Probabilistic Reasoning in Frontier Science

Giulio D’Agostini

Dipartimento di Fisica

Università di Roma La Sapienza

Outline

Applications of probabilistic inference to physics quantities

- pending issues from yesterday

Outline

Applications of probabilistic inference to physics quantities

pending issues from yesterday

- Parametric inference applied to typical detector responses
  - binomial (efficiencies, branching ratios, ‘proportions’)
  - Poisson (counts following “Poisson process”)
  - Gaussian (‘normal errors’, approximation of other pdf)

Outline

Applications of probabilistic inference to physics quantities

pending issues from yesterday

- Parametric inference applied to typical detector responses
  - binomial (efficiencies, branching ratios, ‘proportions’)
  - Poisson (counts following “Poisson process”)
  - Gaussian (‘normal errors’, approximation of other pdf)
- Bayesian inference Vs χ^2 minimization.

Outline

Applications of probabilistic inference to physics quantities

pending issues from yesterday

- Parametric inference applied to typical detector responses
  - binomial (efficiencies, branching ratios, ‘proportions’)
  - Poisson (counts following “Poisson process”)
  - Gaussian (‘normal errors’, approximation of other pdf)
- Bayesian inference Vs χ^2 minimization.
- Some ‘complications’:
  - systematics
  - background
  - measurements at the limit of the detector sensitivity

Outline

Applications of probabilistic inference to physics quantities

pending issues from yesterday

- Parametric inference applied to typical detector responses
  - binomial (efficiencies, branching ratios, ‘proportions’)
  - Poisson (counts following “Poisson process”)
  - Gaussian (‘normal errors’, approximation of other pdf)
- Bayesian inference Vs χ^2 minimization.
- Some ‘complications’:
  - systematics
  - background
  - measurements at the limit of the detector sensitivity
- Propagation of uncertainties

Outline

Applications of probabilistic inference to physics quantities

pending issues from yesterday

- Parametric inference applied to typical detector responses
  - binomial (efficiencies, branching ratios, ‘proportions’)
  - Poisson (counts following “Poisson process”)
  - Gaussian (‘normal errors’, approximation of other pdf)
- Bayesian inference Vs χ^2 minimization.
- Some ‘complications’:
  - systematics
  - background
  - measurements at the limit of the detector sensitivity
- Propagation of uncertainties
- Conclusions

About ‘statistics’

Uncritical use of ‘statistical methods’ can be dangerous!

About ‘statistics’

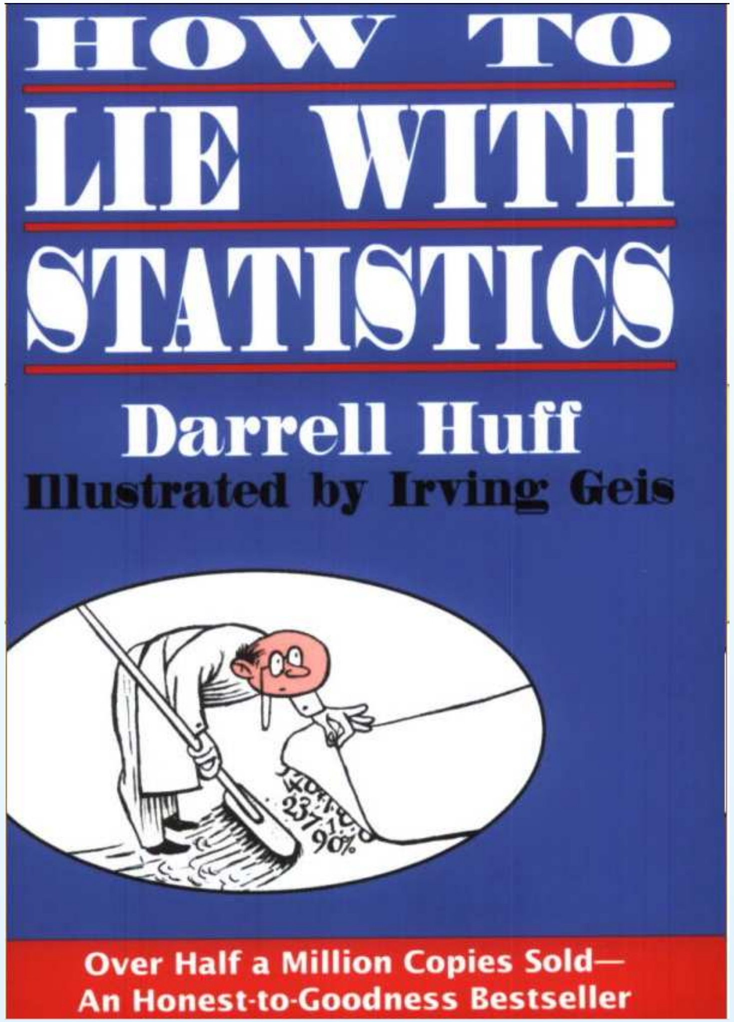
Uncritical use of ‘statistical methods’ can be dangerous!

