Probability in 4 historical steps:

1)CLASSICAL PROBABILITY: Pascal, Fermat et al., around 1650,
to support gamblers and games of dice/playing cards,
“*symmetry of different events*”
- no generalization to continuous case, possible multiple incoherent definitions -

2)AXIOMATICS PROBABILITY: Kolmogorov, 1950
“*axioms and formal theory*”
- no meaning about the actual values of probability -

3)FREQUENTIST PROBABILITY: von Mises et al, 1957
“*limit to infinity of ratio between preferred cases and the whole set of cases*”
- applicable only to observed data -

4)BAYESIAN PROBABILITY: 2nd half XX century (around latest 40 years, or afterwards de Finetti unsubstantial essay in 1974),
“*subjective probability based on Bayes theorem*”
- prior to be chosen -

By applying the rules (axioms by Kolmogorov):

For any element A of the space Ω of events: $P(A) \geq 0 \quad \forall A$

For the whole space Ω of mutually exclusive events: $P(\Omega) = 1$

If the events A and B are disjoint: $P(A \cup B) = P(A) + P(B)$

It follows that: $P(not\ A) = 1 - P(A)$

And the following two operations apply:

“SUM”: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

“PRODUCT” (conditional probability):

$$P(A \cap B) = P(A|B) * P(B)$$

and then, **INDEPENDENT** events : $P(A \cap B) = P(A) * P(B)$

The concept of “probability” is quite old

How to get “meat”

without being killed ?

WHICH “KIND” of PROBABILITY were WE APPLYING ?

What is the probability to “find meat and escape to big killers” whether I would take some kind of actions instead of others ?

I know how to evaluate a certain probability P_A
(how many pards did not die by taking the action A)

P_A(survival; action A)

HOWEVER, that DOES NOT answer the question. I am really interested to evaluate the probability of a certain action to let me survive !
i.e. P_S(action A; survival) i.e. I have to decide whether to take action A or not

Similar questions:
- I know the probability to die whether I smoke, but what is the probability that I smoke whether I am dying ? (a smokers, finally, supposes not to die !)
- I know what is the probability I win to lottery whether it is perfectly run, but what is the probability the lottery is unbiassed whether I win ?
- I can evaluate the probability that I own a soul whether God exists (San Tommaso), but what is the probability that God exists whether I own a soul ?

non sense ? Perhaps yes, for some questions, but Bayes Theorem gives you an answer !

All we discussed till know, in previous lessons, is the **DIRECT PROBABILITY**.

However, Thomas Bayes, during his life (1701-1761) discovered the
INVERSE PROBABILITY.
He never published it, which instead was released in 1763.

For about 2 centuries (200 years !) the theorem was “forgot” and
it was considered “non-sense”.

After all, we are all, individually speaking, thinking in terms of INVERSE PROBABILITY

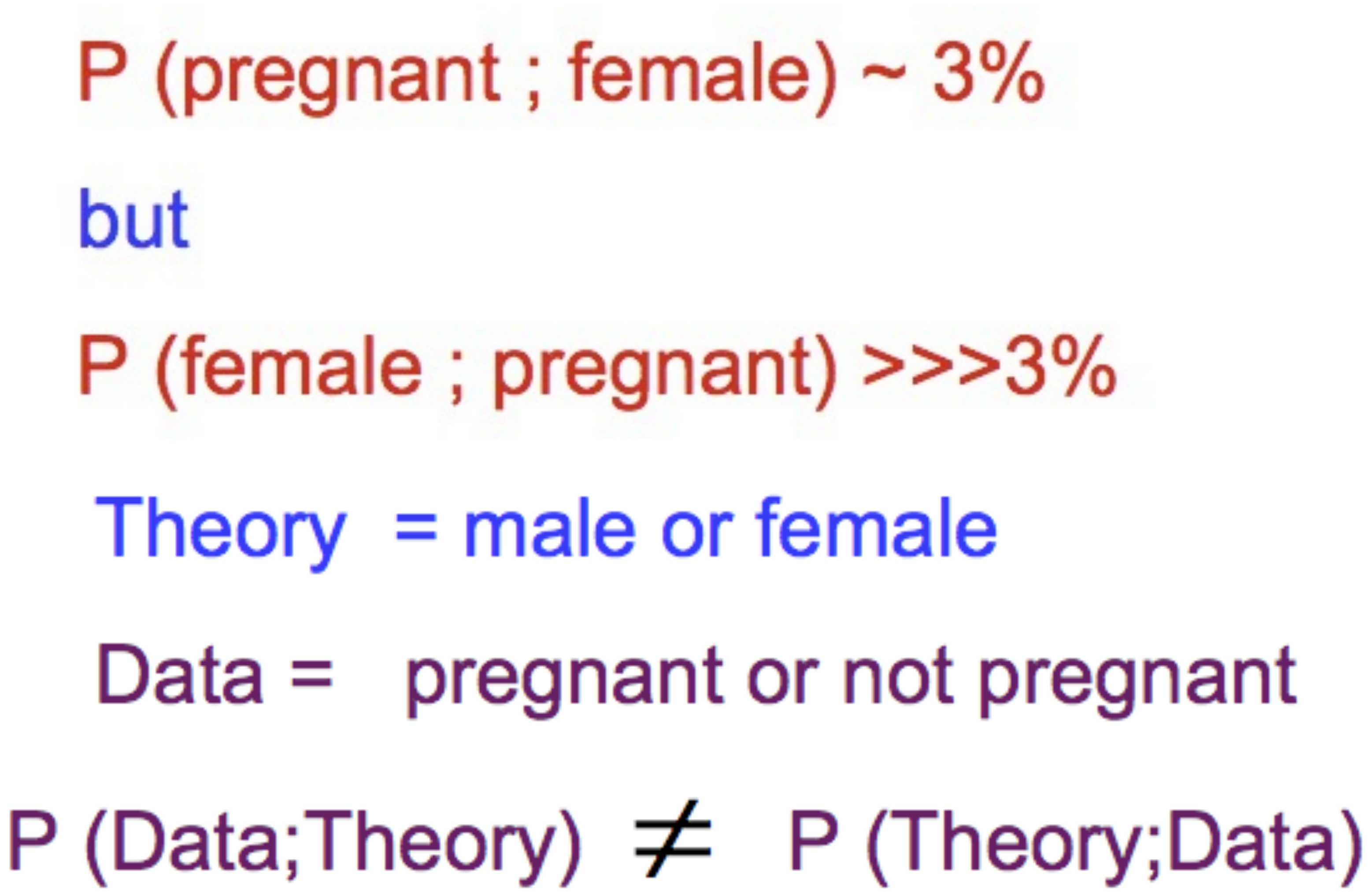
I am not interested to the probability that in the past 5 days it was mild and sunny…
I am interested to the probability that tomorrow be mild and sunny, to go to the beach

I am not interested to the average number of students that will pass exam,
I am interested to the probability to pass the exam myself

I am not interested to the probability that Data supports Dark Matter Hypothesis,
I am interested to the probability that Dark Matter exists

Just for joking (but terrifically true!)

(from Louis Lyons (Oxford), http://indico.cern.ch/conferenceDisplay.py?confId=a063350)



P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Theory = male or female

Data = pregnant or not pregnant

P (Data;Theory) $\neq$ P (Theory;Data)

More seriously:

Once a bridge is fallen the justice has to evaluate the probability

P(engineer’s mistake | bridge fallen)

and it may not evaluate the probability

P(bridge fallen | engineer’s mistake)

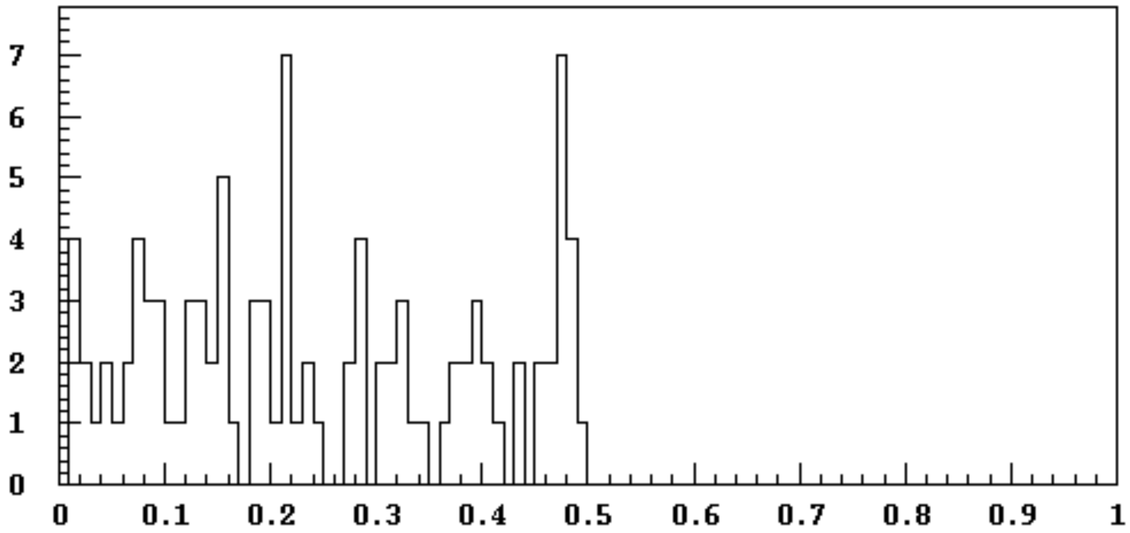
(the latter is a judiciary pursuable mistake in the judicial system!)

Remind also the attempts of some scientific eminent academics to argue against the sentence about the l’Aquila earthquake, which convicted the scientific advisory board of Protezione Civile…

REDUCTIO AD ABSURDUM PROOF

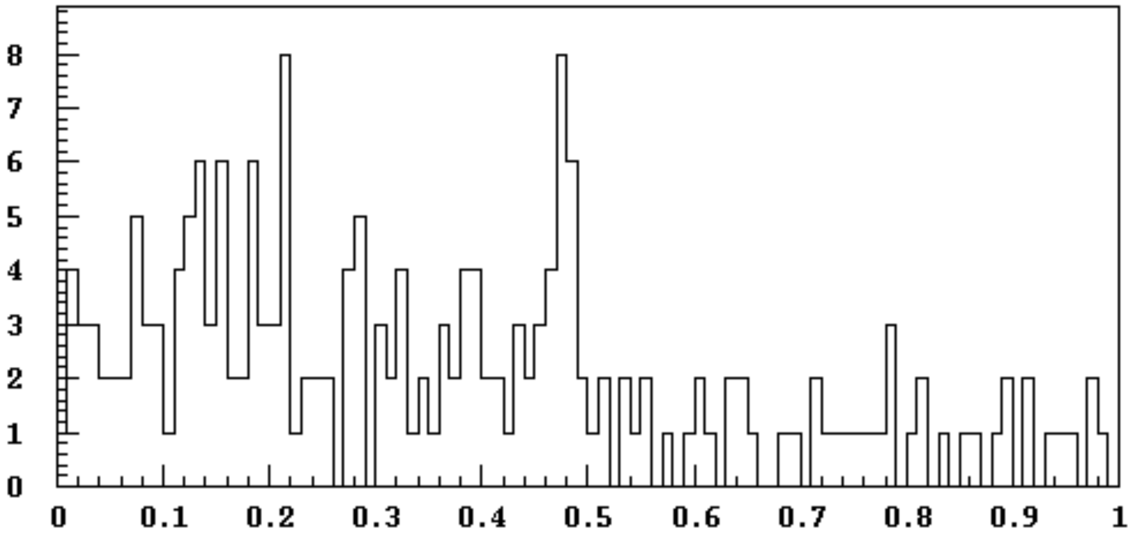
SUPPOSE I play to a perfect roulette: RED & BLACK
(consider a flat distribution in [0., 1.]

IF I count 100 times the RED, ie. I found 100 events in the in region [0., 0.5], how do I gamble for the 101 roll ??



Since the probability does not depend of the previous results, in the next 100 counts I expect a uniform distribution in [0.,1.]
However, I suspect that (almost) everybody would instead expect a distribution packed up in [0.5, 1.]

Actually, in 200 counts I would find a distribution like that:



i.e. a non-uniform PDF !

CLASSICAL PROBABILITY IS NOT ABLE TO SOLVE THE PROBLEM !

The problem lays in the definition of Probability:

PDF (data | physics law)

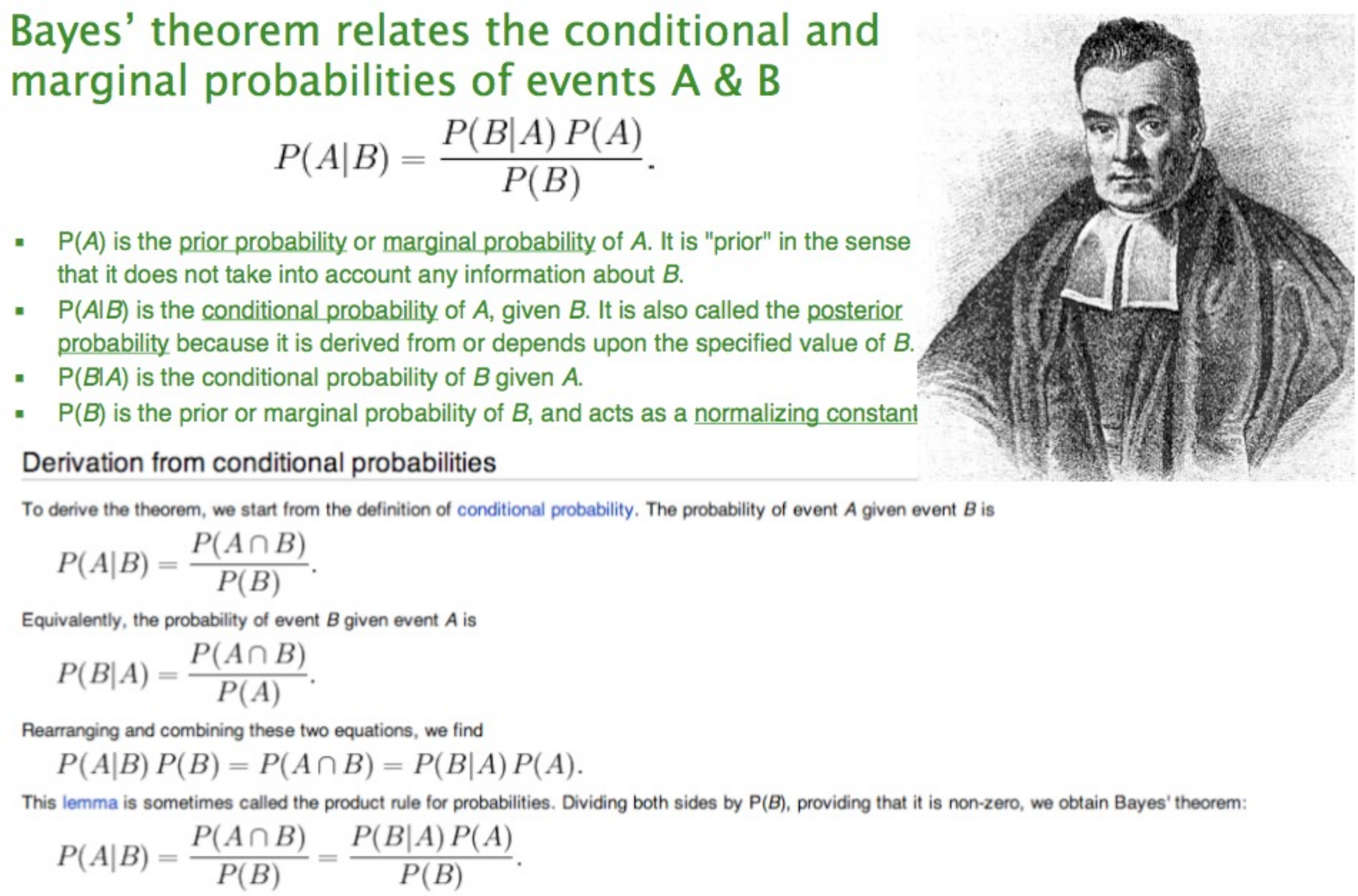
Instead I am trying to compute:

PDF (physics law | data)

Bayes …

Bayes Theorem

(from Kyle Cranmer, NYU, CERN Academic Training, Feb 2-5, 2009, http://indico.cern.ch/conferenceDisplay.py?confId=48425)



Bayes' theorem relates the conditional and marginal probabilities of events A & B

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

- P(*A*) is the prior probability or marginal probability of *A*. It is "prior" in the sense that it does not take into account any information about *B*.
- P(*A*|*B*) is the conditional probability of *A*, given *B*. It is also called the posterior probability because it is derived from or depends upon the specified value of *B*.
- P(*B*|*A*) is the conditional probability of *B* given *A*.
- P(*B*) is the prior or marginal probability of *B*, and acts as a normalizing constant

Derivation from conditional probabilities

To derive the theorem, we start from the definition of conditional probability. The probability of event *A* given event *B* is

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

Equivalently, the probability of event *B* given event *A* is

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

Rearranging and combining these two equations, we find

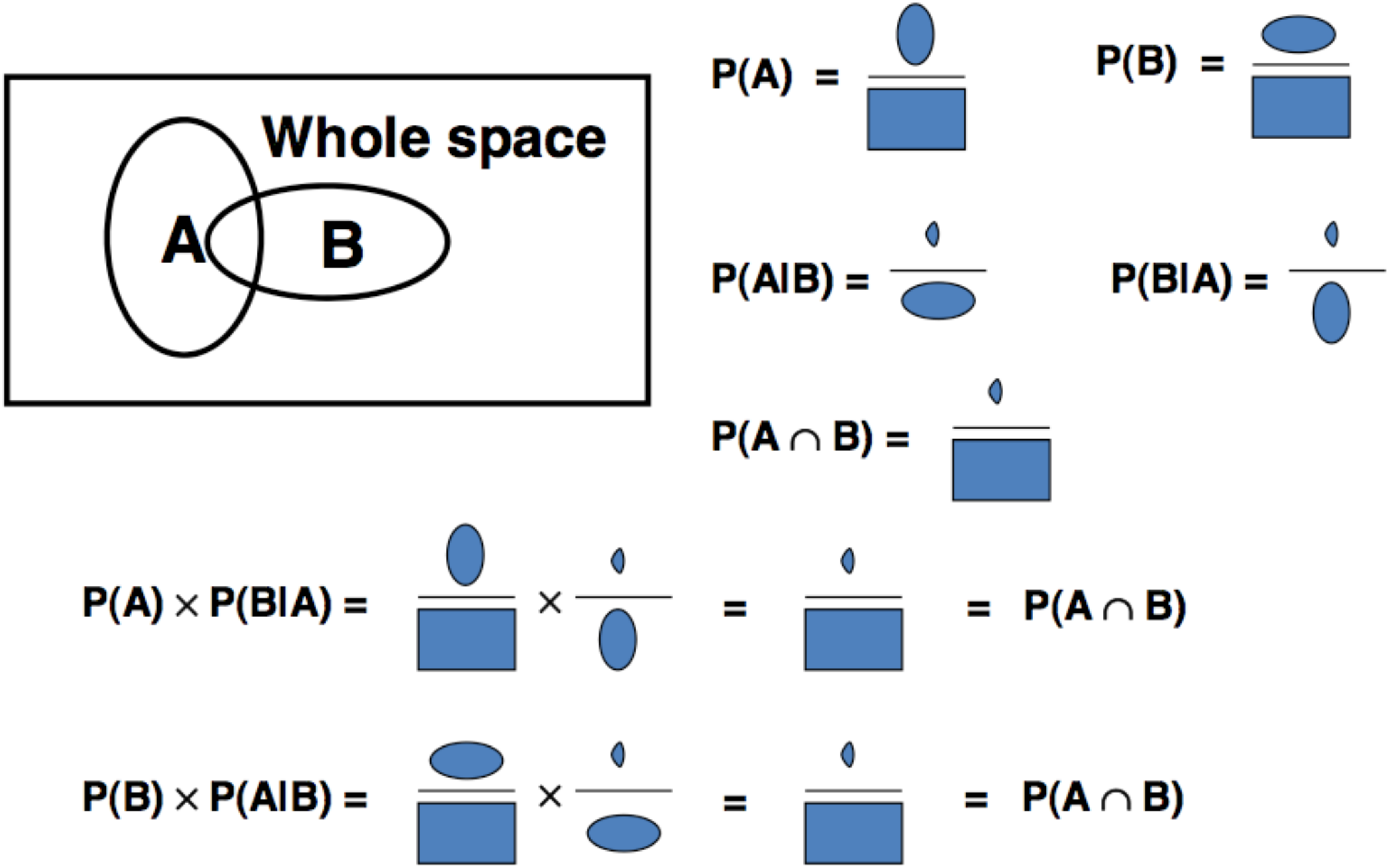
$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

This lemma is sometimes called the product rule for probabilities. Dividing both sides by P(*B*), providing that it is non-zero, we obtain Bayes' theorem:

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

... in pictures (from Bob Cousins)

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



Bob Cousins, CMS, 2008

⇒ **P(A|B) = P(B|A) x P(A) / P(B)**

It is true that we are able to deal only with DATA applied to functions/theories.

