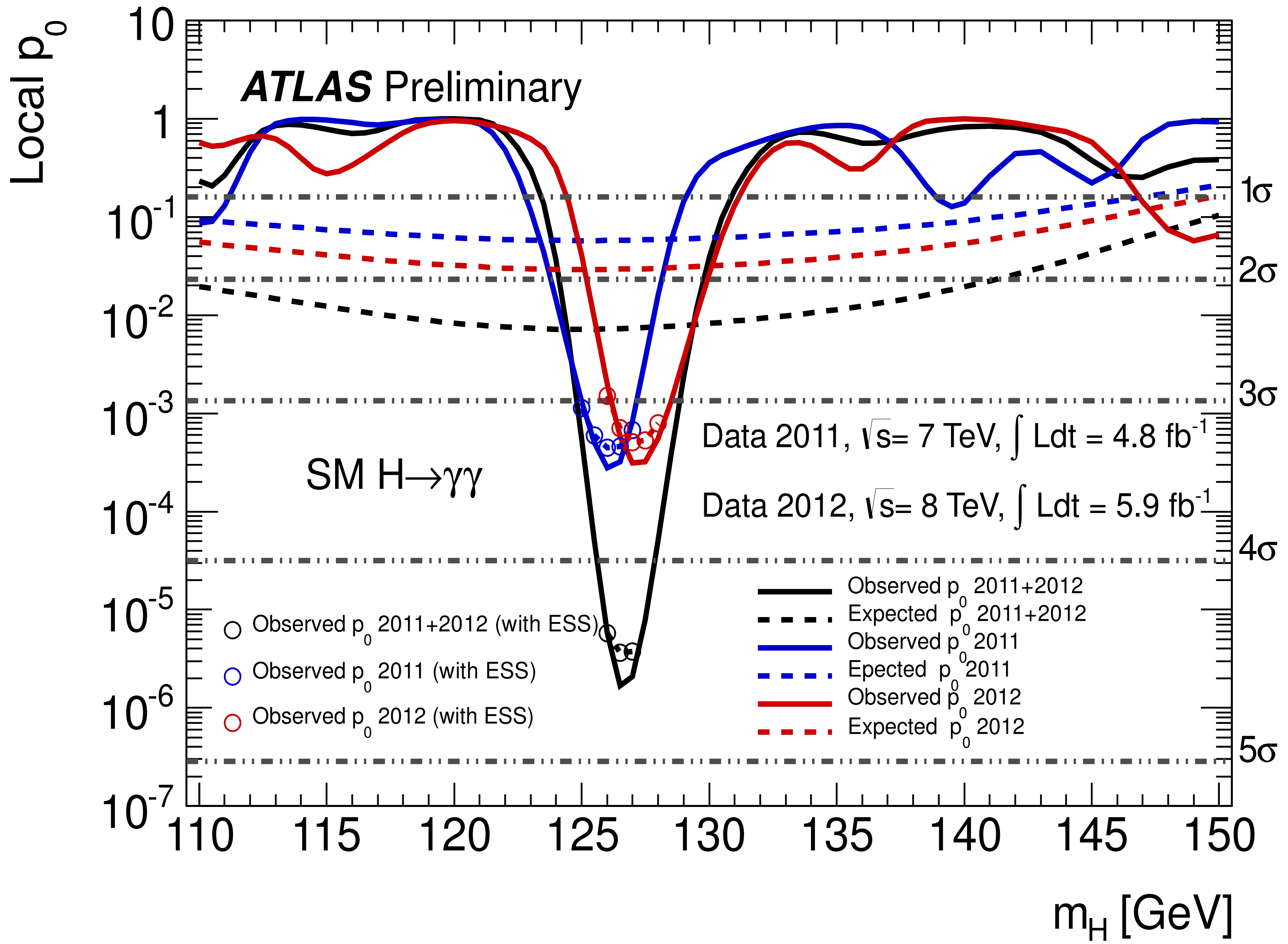
how to read the green & yellow plots

- For every (assumed) value of m_H, we want to find the CL$_s$ upper limit on $\mu \equiv \sigma(H)/\sigma_{\rm SM}(H)$ (*solid curve*)
- Also shown is the 'expected upper limit', determined for each assumed m_H value, under the assumption that we see no excess above background.