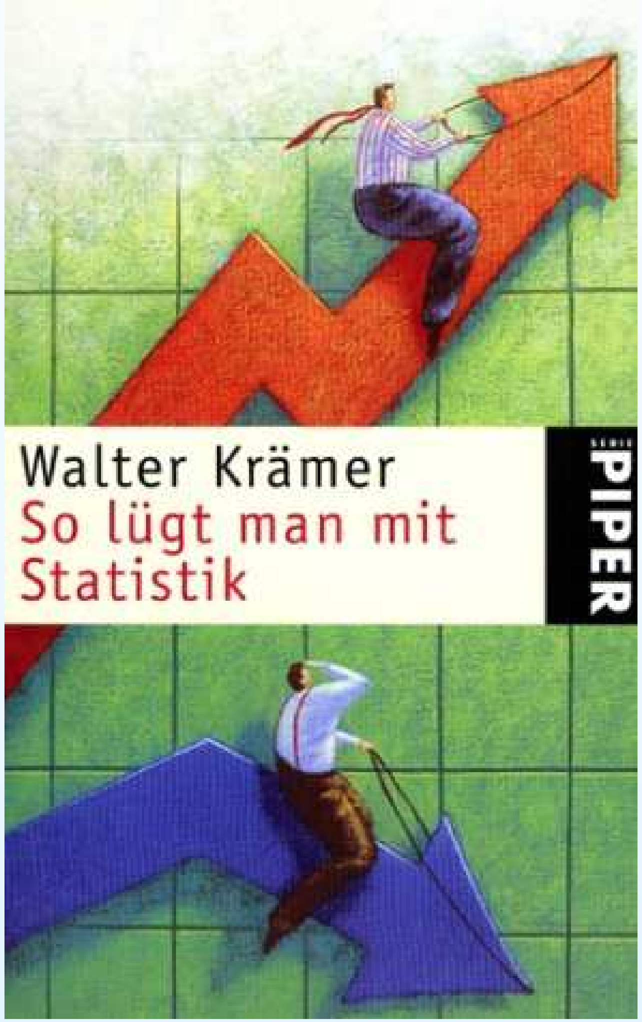
“There are three kinds of lies:
lies, damn lies, and statistics”

(Benjamin Disraeli/Mark Twain)