However, things become more understandable when theory depends of parameters that we may want to extract.
Actually, physicists were used to apply the method of evaluating theoretical parameters, and even adding an estimated error !

That is totally wrong.

Bayesianism versus Frequentism

"Bayesians address the question everyone is interested in, by using assumptions no-one believes *(or obvious)*"

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

(P.G.Hamer cited by Kyle Cranmer)

Bayesian vs Frequentist

(inference, but also statistics, people, approach, idea, opinion, bias, superstition…)

Statistical Inference ("learning"): process of using DATA to *infer* the DISTRIBUTION that generated the data.

How we interpret PROBABILITY

Frequentist: Probability can be interpreted as FREQUENCY:

$$\mathcal{P} = n \,/\, N$$

where n stands for successes and N as the total number of trials

Bayesian: Probability can be interpreted as LACK of KNOWLEDGE, from the Bayes theorem:

$$\mathcal{P}(\mathrm{H}|\mathrm{D}) = \mathcal{P}(\mathrm{D}|\mathrm{H}) * \pi(\mathrm{H}) \,/\, \mathcal{P}(\mathrm{D})$$

Posterior

Prior

normalization

where **H** stands for *hypothesis* and **D** for *data*

Likelihood

Bayes Theorem applied to Physics

$$\overset{A}{P(\theta|\mathbf{X})} = \overset{B}{P(\mathbf{X}|\theta)} * \overset{C}{\pi(\theta)} / \overset{D}{\text{Integral}}$$

θ: parameter (Physics Model)
X: DATA

A: what is the probability that a certain model is described by DATA (posterior) ?

B: how much is likely that DATA describes a certain model ?

C: prior (information/limits about the physics model)

D: normalization

The **PRIOR** problem

Physicists prefer the **UNIFORM** prior, which however owns two drawbacks:

1) To be finite, or $\int \pi(x) \cdot dx = 1$, the range has to be defined
2) It is not invariant for parameter transformation

Of course, both the limits can be partially overcome:

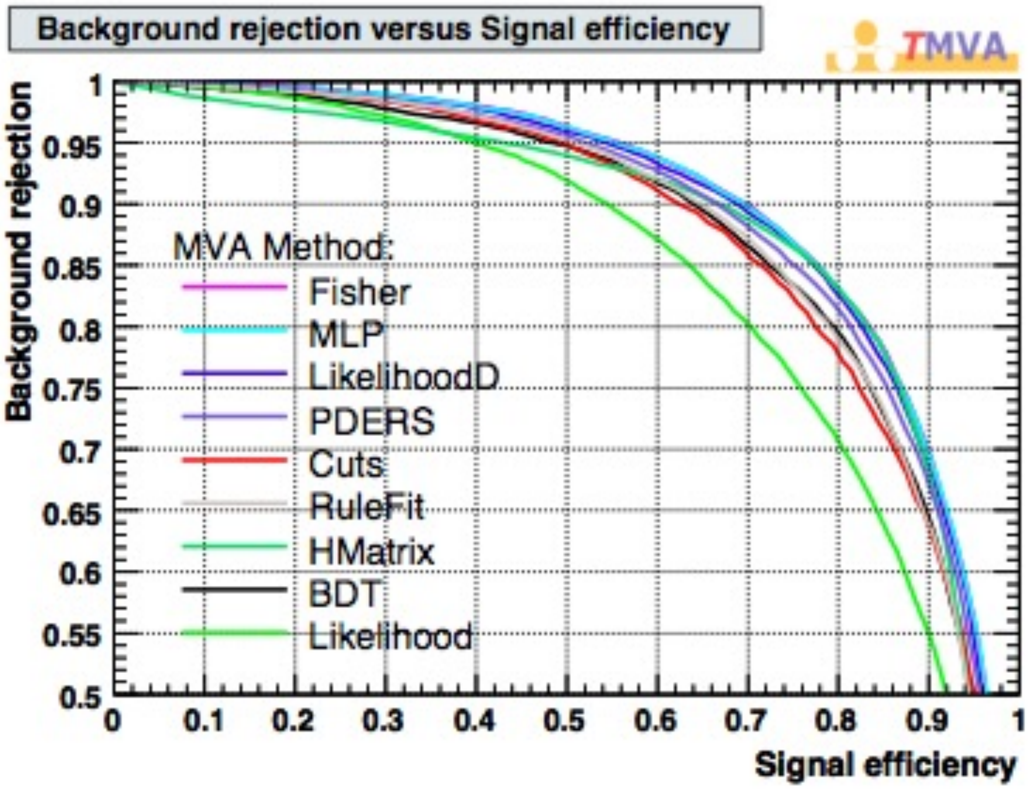
1) Use the denominator to normalize
2) Make the statistical analysis with the ultimate parameter

Another good possibility is to use as PRIOR the whole previous collected information

A non-controversial use of the Bayes theorem occurs sometime…

Identify a subset of events by applying certain conditions/algorithms, e.g.

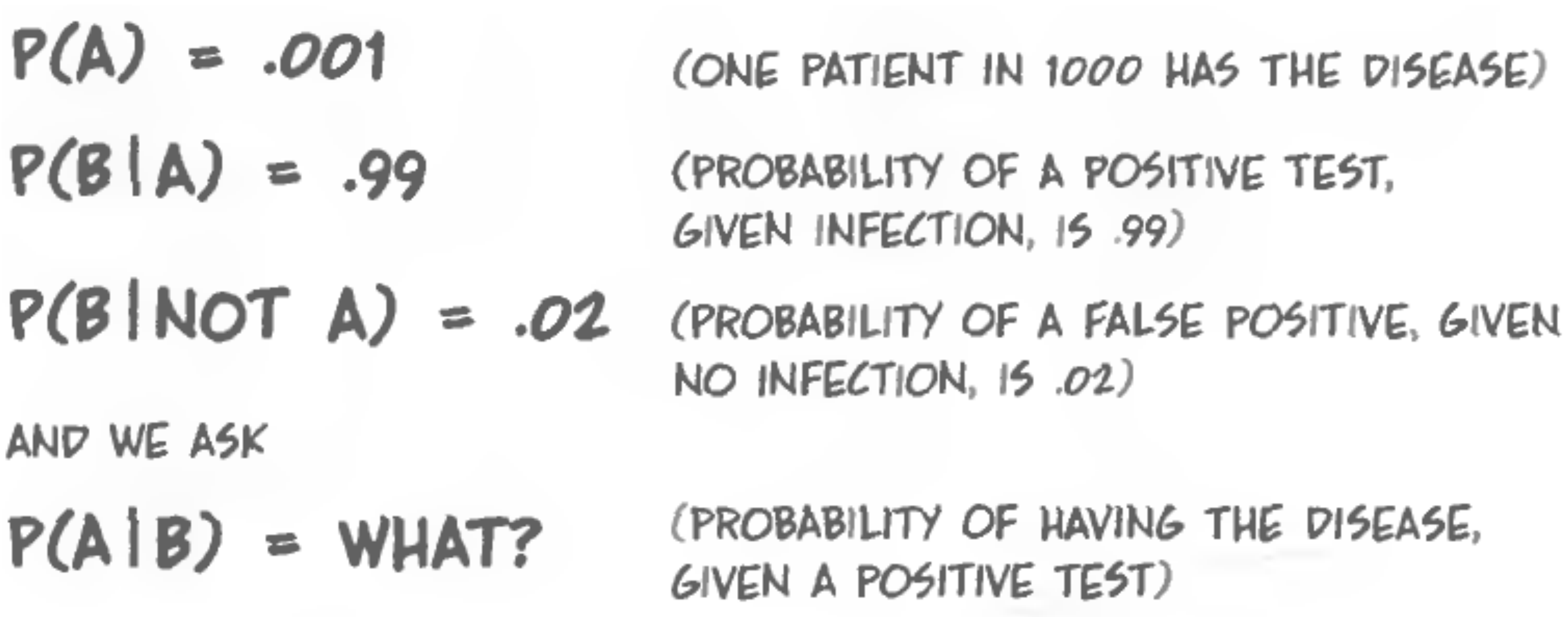
1. Person sickness estimation via medical checks
2. b-quarks estimation via b-tagging algorithm
3. Physics students via question form



In any context one usually comes out with the following pattern:
PURITY vs EFFICIENCY

Perform "measurement" and get:
P(data|signal)=efficiency for signal
P(data|bck)=eff. for background

Extract P(signal|data) from Bayes, but one needs P(signal) !



$P(A) = .001$ (ONE PATIENT IN 1000 HAS THE DISEASE)

$P(B|A) = .99$ (PROBABILITY OF A POSITIVE TEST, GIVEN INFECTION, IS .99)

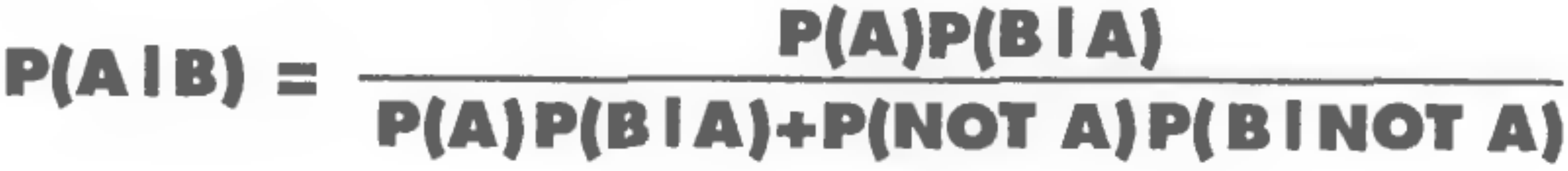
$P(B|\text{NOT } A) = .02$ (PROBABILITY OF A FALSE POSITIVE, GIVEN NO INFECTION, IS .02)

AND WE ASK

$P(A|B) = \text{WHAT?}$ (PROBABILITY OF HAVING THE DISEASE, GIVEN A POSITIVE TEST)

We need to know P(A) !

Then



$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(\text{NOT } A)P(B|\text{NOT } A)}$$

and P(A|B)=0.0472

CAN BE EXPRESSED AS

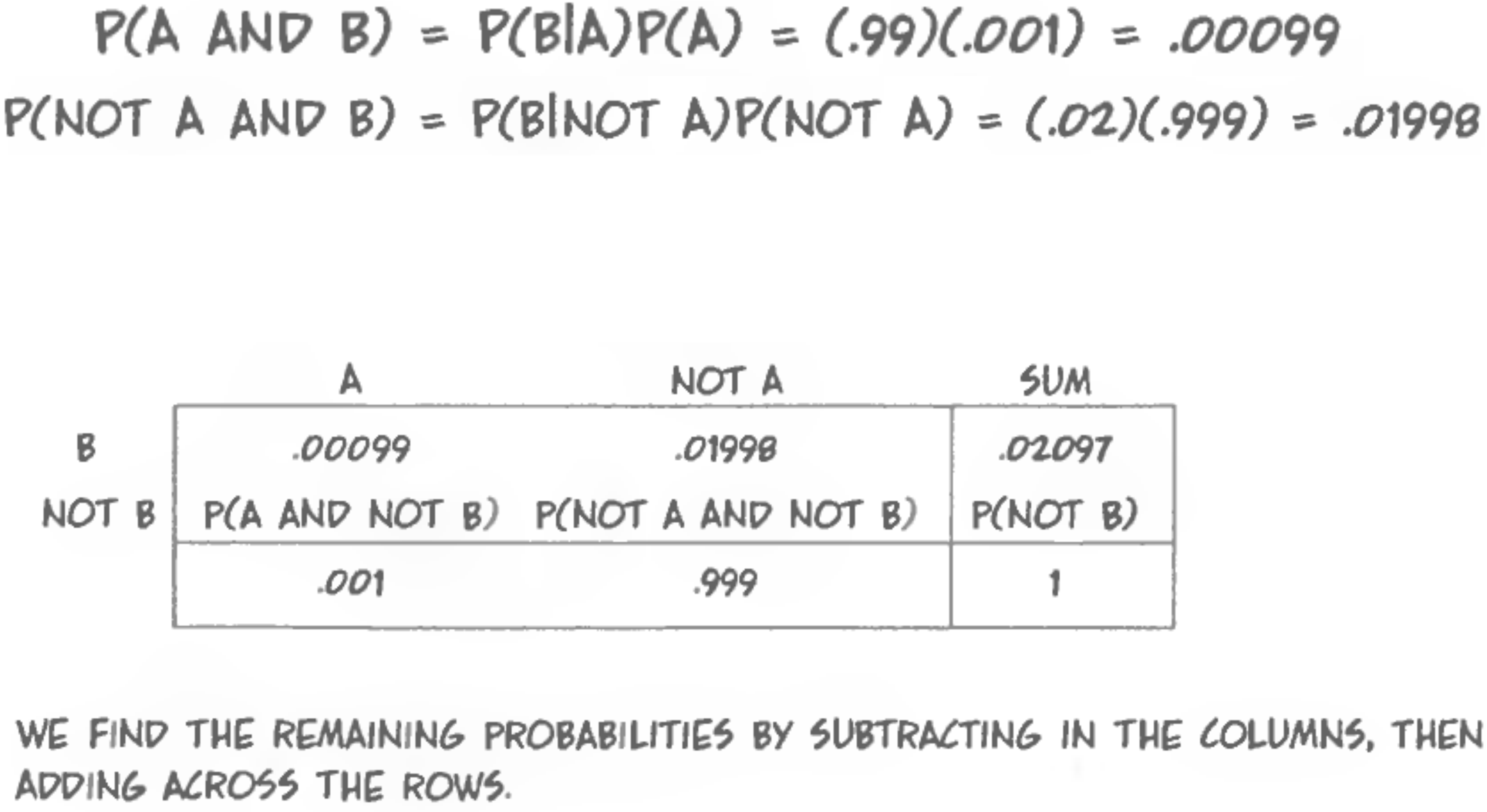
$$\frac{P(A \text{ and } B)}{P(A \text{ and } B)+P(\text{NOT } A \text{ and } B)} = \frac{P(A \text{ and } B)}{P(B)} = P(A|B)$$

	A	NOT A
B	A AND B	NOT A AND B
NOT B	A AND NOT B	NOT A AND NOT B

LET'S FIND THE PROBABILITIES OF EACH EVENT IN THE TABLE:

	A	NOT A	SUM
B	P(A AND B)	P(NOT A AND B)	P(B)
NOT B	P(A AND NOT B)	P(NOT A AND NOT B)	P(NOT B)
	P(A)	P(NOT A)	1

THE PROBABILITIES IN THE MARGINS ARE FOUND BY SUMMING ACROSS ROWS AND DOWN COLUMNS.

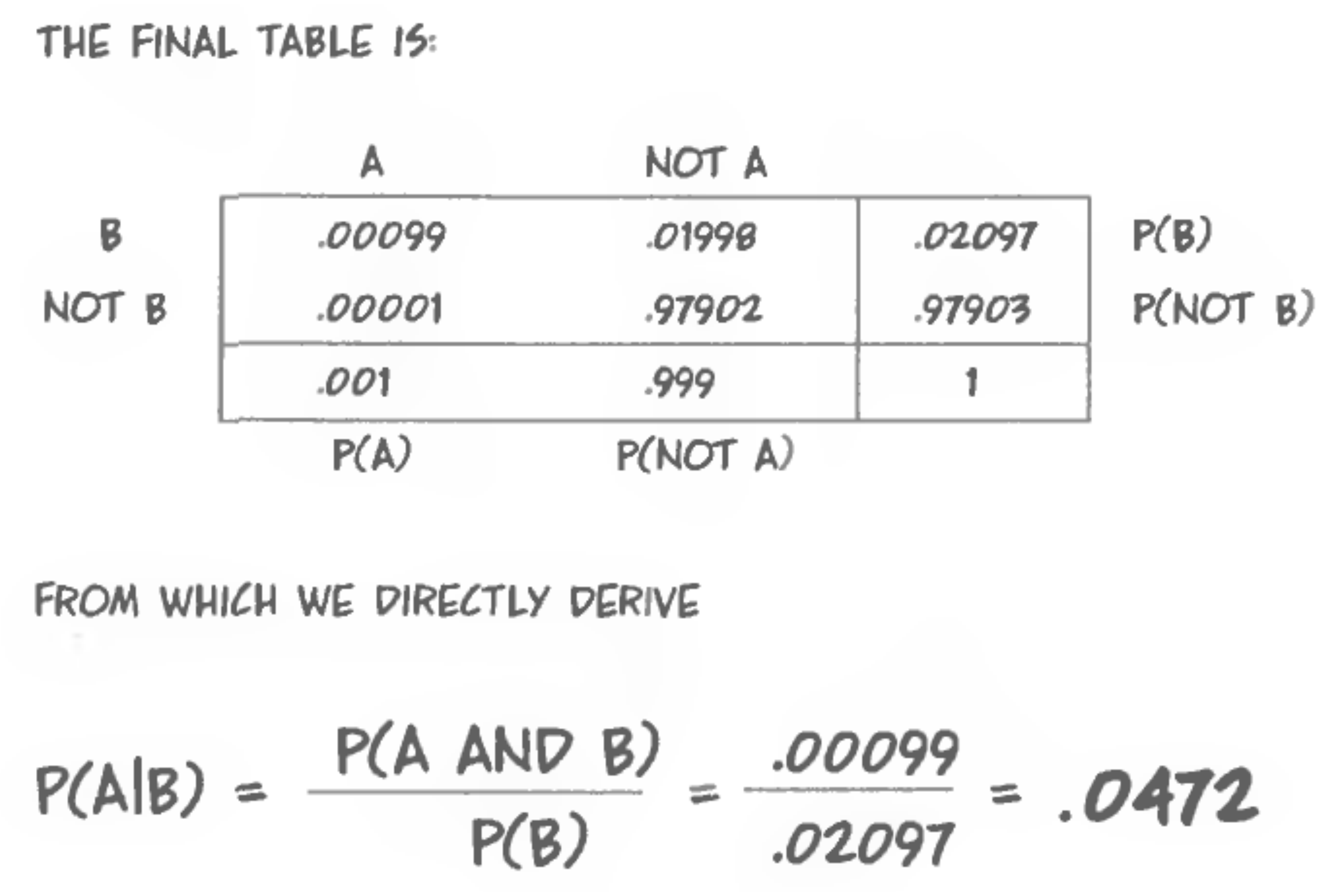


$$P(\text{A AND B}) = P(\text{B}|\text{A})P(\text{A}) = (.99)(.001) = .00099$$

$$P(\text{NOT A AND B}) = P(\text{B}|\text{NOT A})P(\text{NOT A}) = (.02)(.999) = .01998$$

	A	NOT A	SUM
B	.00099	.01998	.02097
NOT B	P(A AND NOT B)	P(NOT A AND NOT B)	P(NOT B)
	.001	.999	1

WE FIND THE REMAINING PROBABILITIES BY SUBTRACTING IN THE COLUMNS, THEN ADDING ACROSS THE ROWS.