how to read the p_0 plots

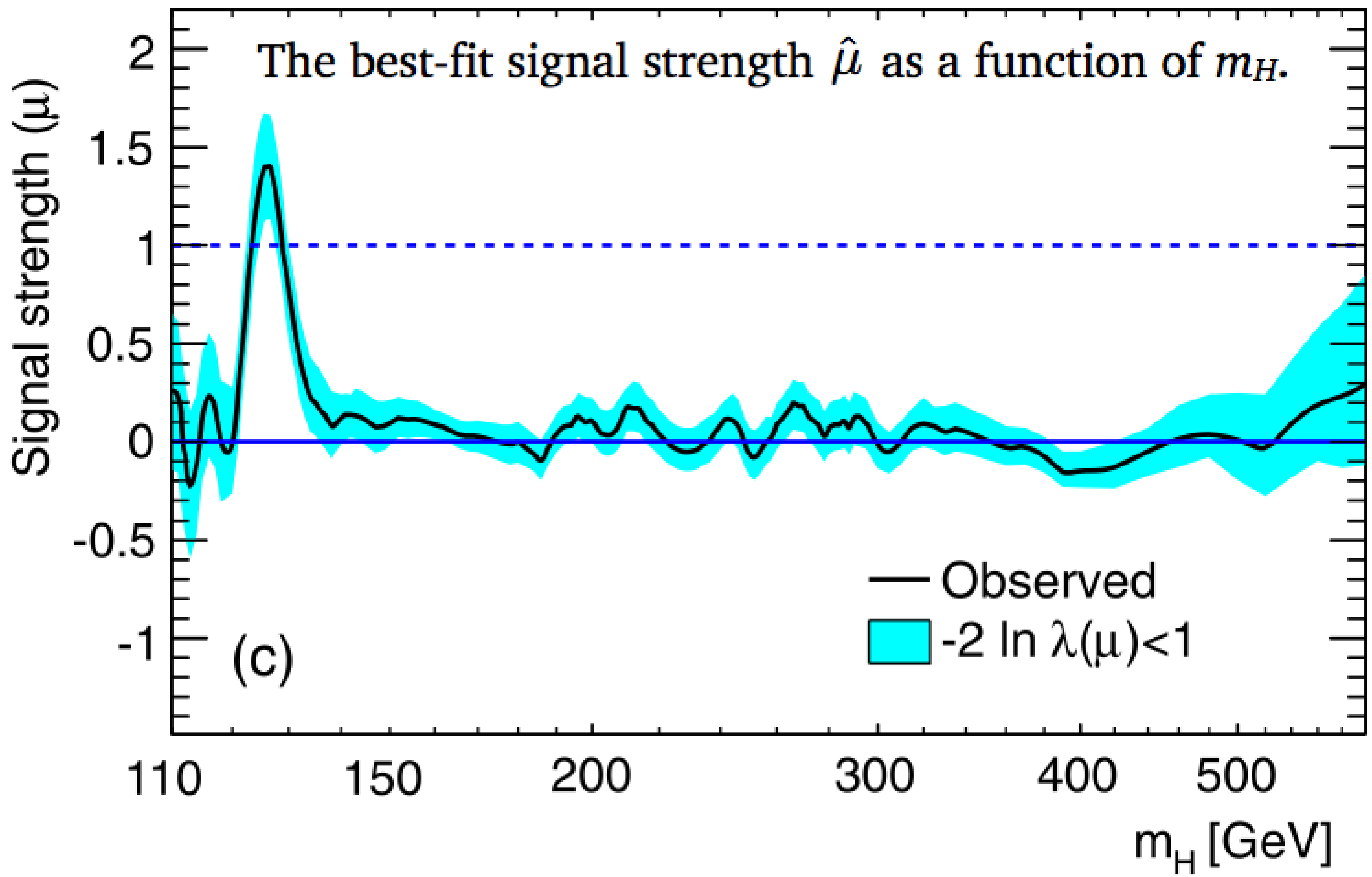
- The **local** p_0 values for a SM Higgs boson as a function of assumed m_H.
- The minimal p_0 (observed) is 2×10^{-6} at $m_H = 126.5$ GeV.
  $\Rightarrow$ local significance of $4.7\sigma \rightarrow$ reduced to 3.6σ after LEE



how to read the "*blue band*" plots

- $\hat{\mu}$ vs. m_H where $\hat{\mu}$ is the signal strength ($= \sigma/\sigma_{\mathrm{SM}}$) estimated by likelihood method[1]. The blue band corresponds to approx. $\pm 1\sigma$ error bar for μ.

ATLAS PLB 716, 1 (2012)



[1] Some details are skipped, for the sake of simplicity

Now that you have the language to talk about stat. interpretation of HEP results (e.g. LHC), it's your job to explore & enjoy them!

Thank you!