Damned lies and statistics

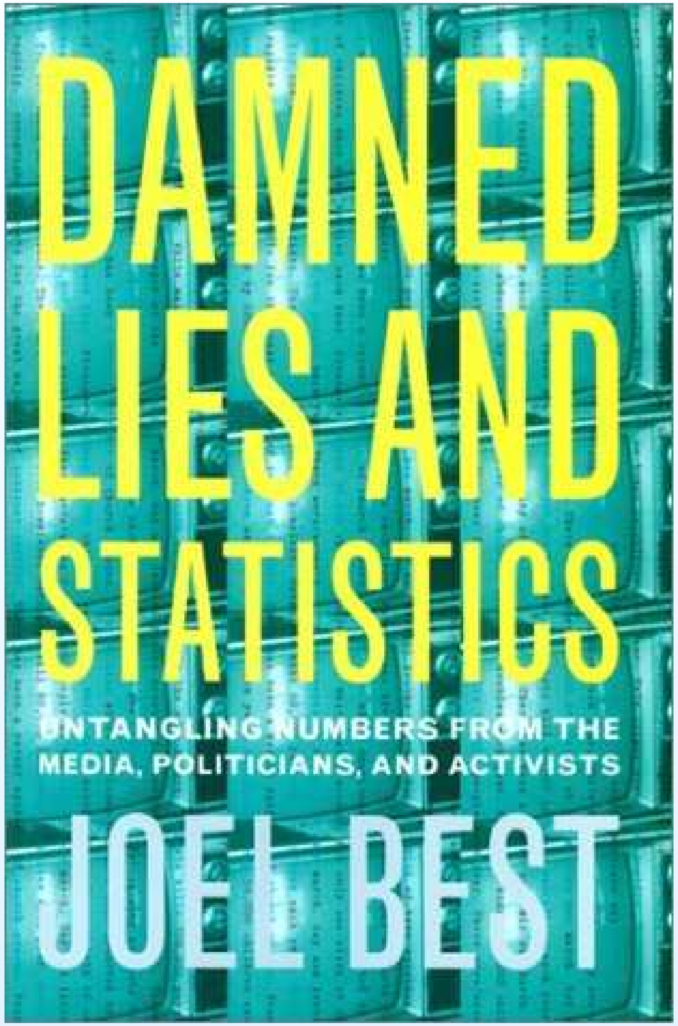
Well known subject

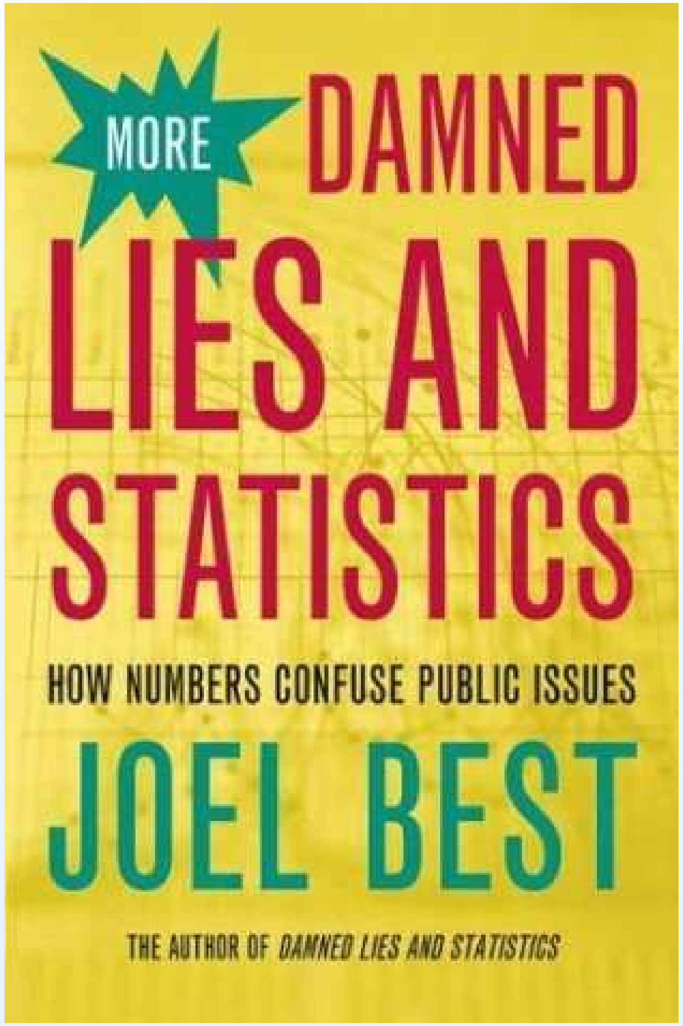




Damned lies and statistics

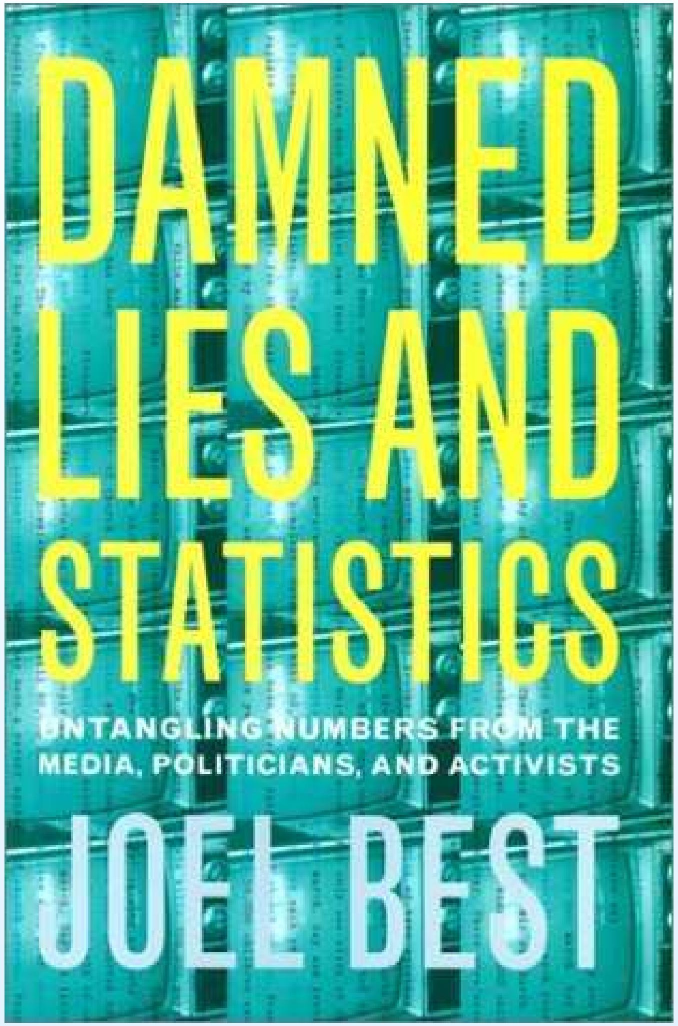
Well known subject, especially in marketing and politics

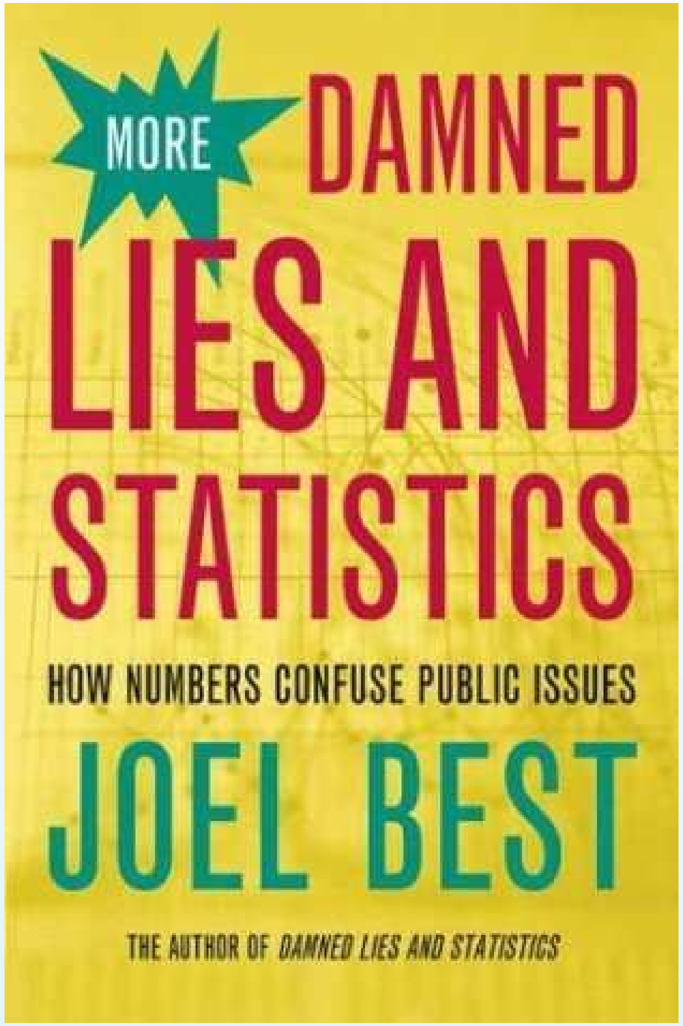




Damned lies and statistics

Well known subject, especially in marketing and politics





but also scientists might get confused!

Summary of 1st lecture

- The main interest in ‘statistics’ of physicists is inference, i.e. how to learn from data

Summary of 1st lecture

- The main interest in ‘statistics’ of physicists is inference, i.e. how to learn from data
- but the ‘prescriptions’ of ‘conventional statistics’ are not satisfactory

Summary of 1st lecture

- The main interest in ‘statistics’ of physicists is inference, i.e. how to learn from data
- but the ‘prescriptions’ of ‘conventional statistics’ are not satisfactory
- They produce confusions in those who want to understand the sense of what they do

Summary of 1st lecture

- The main interest in ‘statistics’ of physicists is inference, i.e. how to learn from data
- but the ‘prescriptions’ of ‘conventional statistics’ are not satisfactory
- They produce confusions in those who want to understand the sense of what they do
- and, anyhow, they are responsible of severe errors in scientific judgments.