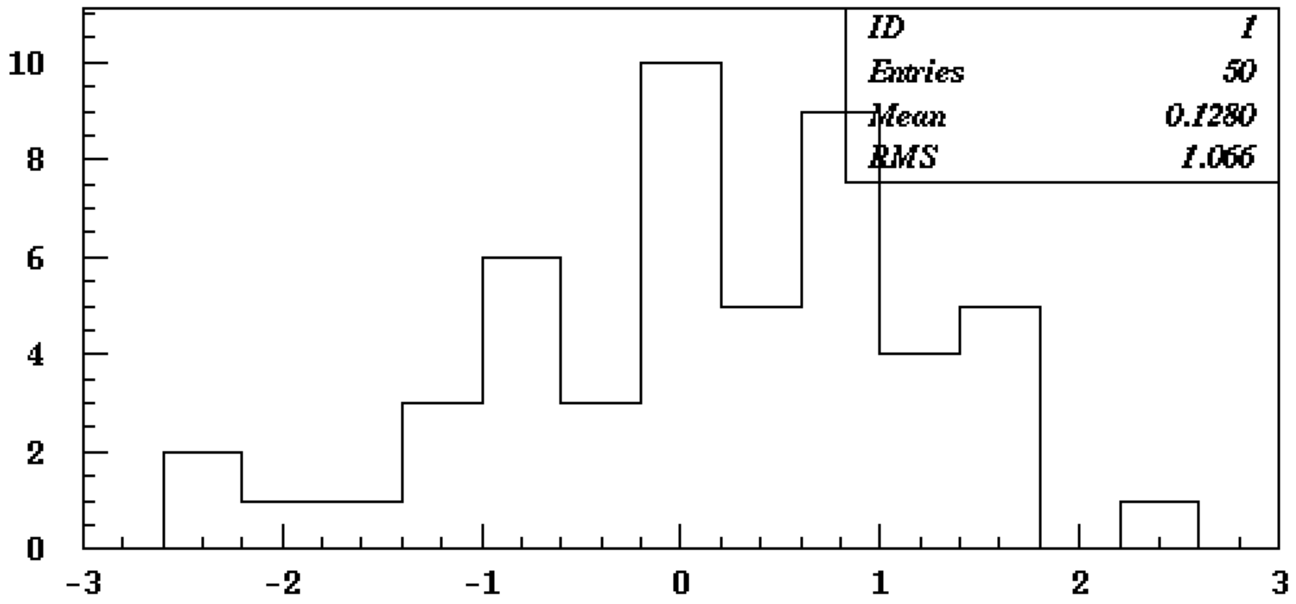
THE FINAL TABLE IS:

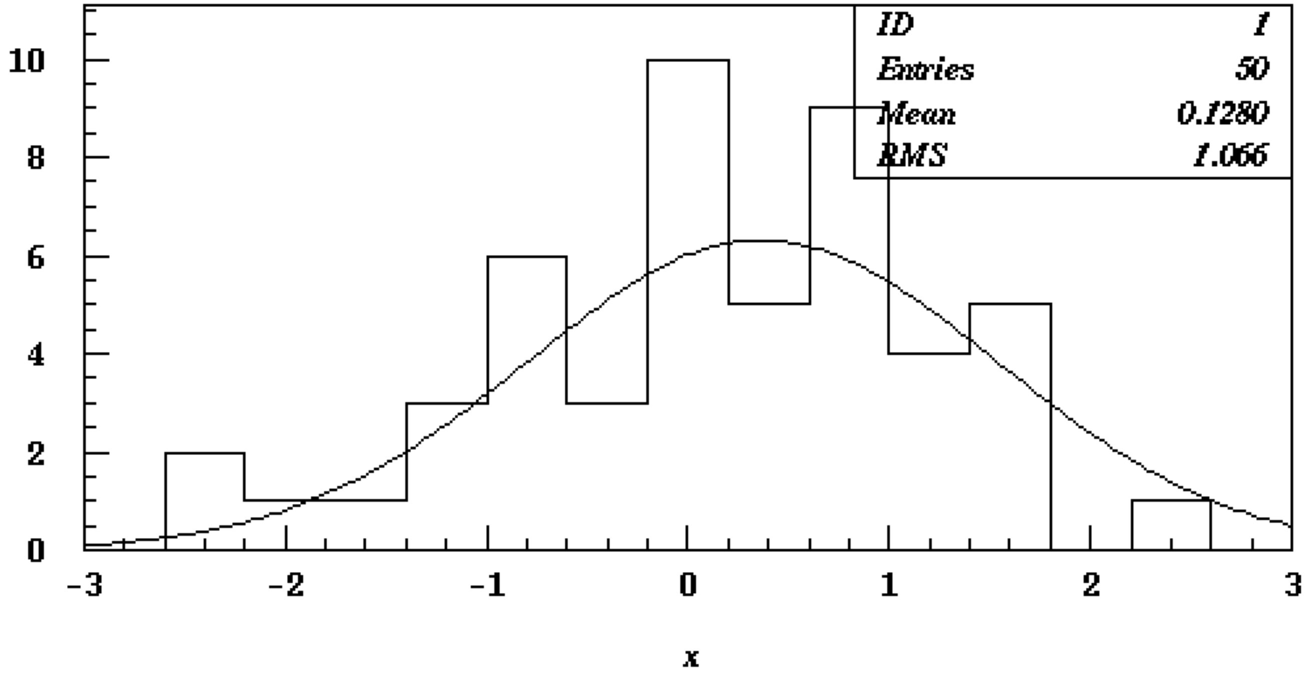
	A	NOT A		
B	.00099	.01998	.02097	P(B)
NOT B	.00001	.97902	.97903	P(NOT B)
	.001	.999	1	
	P(A)	P(NOT A)		

FROM WHICH WE DIRECTLY DERIVE

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{.00099}{.02097} = .0472$$

Bayes against Frequentism: be G(0,1) the theory, and suppose 50 measurements





$\mu=0,$
$\sigma=1$
} THEORY

$x_{med}=0.1280 \pm 0.151$
sqm=1.066

$\mu'=x_{fit}=0.3696 \pm 0.165$
$\sigma'=\sigma_{fit}=1.1689$
$\chi^2=0.852$ (7.668/(12-3))
$Pr(\chi^2)=Pr(7.668;9)=1-0.568$

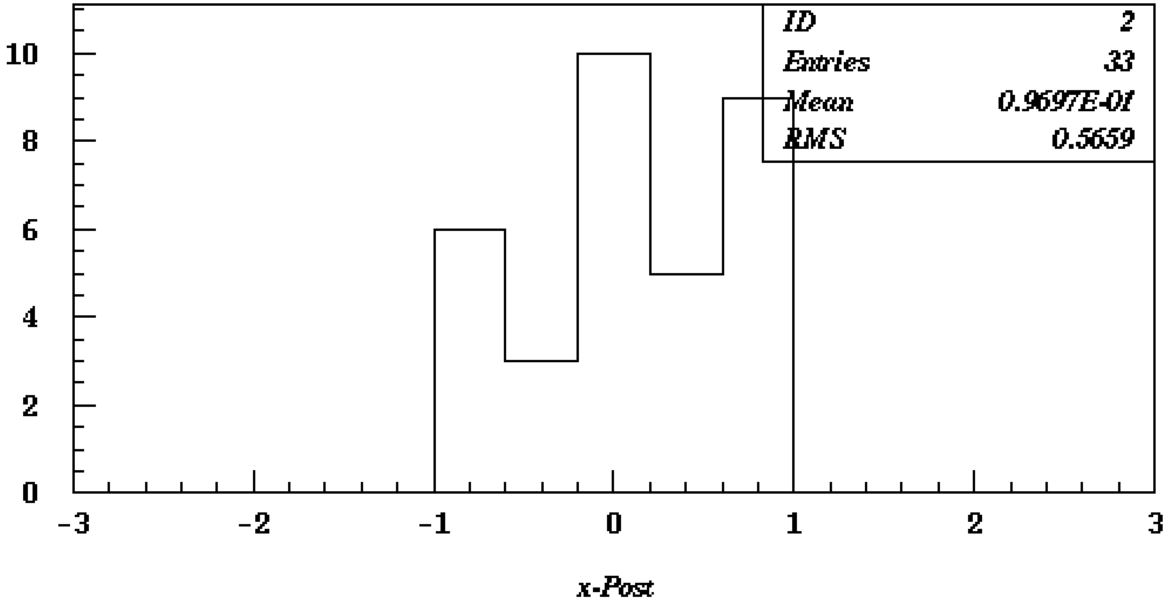
The weighted mean is only slightly better of the fitted one

$x_{med}=0.3679 \pm 0.1586$

Suppose that we already know that -1<μ<1

Apply Bayes $P(\mu';x_i) = G(x_i;\mu',\sigma')\cdot\pi(\mu')$

with π(μ')=uniform in [-1,1]



x_{med}=0.0970 ± 0.0985
sqm=0.566

This is the PDF of μ, i.e. the inverse-probability,
it cannot be constructed via direct–probability densities

Back to the problematic use of the Bayes theorem…

In 1966 **Jeffreys** introduced the “objective” prior, by taking into account the *Fisher’ information*:
in average the amount of “information” from a measurement is given by the second derivative of the Likelihood.

$$\boldsymbol{I(\theta) \equiv -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]} \quad ^*$$

** the “expectation value” is over PDF(x;θ)*

Demonstration:

$$E\left\{\left(\frac{\partial(\ln L)}{\partial\theta}\right)^2\right\} = \int\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)^2 \cdot L = \boxed{\int \frac{1}{L}\left(\frac{\partial L}{\partial\theta}\right)^2}$$

$$E\left\{\frac{\partial^2(\ln L)}{\partial\theta^2}\right\} = \int \frac{\partial}{\partial\theta}\left(\frac{\partial(\ln L)}{\partial\theta}\right)\cdot L = \int \frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)\cdot L = \int\left(-\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 + \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2}\right)\cdot L =$$

$$= -\int \frac{1}{L}\left(\frac{\partial L}{\partial\theta}\right)^2 + \int \frac{\partial^2 L}{\partial\theta^2} = \boxed{-\int \frac{1}{L}\left(\frac{\partial L}{\partial\theta}\right)^2}$$

Surely it holds:

$$E\left\{\frac{\partial(\ln L)}{\partial\theta}\right\}=\int\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)\cdot L=\frac{\partial}{\partial\theta}\int L=0$$

since

$$\int_{-\infty}^{+\infty} L=1$$

The Jeffreys' prior has been defined to prevent any subjective choice, i.e. to be constant in the Fisher Information, without adding more information (i.e. it is supposed to be "uninformative"):

$$\pi_J(\theta) \equiv \sqrt{I(\theta)} = \sqrt{E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

Then for the Normal distribution:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(\theta-x)^2/2\sigma^2} \rightarrow \ln\frac{1}{\sigma\sqrt{2\pi}} - (\theta - x)^2/2\sigma^2 \rightarrow -(\theta - x)/\sigma^2 \rightarrow$$

$$\rightarrow -\frac{1}{\sigma^2}\int f(x)\,dx \rightarrow \sqrt{\frac{1}{\sigma^2}} \rightarrow 1/\sigma$$

For the Poisson distribution and $\theta=\mu$: $\pi_J=1/\sqrt{\mu}$

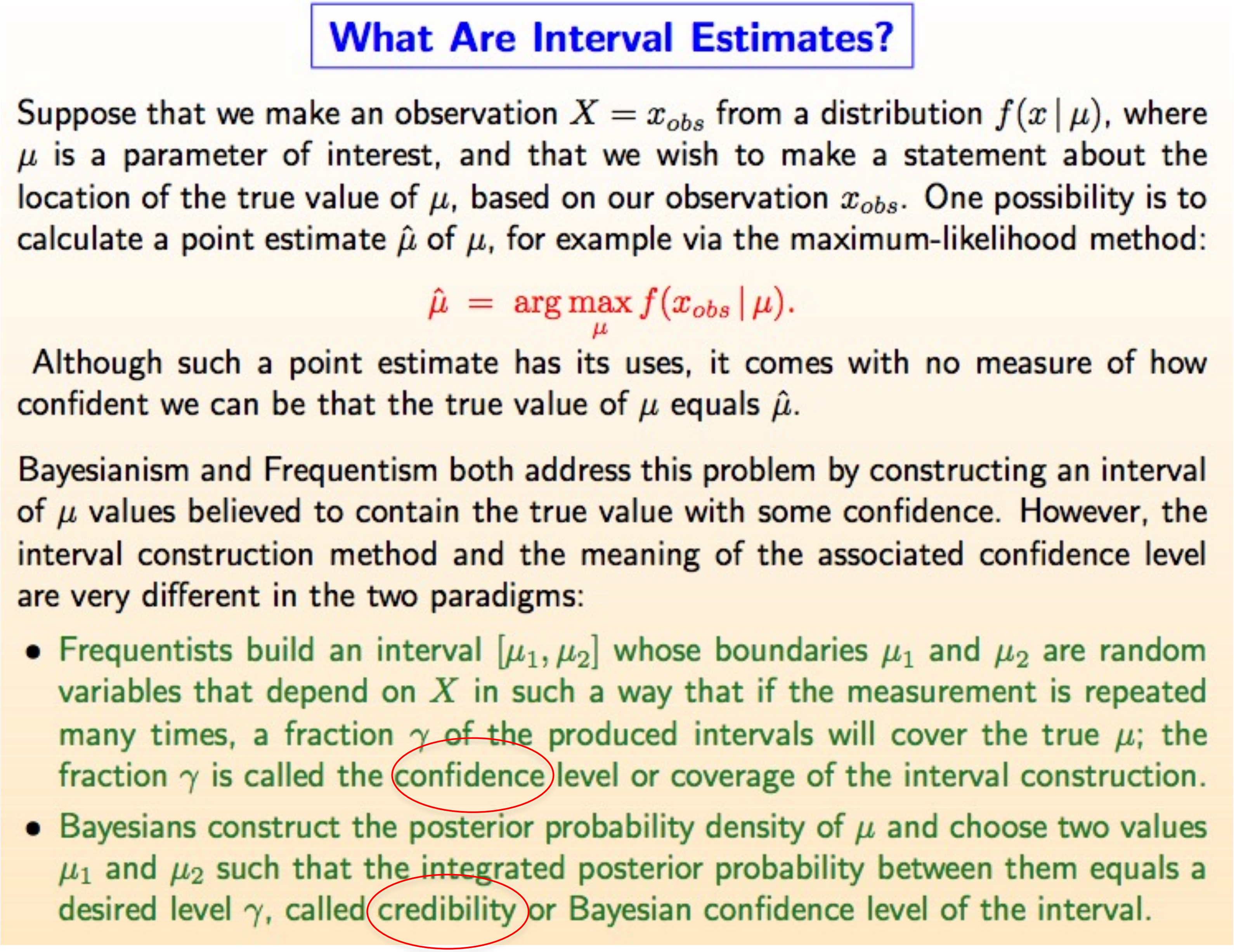
It is invariant under re-parameterization !

However, carefulness has to be taken. Think all the time what you are doing.

-Jeffreys' prior misbehaves for multidimensional parameter
-It depends heavily on the Likelihood, i.e. on the chosen θ parameter and data set
-It violates the Likelihood principle *("all the information is contained in the Likelihood, i.e. the function obtained by applying the data to the PDF")*
since the prior does not depend on the data
-More relevant: it may constraint too much your analysis
(see next example)

CONFIDENCE INTERVALS…………………………..

(L. Demortier in Terascale Stat. School)



What Are Interval Estimates?

Suppose that we make an observation $X = x_{obs}$ from a distribution $f(x \mid \mu)$, where μ is a parameter of interest, and that we wish to make a statement about the location of the true value of μ, based on our observation x_{obs}. One possibility is to calculate a point estimate $\hat{\mu}$ of μ, for example via the maximum-likelihood method:

$$\hat{\mu} \;=\; \arg\max_{\mu} f(x_{obs} \mid \mu).$$

Although such a point estimate has its uses, it comes with no measure of how confident we can be that the true value of μ equals $\hat{\mu}$.

Bayesianism and Frequentism both address this problem by constructing an interval of μ values believed to contain the true value with some confidence. However, the interval construction method and the meaning of the associated confidence level are very different in the two paradigms:

- Frequentists build an interval $[\mu_1, \mu_2]$ whose boundaries μ_1 and μ_2 are random variables that depend on X in such a way that if the measurement is repeated many times, a fraction γ of the produced intervals will cover the true μ; the fraction γ is called the confidence level or coverage of the interval construction.
- Bayesians construct the posterior probability density of μ and choose two values μ_1 and μ_2 such that the integrated posterior probability between them equals a desired level γ, called credibility or Bayesian confidence level of the interval.

Confidence Intervals *(frequentist)*

A $1\text{-}\alpha$ confidence interval for a parameter θ is an interval $C_n=(a,b)$ where $a=a(X_1,\dots, X_n)$ and $b=b(X_1,\dots,X_n)$ are functions of DATA such that

$$\mathcal{P}_\theta(\theta \in C_n) \geq 1\text{-}\alpha, \text{ for all } \theta \in \Theta.$$

In words (a,b) traps θ with probability $1\text{-}\alpha$.

$1\text{-}\alpha$ is called the **coverage** of the *confidence interval* (normally choose $\alpha=0.05$).

NOTE: θ is fixed and C_n is random !

Therefore a *confidence interval* is NOT a probability statement about θ.

if I repeat the experiment over and over
(or I take different DATA SAMPLES),
the intervals will contain the true parameter 95% of the time,
id est 95% of the intervals will trap the true parameter value.

Confidence Intervals *(Bayesian)*

Bayesians can make statements like:

The probability that θ is in **C**$_{\mathbf{n}}$, given the data, is 95%.

Bayesians make inferences about θ (fixed parameter) by producing a probability distribution for θ.

Confidence Intervals can be extracted from these distributions

Given *n* observations, $x_1,\ldots x_n$, and the parameter(model) θ, Likelihood is defined as

$$L_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and the posterior distribution is (up to a normalization factor): $f(\theta;x_n) \propto L(\theta) \times f(\theta)$

Let *C=(a,b)* the interval estimate. *a* and *b* are such that $\int_{-\infty}^{a} f(\theta;x_n)d\theta = \int_{b}^{\infty} f(\theta;x_n)d\theta = \alpha/2$

Therefore $P(\theta \in C;x_n) = \int_a^b f(\theta;x_n)d\theta = 1-\alpha$ and *C* is a 1-α **posterior interval**.

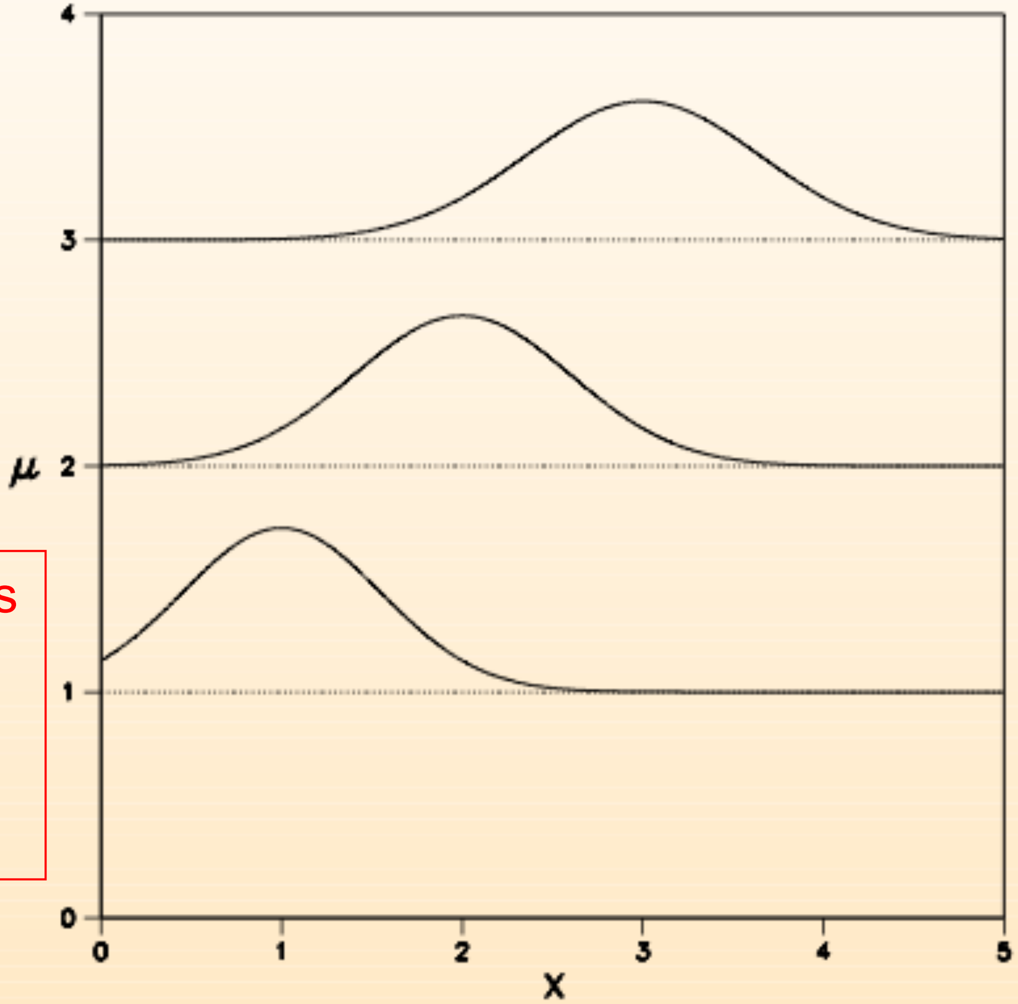
However these Bayesian intervals refer to degree-of-belief probabilities. **It is NOT true that:**
the Bayesian intervals will trap the true parameter 95% of the time!

Credibility Intervals *(Bayesian)*

Frequentist Intervals: the Neyman Construction (1)

J. Neyman, Philos. Trans. R. Soc. London ***767, 333, 1937***

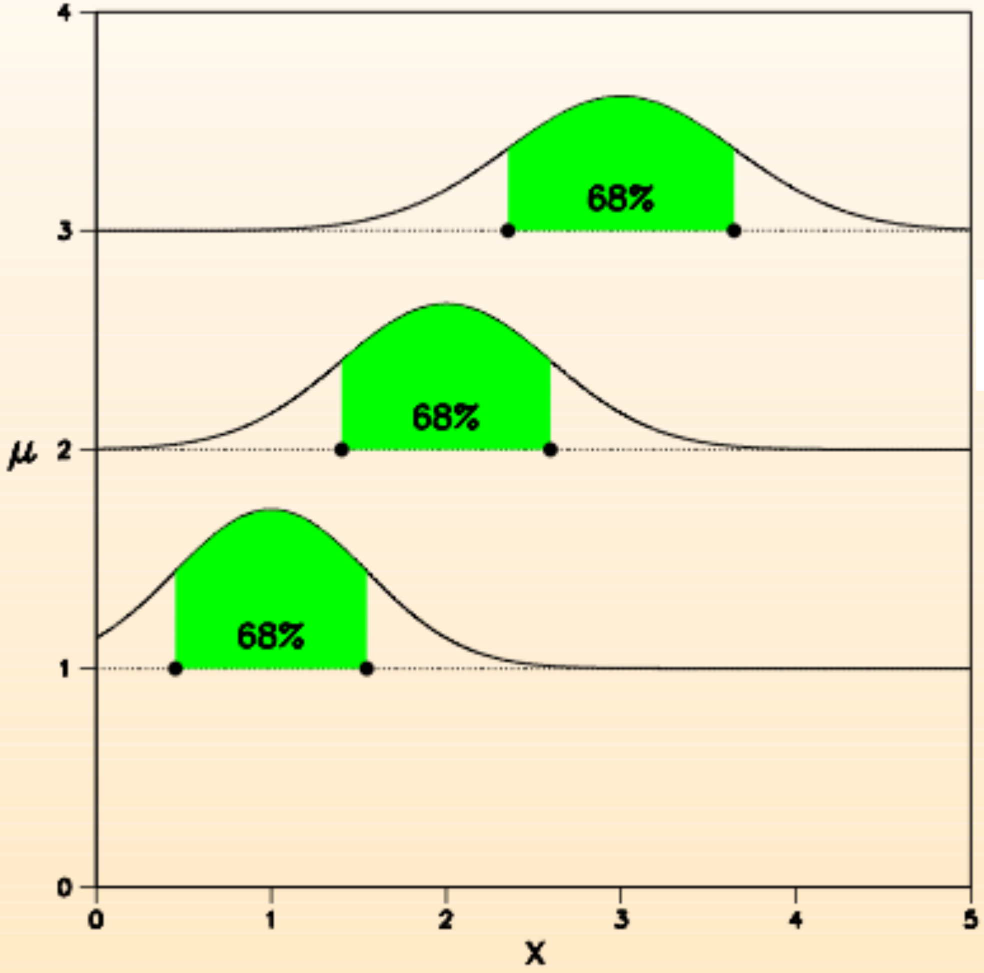
Step 1: Make a graph of the parameter μ versus the data X, and plot the density distribution of X for each value of μ.



Note: μ continous
x discrete.
That creates
inconsitencies
at the boundery.

Frequentist Intervals: the Neyman Construction (2)

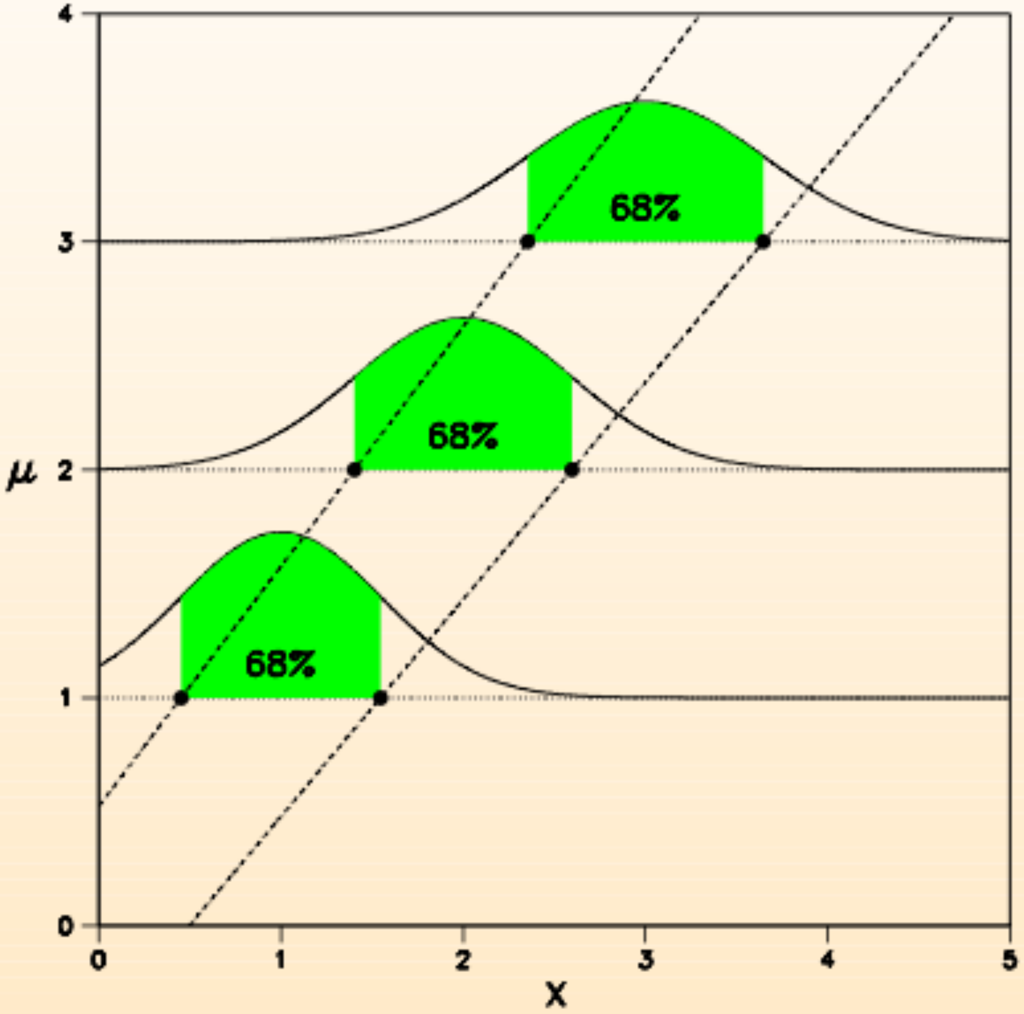
Step 2: For each value of μ, select an interval of X values that has a fixed integrated probability, for example 68%.



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

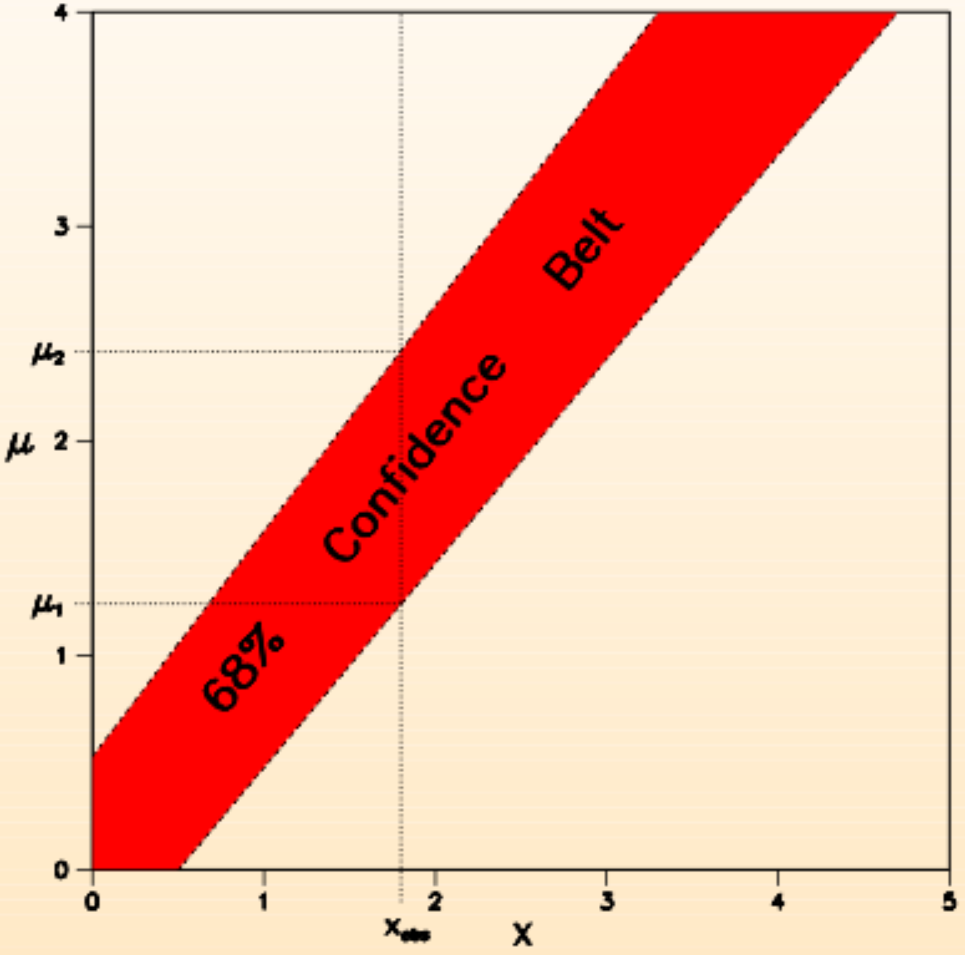
Frequentist Intervals: the Neyman Construction (3)

Step 3: Connect the interval boundaries across μ values.



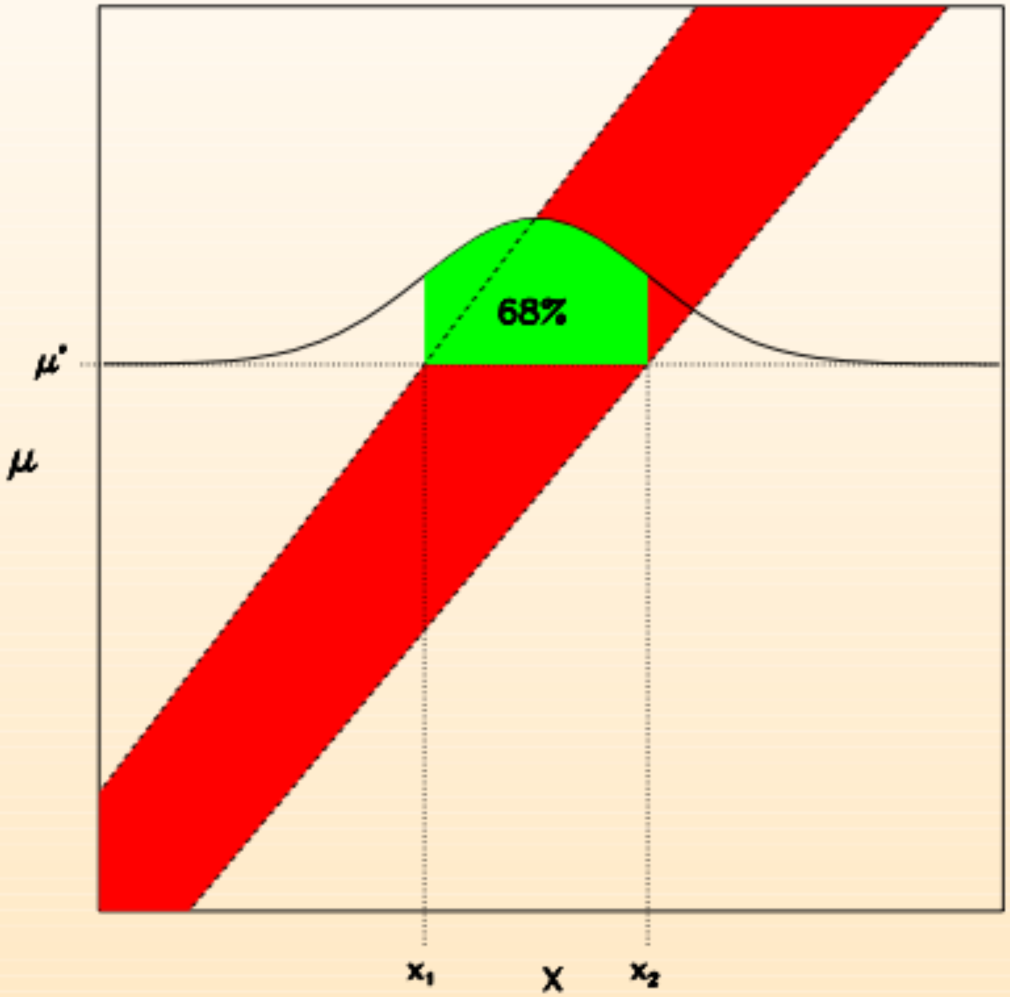
Frequentist Intervals: the Neyman Construction (4)

Step 4: Drop the "scaffolding" and use the resulting confidence belt to construct an interval $[\mu_1, \mu_2]$ for the true value of μ every time you make an observation x_{obs} of X.



Frequentist Intervals: the Neyman Construction (5)

Why does this work?
Suppose $\mu^\star$ is the true value of μ. Then $\mathbb{P}(x_1 \leq X \leq x_2 \,|\, \mu^\star) = 68\%$.
Furthermore, for every $X \in [x_1, x_2]$, the reported μ-interval will contain $\mu^\star$.
Therefore, the probability of covering $\mu^\star$ is 68%.



But now problems comes…

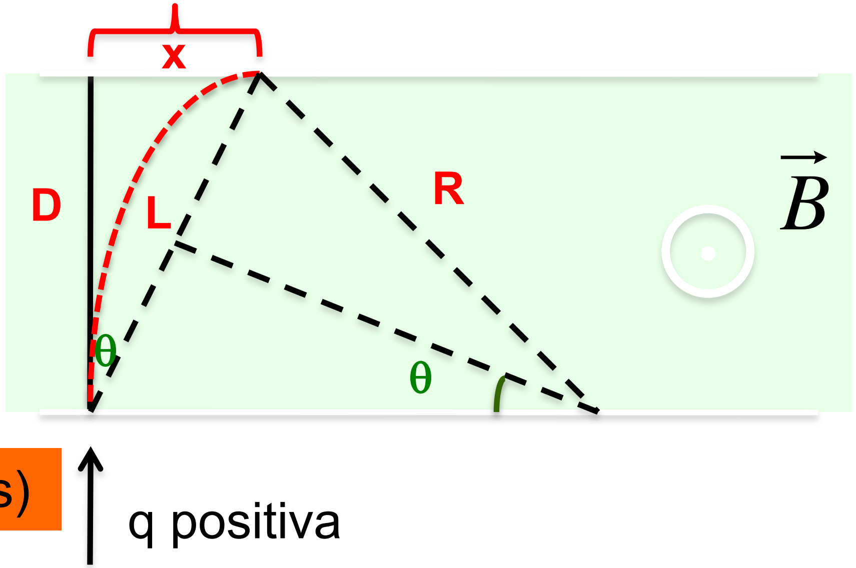
No such mental and technical gymnastic needed for Bayesian intervals:

- Construct the Posterior probability, i.e. the PDF of the parameter
- Choose the C.L.
- Extract the C.I. by your “preferred” rule

$$\vec{F} = q\vec{v} \times \vec{B} = m\vec{a}_{centripeta} = m\frac{v^2}{R}\vec{u}_F$$

mv = p = qBR

p(Gev/c) = 0.3 q(positron) B(Tesla) R(meters)



$$L = \sqrt{D^2 + x^2}, \quad R = \frac{L}{2\sin(\theta)} = \frac{L}{2 \bullet x/L}$$

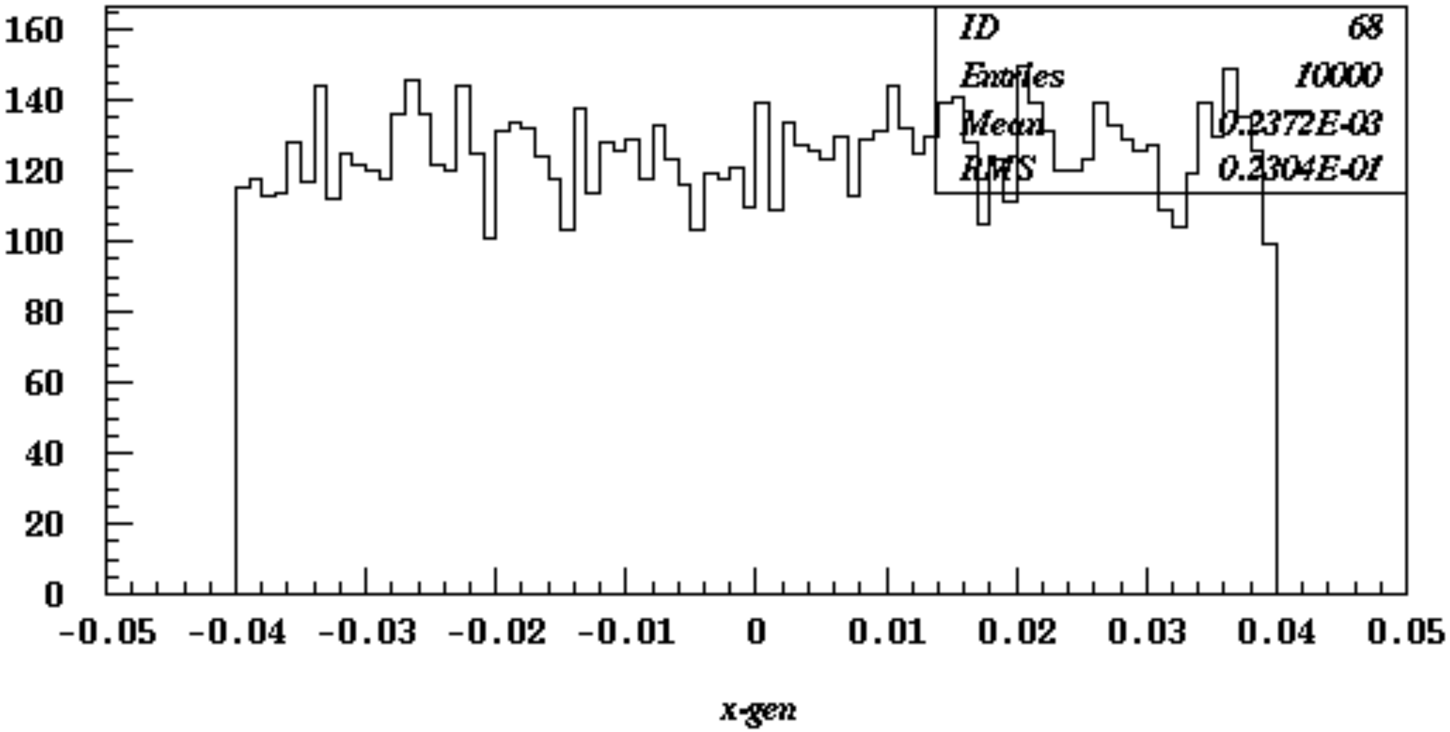
$$\Rightarrow \quad R = \frac{D^2 + x^2}{2x}, \quad p = 0.3 \cdot q \cdot B\frac{D^2 + x^2}{2x}$$

Depending of the sign of the charged particle, x can assume either positive or negative Values. Dispersion on x is constant and it depends of the kind of measurement.

Typical values: q=|1|, B=1.4 Tesla, D=50 cm, x ∈[-4,4] cm, δx=1 cm (Normal distribution)

In reality also the error due to Multiple Scattering is present (field B in material)

A uniform distribution is simulated for charged particles in [-4, 4] cm
And the distribution of residuals of momenta p is computed.