Summary of 1st lecture

- The main interest in ‘statistics’ of physicists is inference, i.e. how to learn from data
- but the ‘prescriptions’ of ‘conventional statistics’ are not satisfactory
- They produce confusions in those who want to understand the sense of what they do
- and, anyhow, they are responsible of severe errors in scientific judgments.
- Trying to start from the very beginning, and focusing on ‘hypotheses tests’, we have seen that

Summary of 1st lecture

- but the ‘prescriptions’ of ‘conventional statistics’ are not satisfactory
- They produce confusions in those who want to understand the sense of what they do
- and, anyhow, they are responsible of severe errors in scientific judgments.
- Trying to start from the very beginning, and focusing on ‘hypotheses tests’, we have seen that
- the very source of uncertainty is the uncertainty in the causal connections:

Summary of 1st lecture

- They produce confusions in those who want to understand the sense of what they do
- and, anyhow, they are responsible of severe errors in scientific judgments.
- Trying to start from the very beginning, and focusing on ‘hypotheses tests’, we have seen that
- the very source of uncertainty is the uncertainty in the causal connections:
- the standard statistical methods to treat the problem, can be seen as the practical attempt to implement the ideal of falsification;

Summary of 1st lecture

- and, anyhow, they are responsible of severe errors in scientific judgments.
- Trying to start from the very beginning, and focusing on ‘hypotheses tests’, we have seen that
- the very source of uncertainty is the uncertainty in the causal connections:
- the standard statistical methods to treat the problem, can be seen as the practical attempt to implement the ideal of falsification;
- falsificationism is a kind of extension of the ‘proof by contradiction’ to the natural science.

Summary of 1st lecture

- Trying to start from the very beginning, and focusing on ‘hypotheses tests’, we have seen that
- the very source of uncertainty is the uncertainty in the causal connections:
- the standard statistical methods to treat the problem, can be seen as the practical attempt to implement the ideal of falsification;
- falsificationism is a kind of extension of the ‘proof by contradiction’ to the natural science.
- But strict falsificationism is just naive,

Summary of 1st lecture

- the standard statistical methods to treat the problem, can be seen as the practical attempt to implement the ideal of falsification;
- falsificationism is a kind of extension of the ‘proof by contradiction’ to the natural science.
- But strict falsificationism is just naive,
- while its statistical implementations are logically flawed.

Summary of 1st lecture

- falsificationism is a kind of extension of the ‘proof by contradiction’ to the natural science.
- But strict falsificationism is just naive,
- while its statistical implementations are logically flawed.
- We ended with some examples from HEP that had quite some resonance some years ago, where fake claims of discoveries can be easily attributed to the general inability of physicists to handle the probability inversion problem, *“the essential problem of the the experimental method”* (Poincaré)

Summary of 1st lecture

- falsificationism is a kind of extension of the 'proof by contradiction' to the natural science.
- But strict falsificationism is just naive,
- while its statistical implementations are logically flawed.
- We ended with some examples from HEP that had quite some resonance some years ago, where fake claims of discoveries can be easily attributed to the general inability of physicists to handle the probability inversion problem, *"the essential problem of the the experimental method"* (Poincaré)
- There is only one way to calculate 'inverse probabilities':

$\longrightarrow$ Use probability theory. $\longrightarrow$ Bayes' theorem

Summary of 1st lecture

- falsificationism is a kind of extension of the 'proof by contradiction' to the natural science.
- But strict falsificationism is just naive,
- while its statistical implementations are logically flawed.
- We ended with some examples from HEP that had quite some resonance some years ago, where fake claims of discoveries can be easily attributed to the general inability of physicists to handle the probability inversion problem, *"the essential problem of the the experimental method"* (Poincaré)
- There is only one way to calculate 'inverse probabilities':

  $\longrightarrow$ Use probability theory. $\longrightarrow$ Bayes' theorem

- But we have first to recover the intuitive idea of probability, rather then XX-th century artefacts.

Where to restart

Starting point for probabilistic reasoning

- Probability means how much we believe something
- Probability values obey the following basic rules

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B) \quad [\,\text{if } P(A \cap B) = \emptyset\,]$
4. $P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(B \,|\, A) \cdot P(A)$

Where to restart

Starting point for probabilistic reasoning

- Probability means how much we believe something
- Probability values obey the following basic rules

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B) \quad$ [if $P(A \cap B) = \emptyset$]
4. $P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(B \,|\, A) \cdot P(A)\,,$

That includes 'direct probability problems' (propagation of uncertainties) and also probabilistic inference (or 'inverse probability'), based on the symmetric reconditioning formula, that, though under several variations, goes under the name of Bayes theorem.

The Bayes 'formulae'

Main link between conditional probabilities of effects and conditional probabilities of hypotheses.

$$\boxed{P(C_j, E_i) = P(E_i \,|\, C_j)\, P(C_j) = P(C_j \,|\, E_i)\, P(E_i)}$$

From which different ways to write Bayes theorem follow:

$$\begin{aligned}
\frac{P(H_j \,|\, E_i)}{P(H_j)} &= \frac{P(E_i \,|\, H_j)}{P(E_i)} \\
P(H_j \,|\, E_i) &= \frac{P(E_i \,|\, H_j)}{P(E_i)}\, P(H_j) \\
P(H_j \,|\, E_i) &= \frac{P(E_i \,|\, H_j) \cdot P(H_j)}{\sum_j P(E_i \,|\, H_j) \cdot P(H_j)} \\
P(H_j \,|\, E_i) &\propto P(E_i \,|\, H_j) \cdot P(H_j) \qquad * * *
\end{aligned}$$

The Bayes 'formulae'

Main link between conditional probabilities of effects and conditional probabilities of hypotheses.

$$\boxed{P(C_j, E_i) = P(E_i \,|\, C_j)\, P(C_j) = P(C_j \,|\, E_i)\, P(E_i)}$$

From which different ways to write Bayes theorem follow:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

$$P(H_j \,|\, E_i) = \frac{P(E_i \,|\, H_j)}{P(E_i)}\, P(H_j)$$

$$P(H_j \,|\, E_i) = \frac{P(E_i \,|\, H_j) \cdot P(H_j)}{\sum_j P(E_i \,|\, H_j) \cdot P(H_j)}$$

$$P(H_j \,|\, E_i) \propto P(E_i \,|\, H_j) \cdot P(H_j) \qquad * * *$$

$$\frac{P(H_j \,|\, E_i)}{P(H_k \,|\, E_i)} = \frac{P(E_i \,|\, H_j)}{P(E_i \,|\, H_k)} \cdot \frac{P(H_j)}{P(H_k)} \qquad * * *$$

And their sequential use

$\boxed{\text{The posterior becomes the prior of the next inference}}$

For conditionally independent E_i:

$$P(H_j \,|\, E^{(1)}, E^{(2)}) \;\;\propto\;\; P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j)$$

And their sequential use

$$\boxed{\text{The posterior becomes the prior of the next inference}}$$

For conditionally independent E_i:

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
P(H_j \,|\, \text{data}) &\propto P(\text{data} \,|\, H_j) \cdot P_0(H_j) \\
P(H_j \,|\, \text{data}) &\propto P(\text{data}_1 \,|\, H_j) \cdot P(\text{data}_2 \,|\, H_j) \cdot \;\ldots\; \cdot P_0(H_j)
\end{aligned}
$$

And their sequential use

$\boxed{\text{The posterior becomes the prior of the next inference}}$

For conditionally independent E_i:

$$P(H_j \,|\, E^{(1)}, E^{(2)}) \;\propto\; P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j)$$

$$P(H_j \,|\, \text{data}) \;\propto\; P(\text{data} \,|\, H_j) \cdot P_0(H_j)$$

$$P(H_j \,|\, \text{data}) \;\propto\; P(\text{data}_1 \,|\, H_j) \cdot P(\text{data}_2 \,|\, H_j) \cdot \;\ldots\; \cdot P_0(H_j)$$