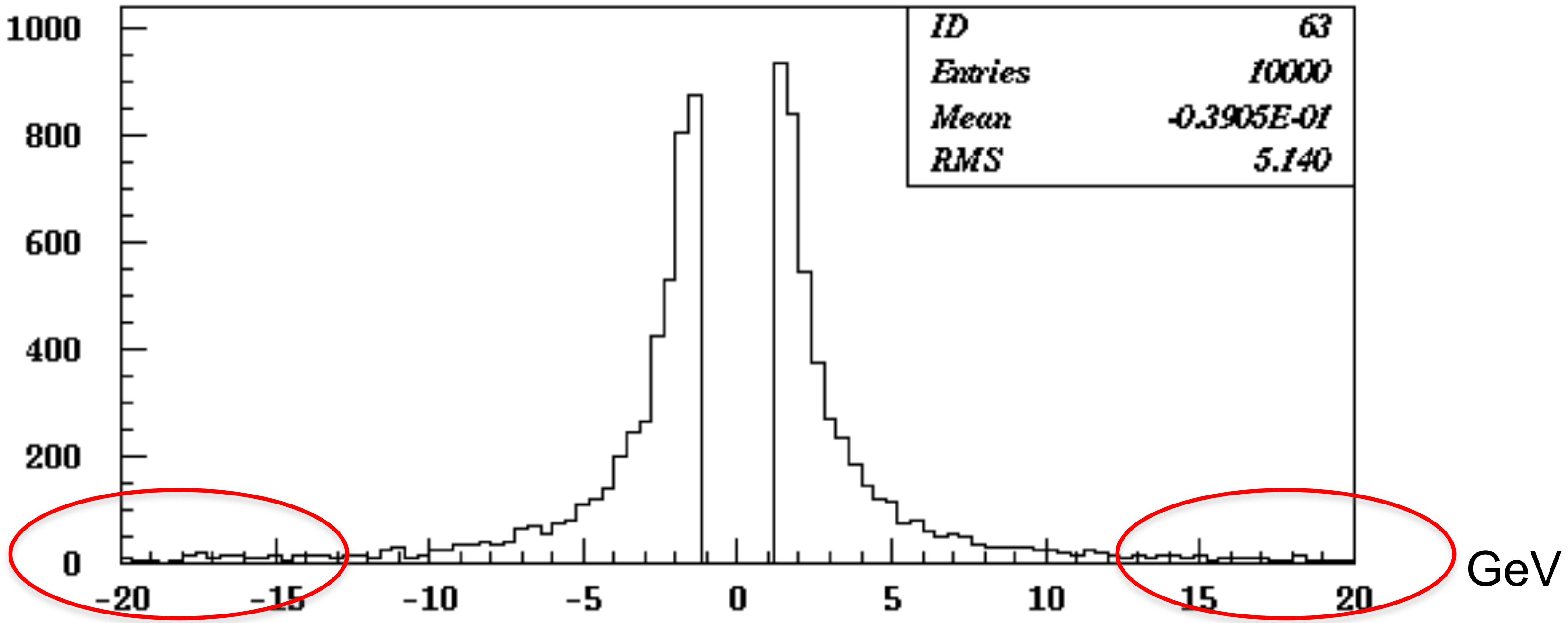


For each generated value of x the corresponding momentum is computed :

p(true)=f(x-random in [-4,4])

p(measured)=p=f(x-random in [-4,4]+$\delta x_{\mathrm{Gauss}}$)

Distribution of the (true) momenta generated:

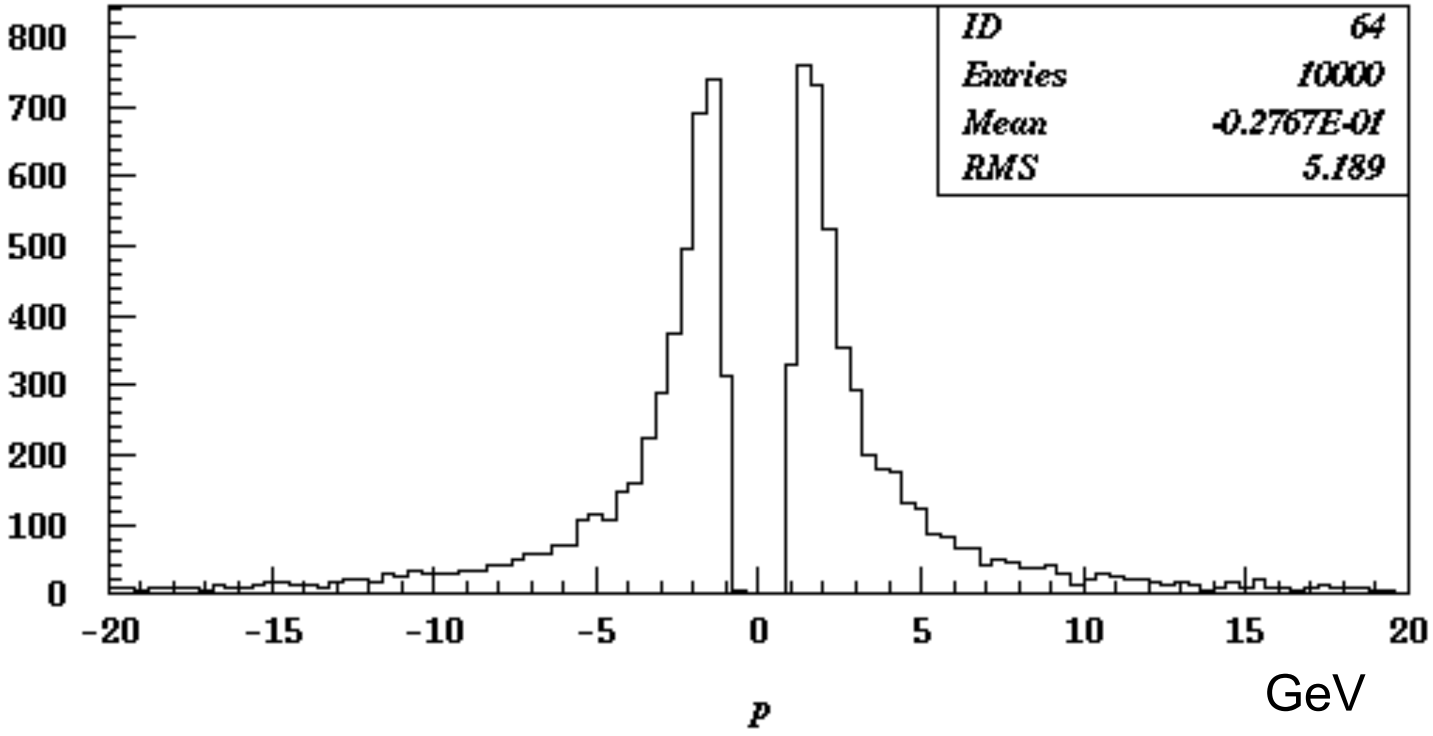


Large momenta of negative charge, small x ≤ 0

Large momenta of positive charge, small x ≥ 0

|p| min for |x| = 0.04 m

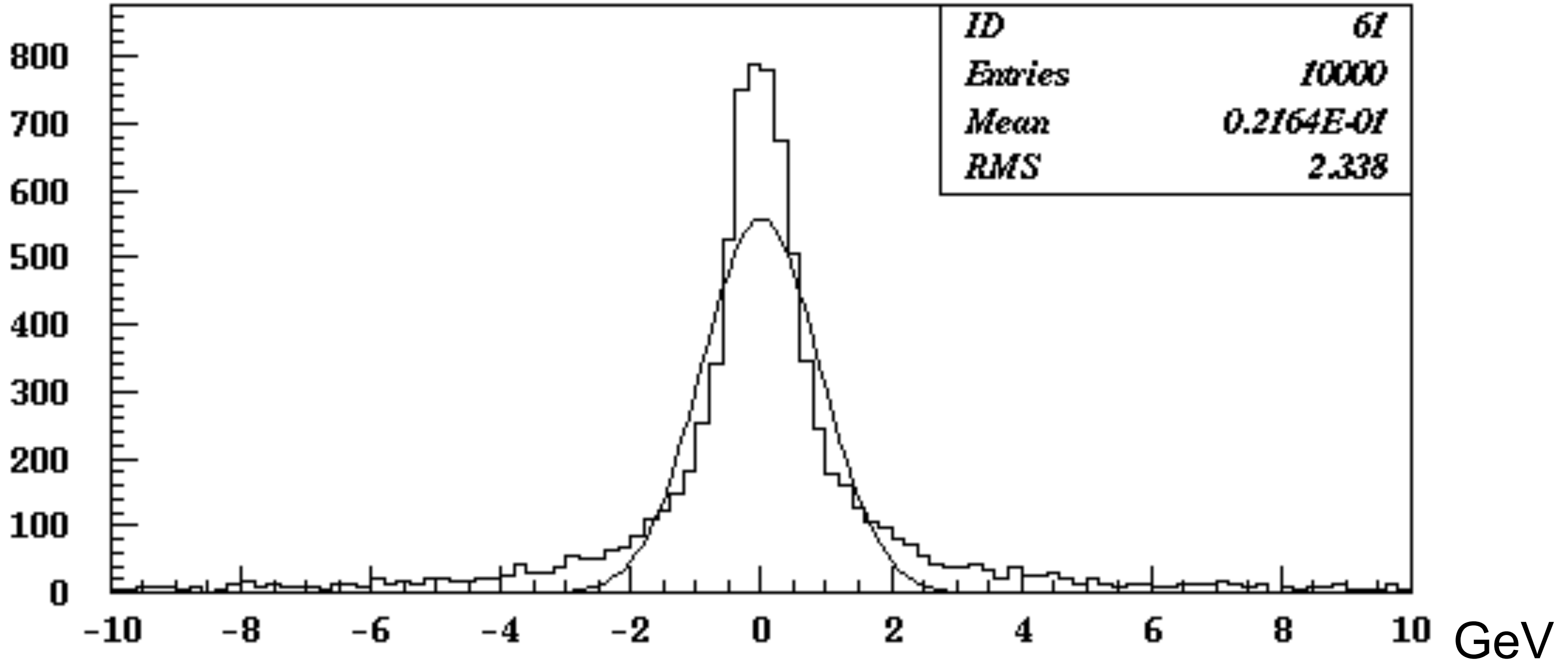
Distribution of the “measured” momenta,
i.e. including the Gaussian error of the measure
on the “measurement” of the position x of the trace



GeV

Distribution of residuals:

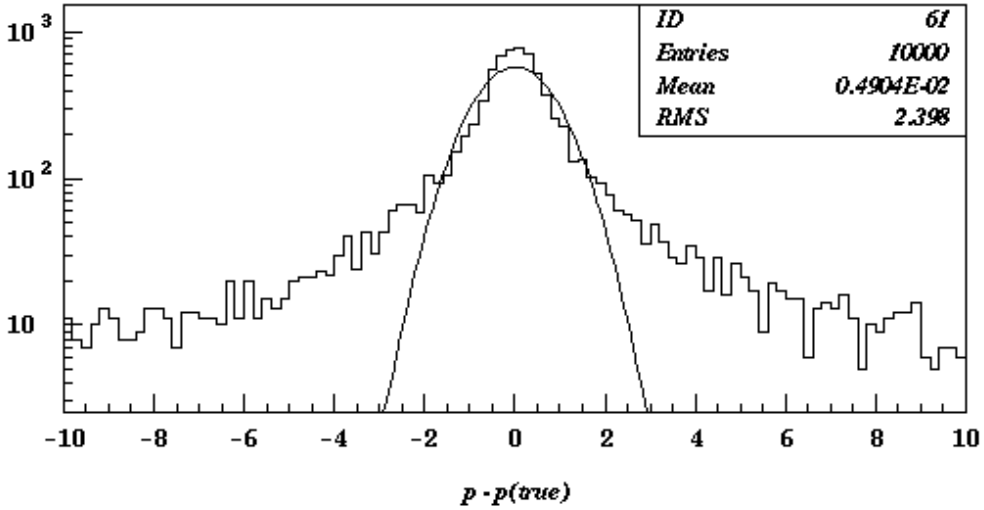
The distribution is NOT a Gaussian !



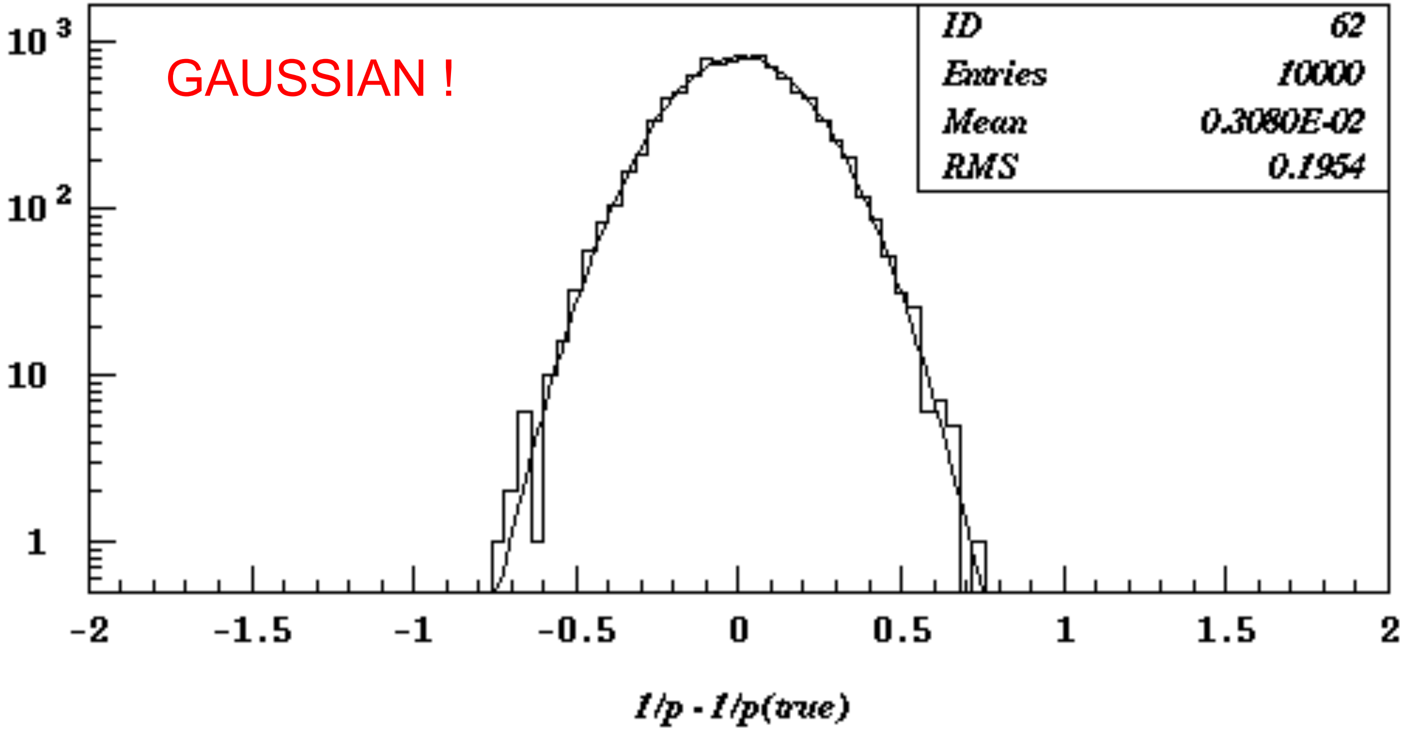
And actually the Theorem of Central Limit cannot be applied

(as $p \propto 1/x$ and $\delta p \propto \delta x/x^2$)

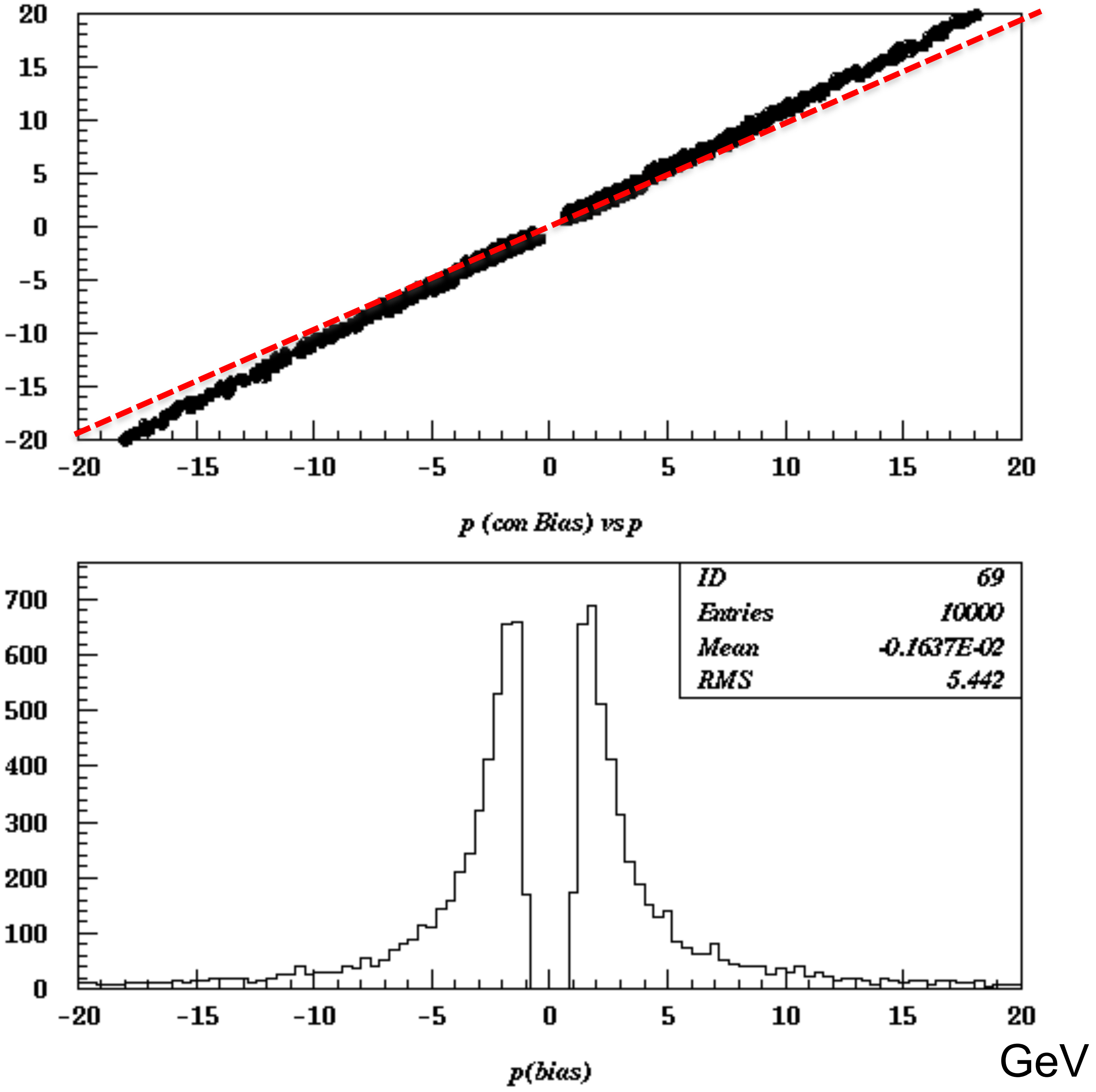
Note: the curve corresponds to a Gaussian fit with free mean, variance and normalization *in semilogaritmic scale…*



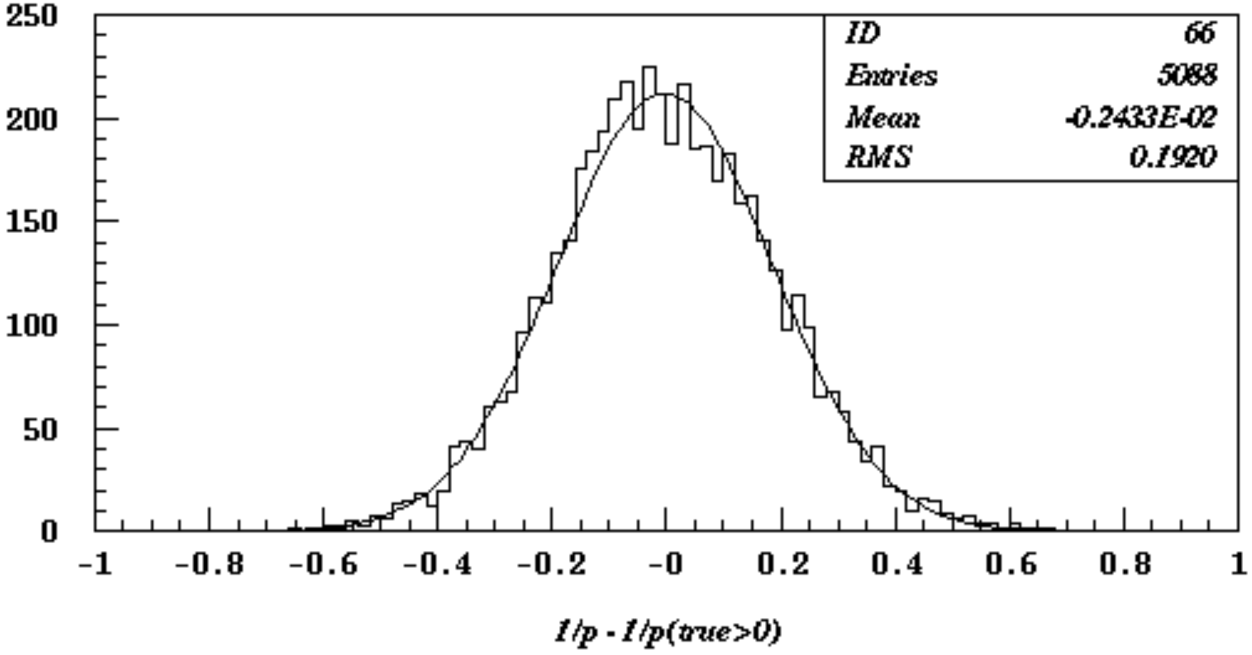
However if the distribution of residuals 1/p is considered:

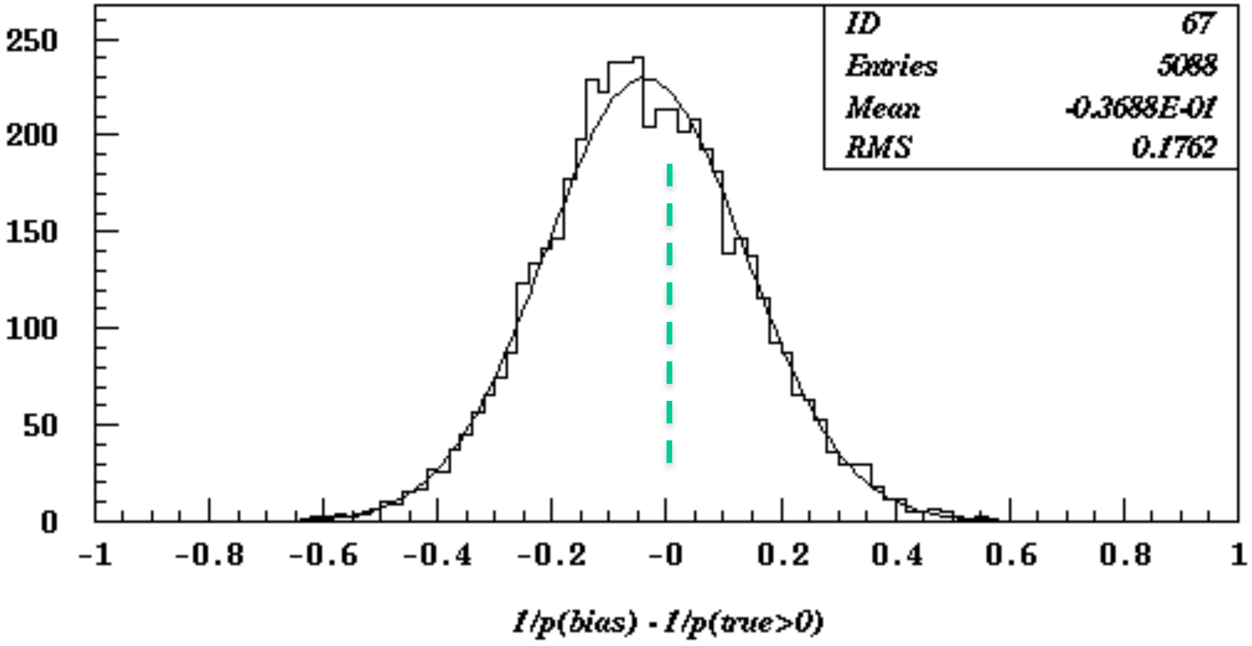


If B is mistaken by +10% (BIAS) then we obtain:

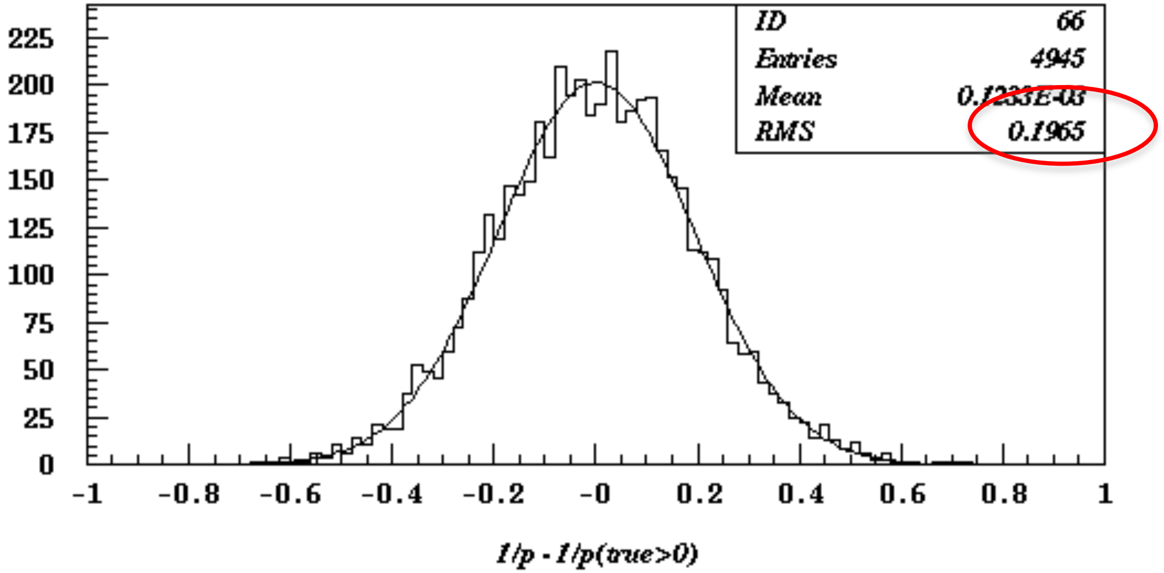


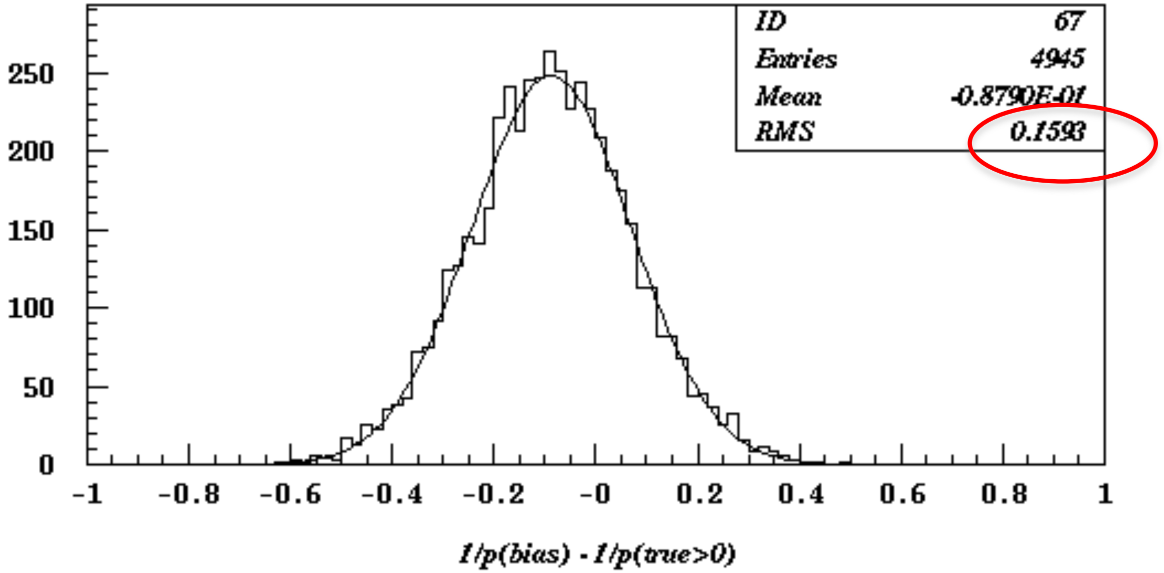
With a Bias of 10% on the Magnetic Field:





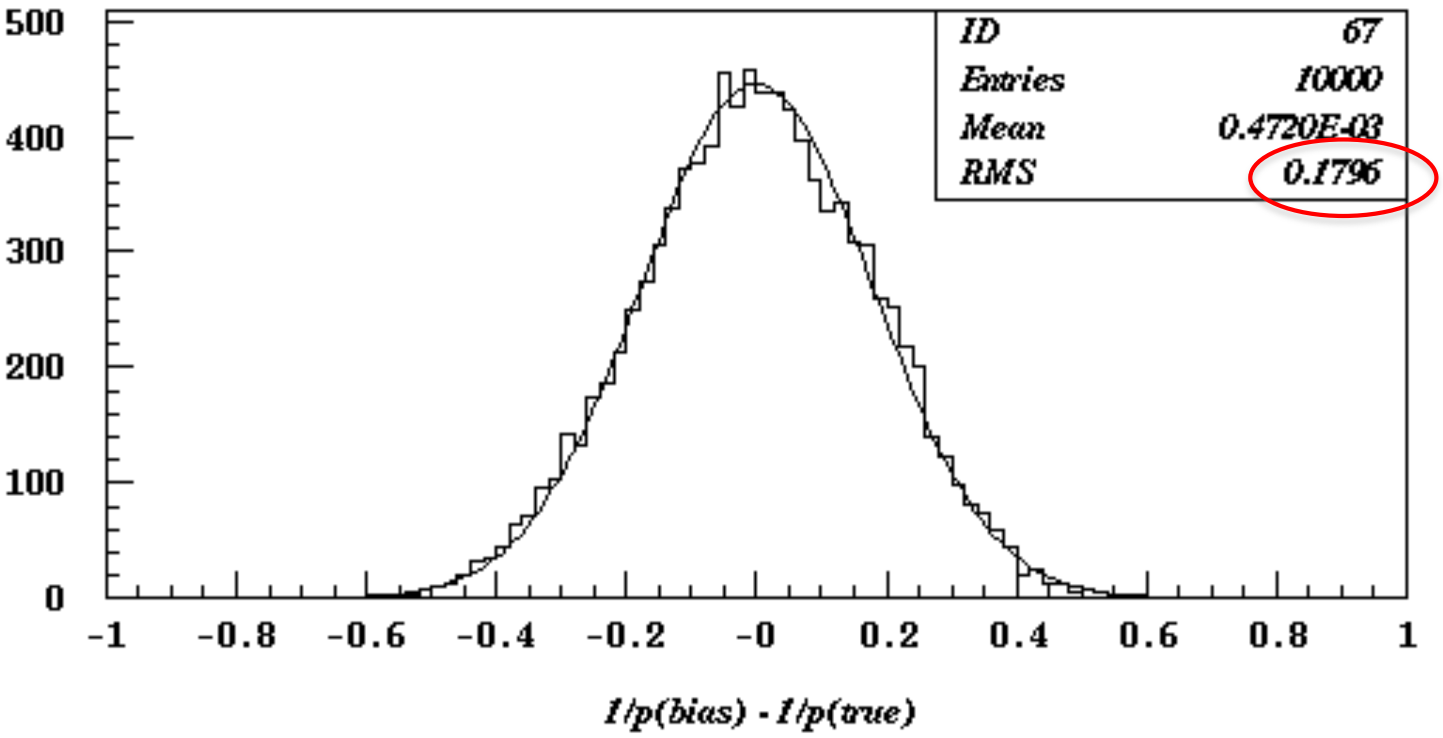
Note that in this case the BIAS affects also the variance.
That is well observed with a Bias of 30% on the Magnetic Field



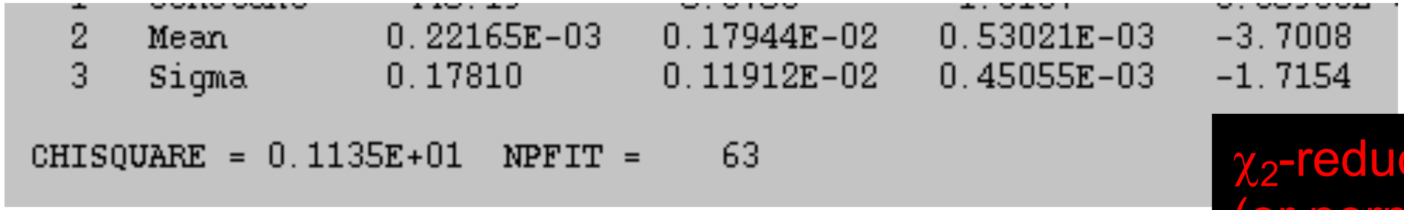


SINCE the relational dependence between p and x is far away from linearity
(even if p e B are proportional between them)

Finally note that the BIAS is evident only if one consider only p>0.
In case one does not distinguish between positive and negative charges the result is:



Variance with bias effect, anyhow



```
  2   Mean        0.22165E-03    0.17944E-02    0.53021E-03    -3.7008
  3   Sigma       0.17810        0.11912E-02    0.45055E-03    -1.7154

 CHISQUARE = 0.1135E+01   NPFIT =     63
```

χ_2-reduced=1.13
(or normalized)

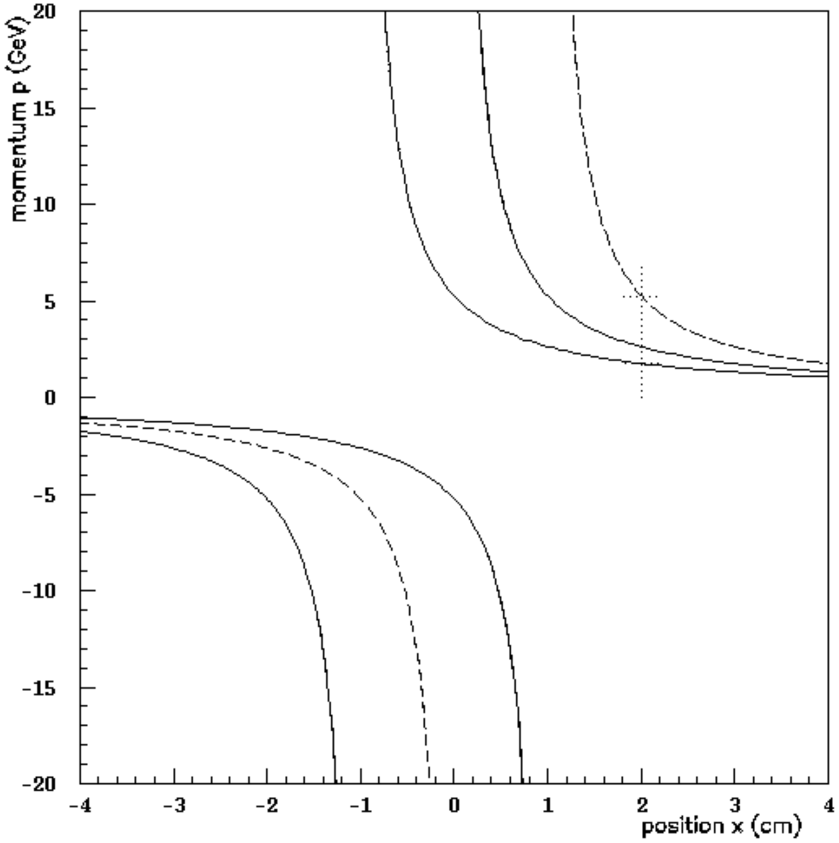
Fit at 60 Degrees of Freedom (63 bins ≠ 0 and 3 free parameters)

Program used: PAW, library of CERN, (in FORTRAN ambiance)
old of about 15 years: http://paw.web.cern.ch/paw/ .
There is the more recent tool, ROOT (in ambiance C++): http://root.cern.ch/drupal/

Script in http://www.pd.infn.it/~stanco/didattica/Stat-An-Dati/carica.kumac
that use the file :
http://www.pd.infn.it/~stanco/didattica/Stat-An-Dati/carica.for

Now that we properly understood the physics and the data analysis, one may try more sophisticated analysis.
One measures **x** but, at the end, one is interested to quote the momentum **p**, within a proper interval, i.e. an interval corresponding to a certain *Confidence Level* and with the proper coverage (68% C.L. should really be 68% !)

For a **frequentist**, all the information is contained in the Likelihood, which is invariant by transformation.The dispersion of 1 cm in **x** can be used to compute the 68% *Confidence Interval* via the Neyman belt-construction
(note: the usual error transformation is a valid approximation ONLY for small errors on x, i.e. small Δx, since it is based on linearization)

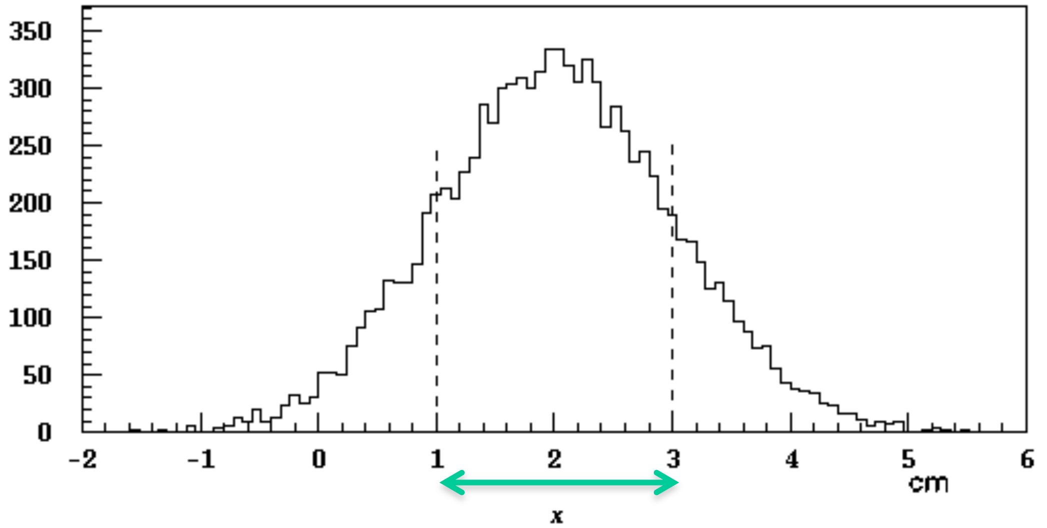


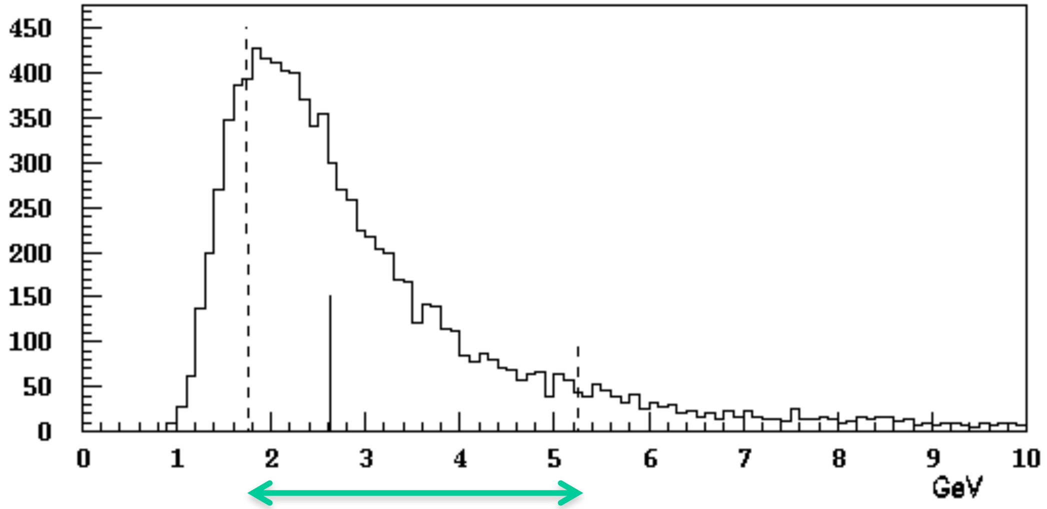
Of course, a good measurement can be obtained only for low momenta, i.e. |x|>2 cm (2 sigma limit).
Note that for small x there are two disjoint regions of validity for the momentum

Note that the Confidence Interval is asymmetric with respect to the “best value”.
E.g. for x=2 cm ➔ p=$2.62^{+2.63}_{-0.87}$ GeV, obtained from $p \propto \frac{1}{x \pm \sigma}$

To have a more careful look at the procedure it is worth to make a simulation of a single data point.

Take the mean measurement x=2 cm and simulate 10,000 measurements with 1 cm dispersion, then plot the corresponding PDF for the momentum:

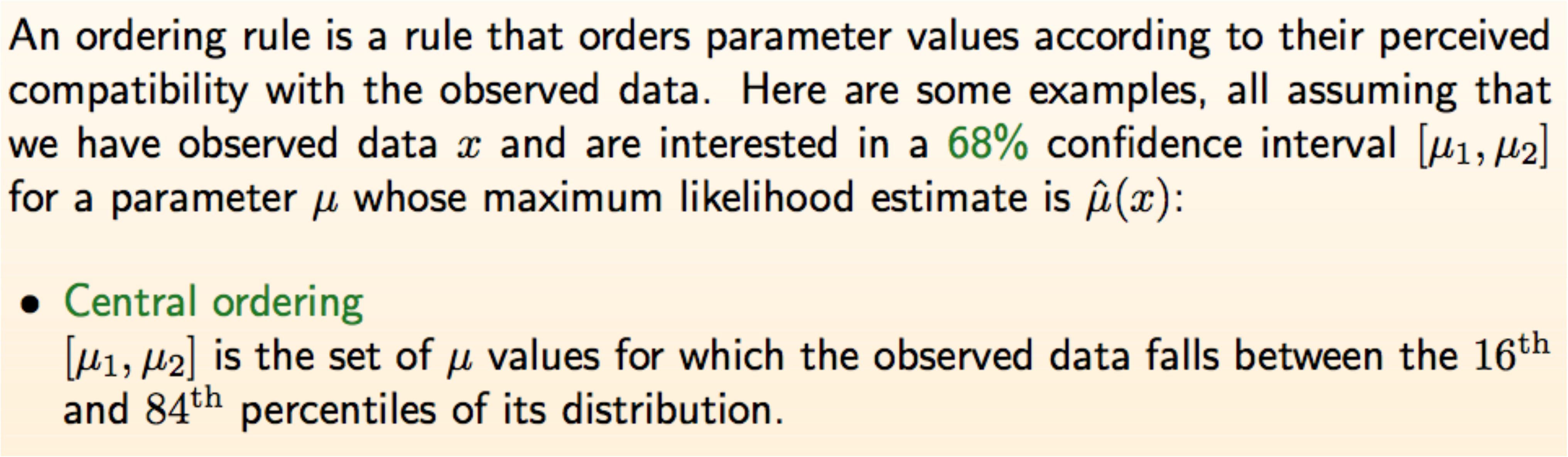




$p=2.62^{+2.63}_{-0.87}$ GeV

The C.I. (*Confidence Interval*) for the momentum is computed around the “mean”. One may like to choose a different way to define it ! i.e. a different **ORDERING** rule

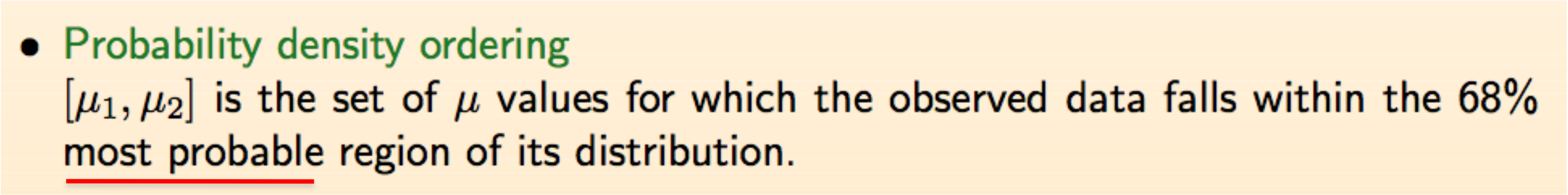
The choice of the ORDERING rule



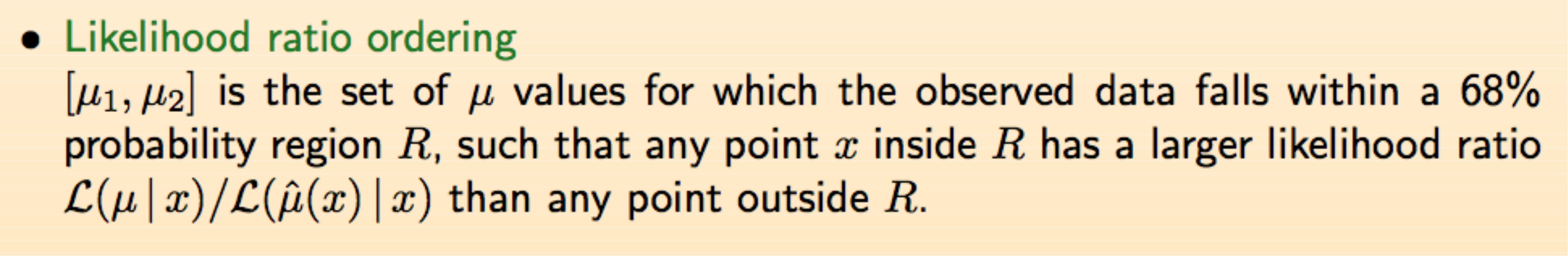
An ordering rule is a rule that orders parameter values according to their perceived compatibility with the observed data. Here are some examples, all assuming that we have observed data x and are interested in a 68% confidence interval $[\mu_1, \mu_2]$ for a parameter μ whose maximum likelihood estimate is $\hat{\mu}(x)$:

- Central ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls between the 16^{th} and 84^{th} percentiles of its distribution.

The latter is the rule chosen in the previous slide, but here are some other examples:

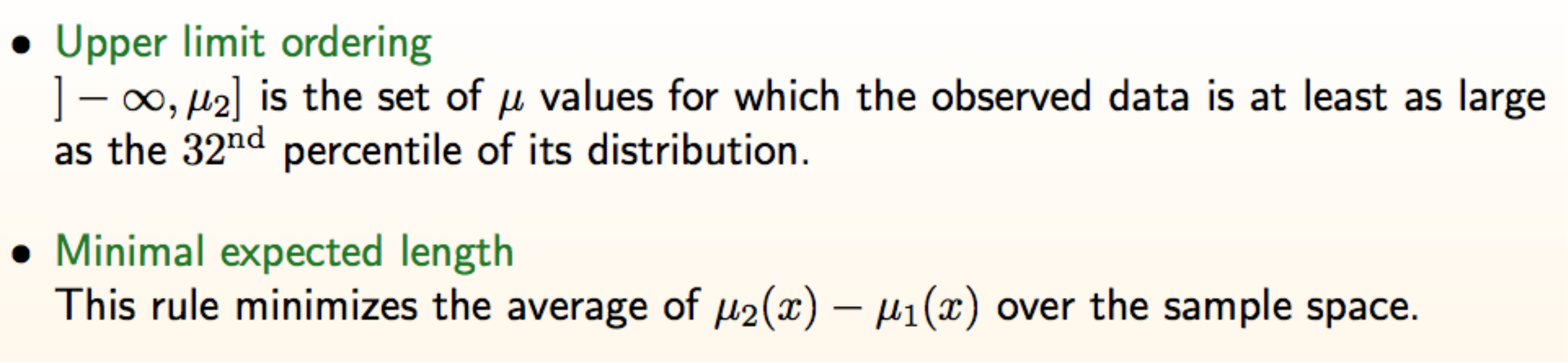


- Probability density ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within the 68% most probable region of its distribution.



- Likelihood ratio ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within a 68% probability region R, such that any point x inside R has a larger likelihood ratio $\mathcal{L}(\mu \,|\, x)/\mathcal{L}(\hat{\mu}(x) \,|\, x)$ than any point outside R.

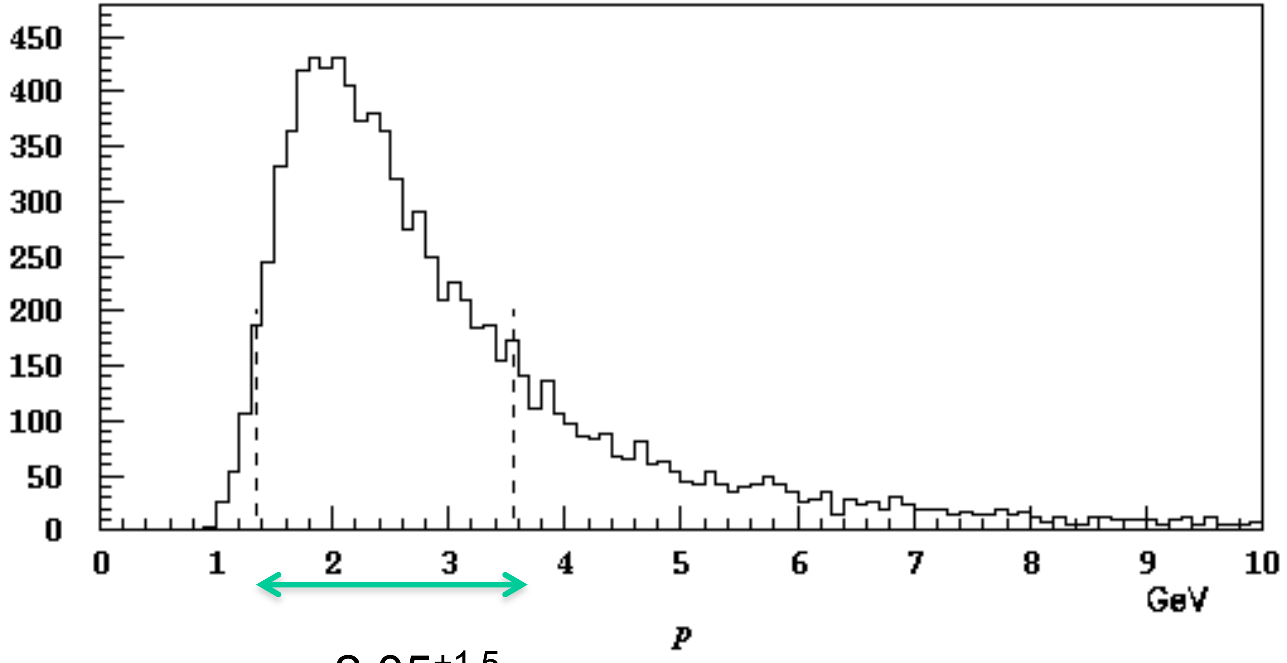
This the rule chosen by Feldman-Cousins in their (in)famous method .



- Upper limit ordering
  $]-\infty, \mu_2]$ is the set of μ values for which the observed data is at least as large as the 32^{nd} percentile of its distribution.
- Minimal expected length
  This rule minimizes the average of $\mu_2(x) - \mu_1(x)$ over the sample space.

As a matter of fact, choose your own rule relying on the kind of measurement you are doing !
(and report it in your scientific paper)

My choice for the momentum measurement is the maximum probability (or the Likelihood ratio, the same thing in this example)



$p=2.05^{+1.5}_{-0.7}$

Rather different C.I.

p in [1.55-3.75]
against
p in [1.75-5.25]

at the same C.L. of 68%

a useful trick: Monte Carlo, Monte Carlo, Monte Carlo…

We may also want to look at the **Bayes** posterior by using the Jeffreys' prior

The likelihood for the momentum **p** is $L(p; p_{best}) \propto e^{-\frac{1}{2}\left(\frac{1/p - 1/p_{best}}{\sigma}\right)^2}$

where sigma is the estimated dispersion on the curvature, i.e. 0.1954 GeV^{-1} (see bottom plot of slide 32)

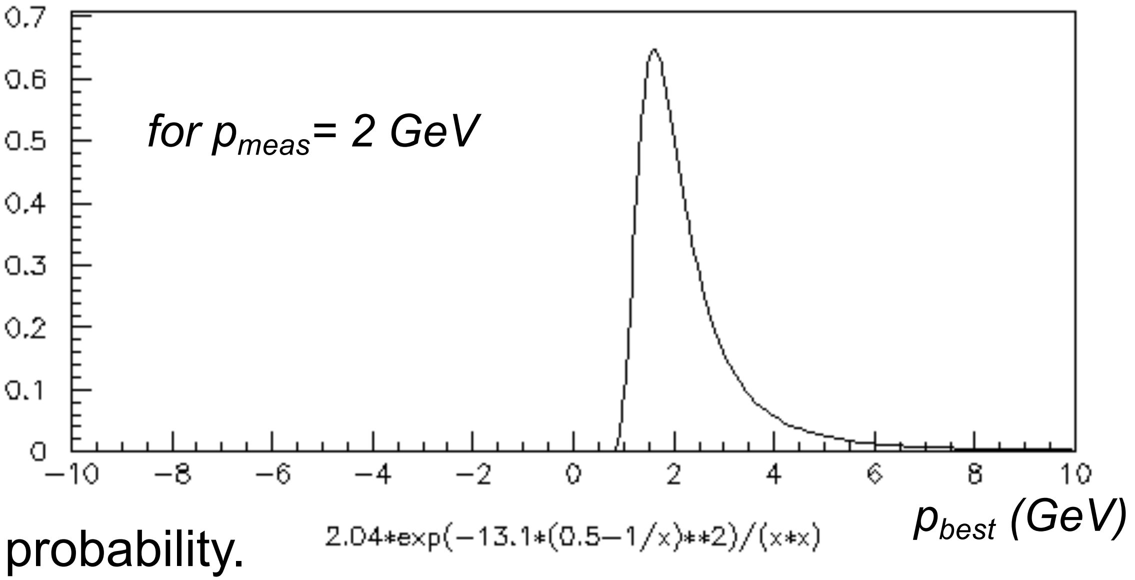
The Jeffreys' prior is $\pi_J(p_{best}) \propto \dfrac{1}{p_{best}^2}$ (p_{best} is a variable !)

since $\dfrac{1}{\sigma^2}\left\{-\dfrac{3}{p_{best}^4} + \dfrac{2}{p_{best}^3}\int \dfrac{1}{p} f\left(\dfrac{1}{p}\right) d\left(\dfrac{1}{p}\right)\right\} = \dfrac{1}{\sigma^2}\left\{-\dfrac{3}{p_{best}^4} + \dfrac{2}{p_{best}^4}\right\}$

And the *properly normalized* posterior becomes: $P(p_{best}; p) = \dfrac{e^{-\frac{1}{2}\left(\frac{1/p - 1/p_{best}}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma\, p_{best}^2}$

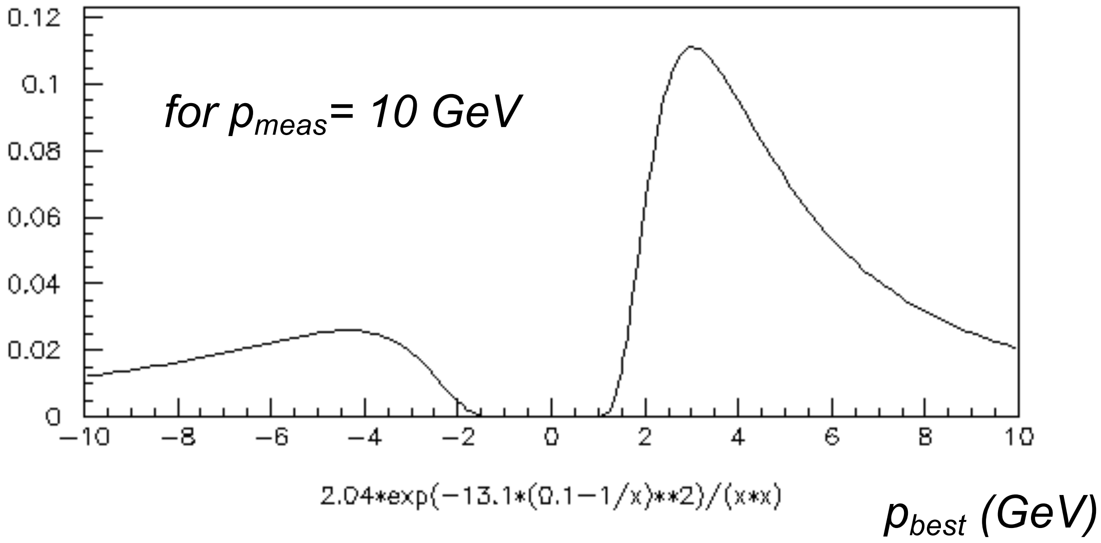
Which kind of PDF is the previous posterior ?

$$P(p_{best};p)=\frac{e^{-\frac{1}{2}\left(\frac{1/p-1/p_{best}}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma p_{best}^2}$$


for p_{meas}= 2 GeV

p_{best} (GeV)

Indeed it gives good insight of the probability.

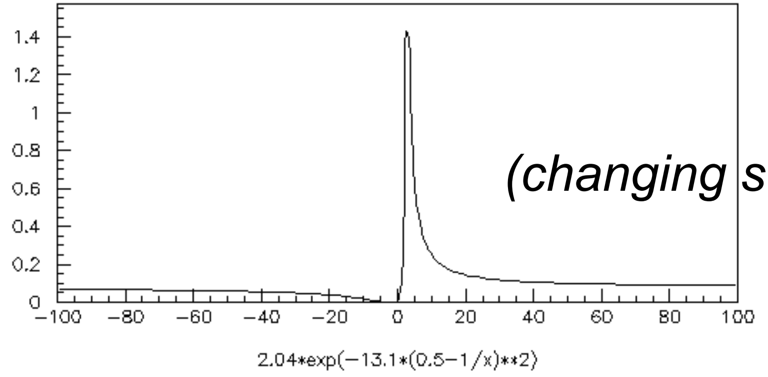

for p_{meas}= 10 GeV

p_{best} (GeV)

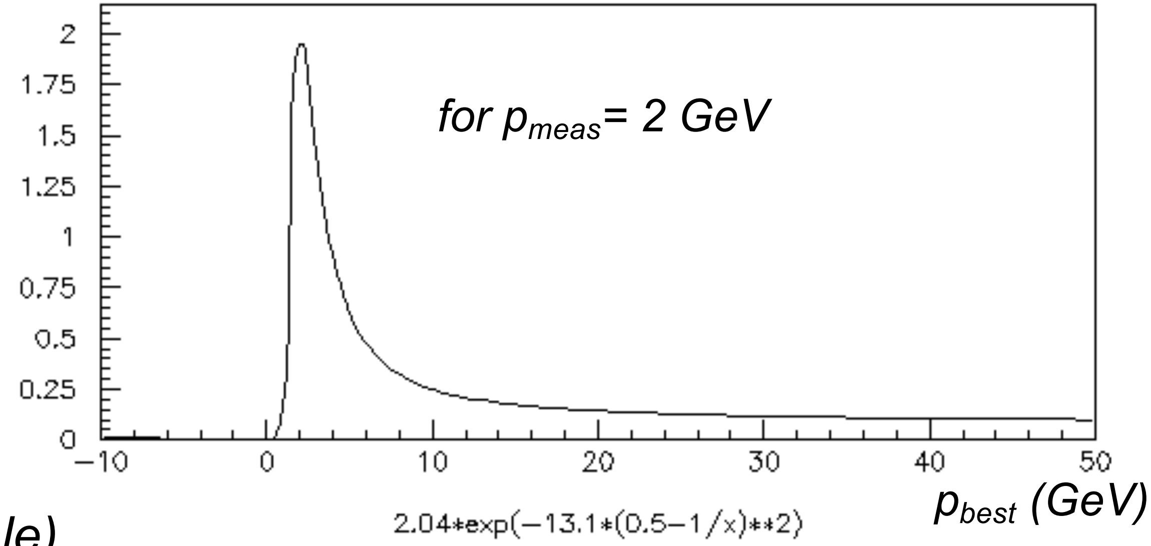
and good estimate of the probability
to compute the wrong charge
but it is a very biased estimate of p_{best}
(the two maxima never exceed $1/\sigma\sqrt{2}$
i.e. 3.6 GeV)

While the uniform Prior works rather badly…

$$P(p_{best};p)=\frac{e^{-\frac{1}{2}\left(\frac{1/p-1/p_{best}}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma}$$



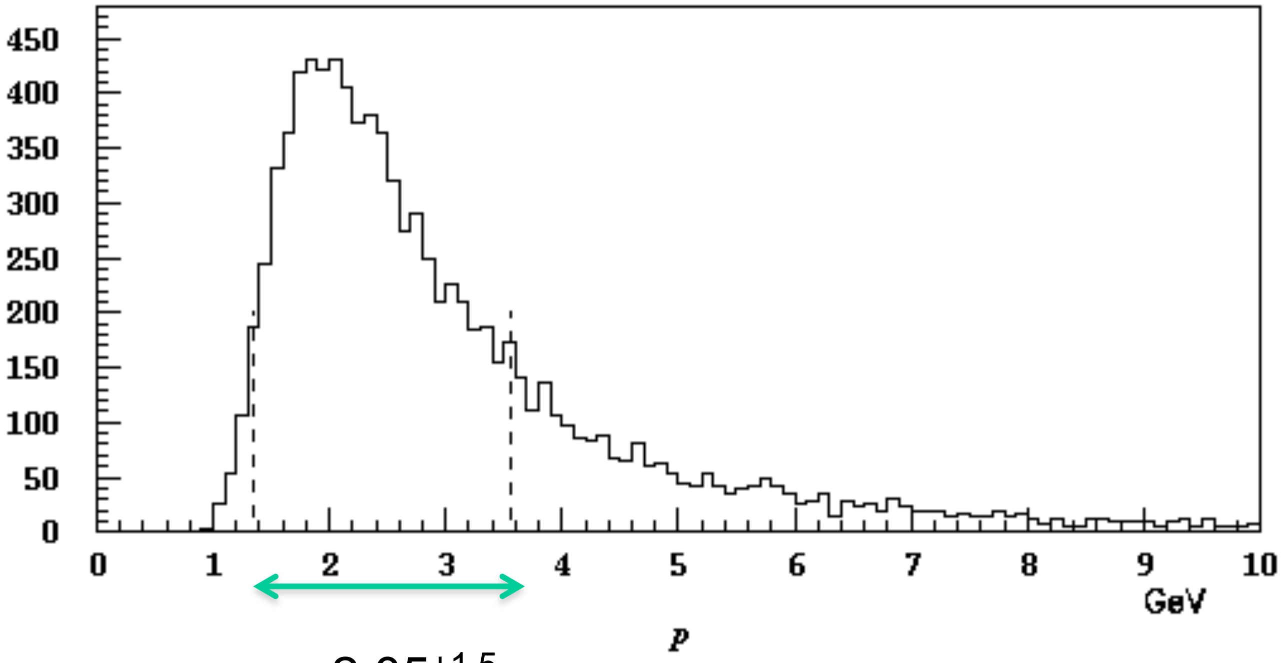
(changing scale)



for p_{meas}= 2 GeV

To me, this is an example where the Frequentist approach works better than the Baysian one.
In the Frequentist approach it is easier to understand the constraints of the measurements and put a priori cuts to define properly the region of good measure, i.e. it is an example where to deal first with data, without assuming them to include all the information.

A note on the handling of the Confidence Interval chosen in slide 42:



$p=2.05^{+1.5}_{-0.7}$

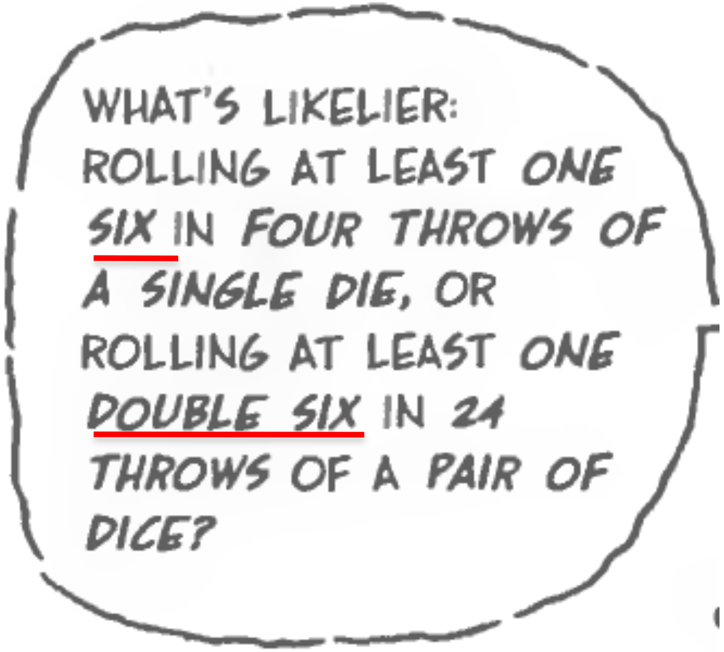
The final decision depends on what one is more interested to (CI or Central Value).
In case of CV the estimation has to be consistent (i.e. with the tendency to the true value).
In frequentist language the previous choice is not consistent.
In Bayesian it corresponds to use a flat distribution (Posterior=Likelihood) and it becomes more “understandable”, even if with the drawbacks of previous slide.

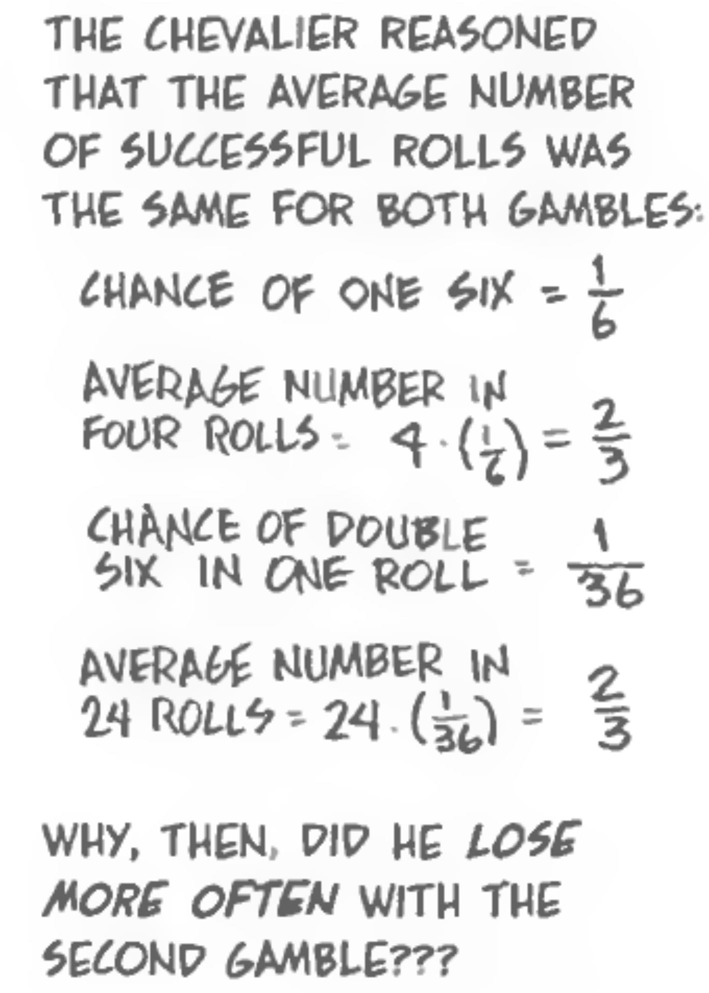
BACKUP SLIDES

PROBABILITY

Chevalier de Méré → Blaise Pascal & Pierre de Fermat (1654)

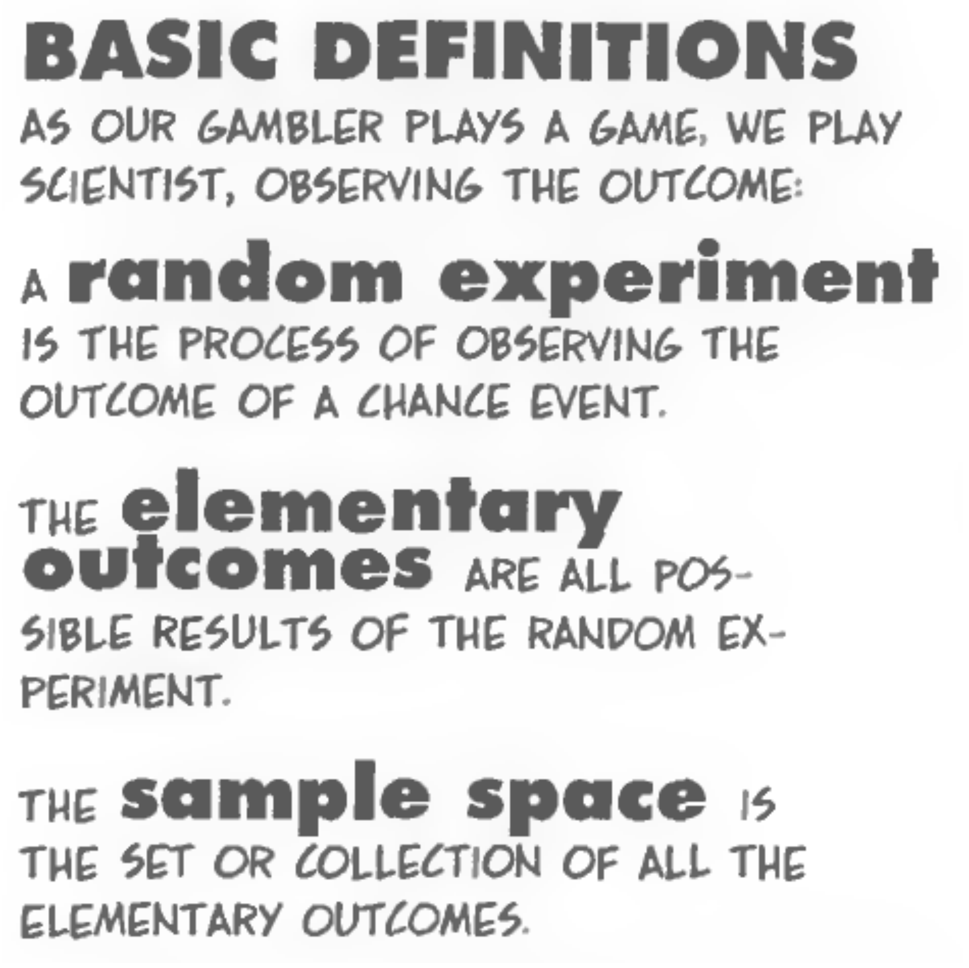
de Méré looked at the 2 cases: 1) 4 rolls of a dice
2) 24 rolls of 2 dices


WHAT'S LIKELIER: ROLLING AT LEAST ONE SIX IN FOUR THROWS OF A SINGLE DIE, OR ROLLING AT LEAST ONE DOUBLE SIX IN 24 THROWS OF A PAIR OF DICE?


THE CHEVALIER REASONED THAT THE AVERAGE NUMBER OF SUCCESSFUL ROLLS WAS THE SAME FOR BOTH GAMBLES:
CHANCE OF ONE SIX = $\frac{1}{6}$
AVERAGE NUMBER IN FOUR ROLLS = $4 \cdot (\frac{1}{6}) = \frac{2}{3}$
CHANCE OF DOUBLE SIX IN ONE ROLL = $\frac{1}{36}$
AVERAGE NUMBER IN 24 ROLLS = $24 \cdot (\frac{1}{36}) = \frac{2}{3}$
WHY, THEN, DID HE LOSE MORE OFTEN WITH THE SECOND GAMBLE???

Pascal and Fermat exchanged some letters and in few months basic of the new Science was settled (1654).

Mistake: formulation of multiplicative law on the “different cases”
(the “event” corresponds to the possible outcome, NOT to the roll)



BASIC DEFINITIONS

AS OUR GAMBLER PLAYS A GAME, WE PLAY SCIENTIST, OBSERVING THE OUTCOME:

A **random experiment** IS THE PROCESS OF OBSERVING THE OUTCOME OF A CHANCE EVENT.

THE **elementary outcomes** ARE ALL POSSIBLE RESULTS OF THE RANDOM EXPERIMENT.

THE **sample space** IS THE SET OR COLLECTION OF ALL THE ELEMENTARY OUTCOMES.

We have instead to define the set of possible outcomes and to construct the basic laws.
We introduce the concept of “event” to which apply the probability and its multiplicative law (to be defined).

EVENT : RANDOM VARIABLE

Bayesian Hypothesis Testing (1)

The Bayesian approach to hypothesis testing is to calculate posterior probabilities for all hypotheses in play. When testing H_0 versus H_1, Bayes' theorem yields:

$$\pi(H_0 \,|\, x) \;=\; \frac{p(x \,|\, H_0)\,\pi_0}{p(x \,|\, H_0)\,\pi_0 \;+\; p(x \,|\, H_1)\,\pi_1}, \qquad \Rightarrow \text{p(x)=Int}(H_0,H_1)$$

$$\pi(H_1 \,|\, x) \;=\; 1 \;-\; \pi(H_0 \,|\, x),$$

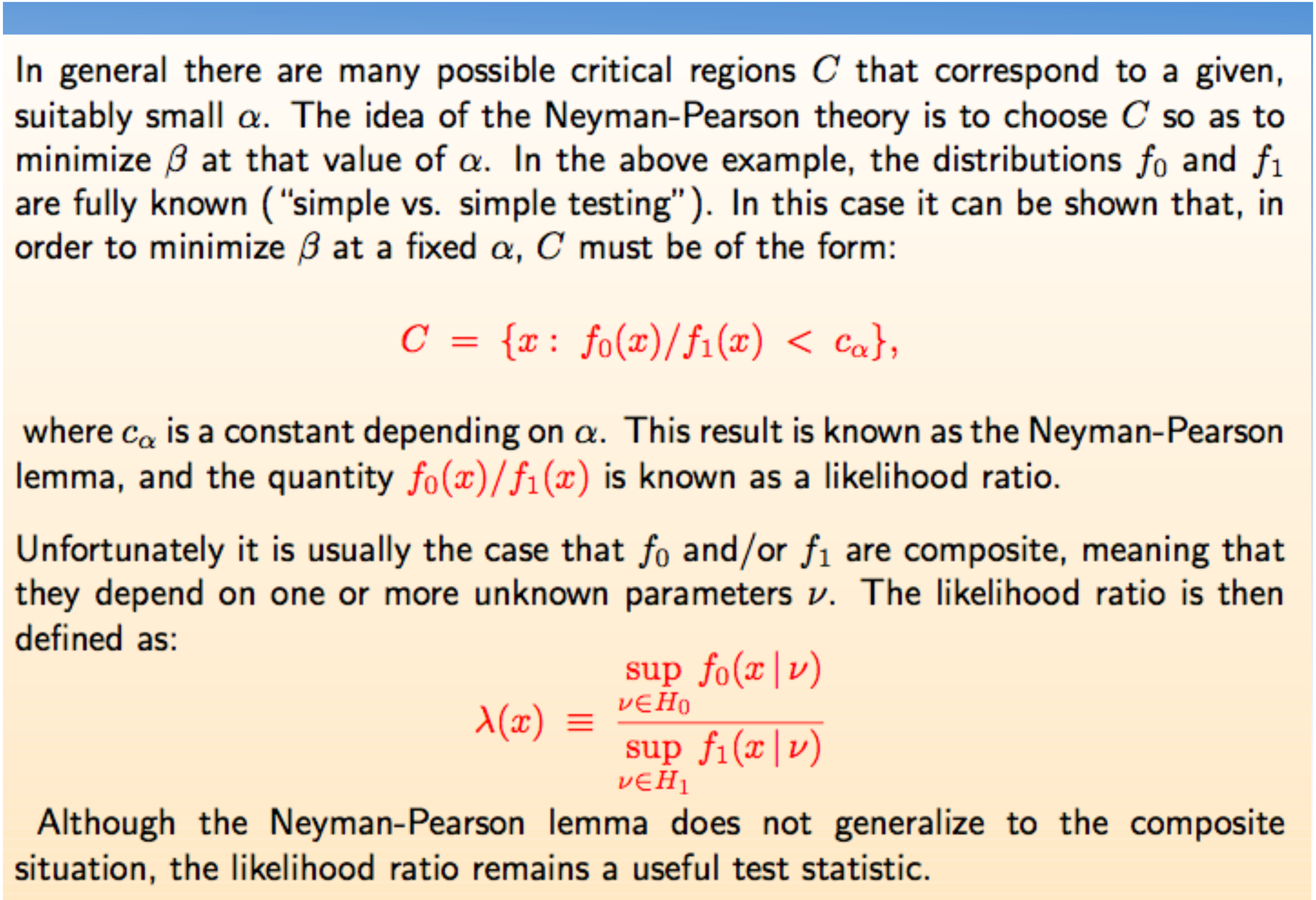
where π_i is the prior probability of H_i, $i = 0, 1$.

If $\pi(H_0 \,|\, x) < \pi(H_1 \,|\, x)$, one rejects H_0 and the posterior probability of error is $\pi(H_0 \,|\, x)$. Otherwise H_0 is accepted and the posterior error probability is $\pi(H_1 \,|\, x)$.

In contrast with frequentist Type-I and Type-II errors, Bayesian error probabilities are fully conditioned on the observed data. It is often interesting to look at the evidence against H_0 provided by the data alone. This can be done by computing the ratio of posterior odds to prior odds and is known as the Bayes factor:

$$B_{01}(x) \;=\; \frac{\pi(H_0 \,|\, x)/\pi(H_1 \,|\, x)}{\pi_0/\pi_1}$$

In the absence of unknown parameters, $B_{01}(x)$ is a likelihood ratio.



In general there are many possible critical regions C that correspond to a given, suitably small α. The idea of the Neyman-Pearson theory is to choose C so as to minimize β at that value of α. In the above example, the distributions f_0 and f_1 are fully known ("simple vs. simple testing"). In this case it can be shown that, in order to minimize β at a fixed α, C must be of the form:

$$C = \{x : f_0(x)/f_1(x) < c_\alpha\},$$

where c_α is a constant depending on α. This result is known as the Neyman-Pearson lemma, and the quantity $f_0(x)/f_1(x)$ is known as a likelihood ratio.

Unfortunately it is usually the case that f_0 and/or f_1 are composite, meaning that they depend on one or more unknown parameters ν. The likelihood ratio is then defined as:

$$\lambda(x) \equiv \frac{\sup\limits_{\nu \in H_0} f_0(x \,|\, \nu)}{\sup\limits_{\nu \in H_1} f_1(x \,|\, \nu)}$$

Although the Neyman-Pearson lemma does not generalize to the composite situation, the likelihood ratio remains a useful test statistic.

The Neyman-Pearson Theory of Testing (3)

The Neyman-Pearson theory of testing is most useful in quality-control applications, when a given test has to be repeated on a large sample of identical items. In HEP we use this technique to select events. For example, if we want to measure the mass of the top quark, for each event in some appropriate trigger stream we set H_0 to the hypothesis that the event contains a top quark, and choose cuts that minimize the background contamination (β) for a given signal efficiency ($1 - \alpha$).

On the other hand, this approach to testing is not very satisfactory when dealing with one-time testing situations, for example when testing a hypothesis about a new phenomenon such as the Higgs boson or SUSY. This is because the result of a Neyman-Pearson test is either "accept H_0" or "reject H_0", without consideration for the strength of evidence contained in the data. In fact, the level of confidence in the decision resulting from the test is already known *before* the test: it is either $1 - \alpha$ or $1 - \beta$.

There are several ways to address this problem: the frequentist approach uses p values exclusively, whereas the Bayesian one works with posterior hypothesis probabilities, Bayes factors, and p values.

Asymptotic Distribution of the Likelihood Ratio Statistic (1)

The likelihood ratio statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta \setminus \Theta_0$ is

$$\lambda(x_{obs}) \equiv \frac{\sup\limits_{\Theta_0} \mathcal{L}(\theta \,|\, x_{obs})}{\sup\limits_{\Theta} \mathcal{L}(\theta \,|\, x_{obs})} = \frac{\mathcal{L}(\hat{\theta}_0 \,|\, x_{obs})}{\mathcal{L}(\hat{\theta} \,|\, x_{obs})},$$

where $\hat{\theta}_0$ is the maximum likelihood estimate (MLE) under H_0 and $\hat{\theta}$ is the unrestricted MLE.

Note that $0 \leq \lambda(X) \leq 1$. A likelihood ratio test is a test whose rejection region has the form $\{x : \ \lambda(x) \leq c\}$, where c is a constant between 0 and 1.

To calculate p values based on $\lambda(X)$ one needs the distribution of $\lambda(X)$ under H_0:

Under suitable regularity conditions it can be shown that the *asymptotic* distribution of $-2 \ln \lambda(X)$ under H_0 is chisquared with $\nu - \nu_0$ degrees of freedom, where $\nu = \dim \Theta$ and $\nu_0 = \dim \Theta_0$.