Similarly, for the Bayes theorem written in terms of odd ratios:

$$\frac{P(H_j \,|\, \text{data})}{P(H_k \,|\, \text{data})} = \frac{P(\text{data}_1 \,|\, H_j)}{P(\text{data}_1 \,|\, H_k)} \cdot \frac{P(\text{data}_2 \,|\, H_j)}{P(\text{data}_2 \,|\, H_k)} \cdot \;\ldots\; \cdot \frac{P(H_j)}{P(H_k)}$$

And their sequential use

The posterior becomes the prior of the next inference

For conditionally independent E_i:

$$P(H_j\,|\,E^{(1)},E^{(2)}) \;\propto\; P(E^{(2)}\,|\,H_j)\cdot P(E^{(1)}\,|\,H_j)\cdot P_0(H_j)$$

$$P(H_j\,|\,\text{data}) \;\propto\; P(\text{data}\,|\,H_j)\cdot P_0(H_j)$$

$$P(H_j\,|\,\text{data}) \;\propto\; P(\text{data}_1\,|\,H_j)\cdot P(\text{data}_2\,|\,H_j)\cdot\;\ldots\;\cdot P_0(H_j)$$

Similarly, for the Bayes theorem written in terms of odd ratios:

$$\frac{P(H_j\,|\,\text{data})}{P(H_k\,|\,\text{data})} = \frac{P(\text{data}_1\,|\,H_j)}{P(\text{data}_1\,|\,H_k)}\cdot\frac{P(\text{data}_2\,|\,H_j)}{P(\text{data}_2\,|\,H_k)}\cdot\;\ldots\;\cdot\frac{P(H_j)}{P(H_k)}$$

(And, obviously, if the data sets are not independent, one has to apply the chain rule $P(A,\,B\,C,\ldots) = P(A)\cdot P(B\,|\,A)\cdot P(C\,|\,A,\,B)\ldots$*)*

Exercise: particle identification

A particle detector has a μ identification efficiency of $95\,\%$, and a probability of identifying a π as a μ of $2\,\%$. If a particle is identified as a μ, then a trigger is fired.
The particle beam is a mixture of $90\,\%$ π and $10\,\%$ μ,

Exercise: particle identification

A particle detector has a μ identification efficiency of $95\,\%$, and a probability of identifying a π as a μ of $2\,\%$. If a particle is identified as a μ, then a trigger is fired.
The particle beam is a mixture of $90\,\%$ π and $10\,\%$ μ,

- What is the probability that a trigger is really fired by a μ?
- What is the signal-to-noise (S/N) ratio?

Exercise: particle identification

A particle detector has a μ identification efficiency of $95\,\%$, and a probability of identifying a π as a μ of $2\,\%$. If a particle is identified as a μ, then a trigger is fired.
The particle beam is a mixture of $90\,\%$ π and $10\,\%$ μ,

- What is the probability that a trigger is really fired by a μ?
- What is the signal-to-noise (S/N) ratio?

$$\begin{aligned} P(\mu\,|\,T) &= \frac{P(T\,|\,\mu)\,P_\circ(\mu)}{P(T\,|\,\mu)\,P_\circ(\mu)+P(T\,|\,\pi)\,P_\circ(\pi)} \\ &= \frac{0.95\times 0.1}{0.95\times 0.1+0.02\times 0.9}=0.84\,, \end{aligned}$$

Exercise: particle identification

A particle detector has a μ identification efficiency of $95\,\%$, and a probability of identifying a π as a μ of $2\,\%$. If a particle is identified as a μ, then a trigger is fired.
The particle beam is a mixture of $90\,\%$ π and $10\,\%$ μ,

- What is the probability that a trigger is really fired by a μ?
- What is the signal-to-noise (S/N) ratio?

$$P(\mu\,|\,T) = \frac{P(T\,|\,\mu)\,P_\circ(\mu)}{P(T\,|\,\mu)\,P_\circ(\mu) + P(T\,|\,\pi)\,P_\circ(\pi)}$$
$$= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.02 \times 0.9} = 0.84\,,$$

Signal-to-noise ratio

$$\frac{P(\mu\,|\,T)}{P(\pi\,|\,T)} = 5.3$$

Signal-to-noise ratio

Rewrite S/N as Bayes factor times prior odds:

$$
\begin{aligned}
S/N &= \frac{P(S\,|\,E)}{P(N\,|\,E)} = \frac{P(E\,|\,S)}{P(E\,|\,N)} \cdot \frac{P_\circ(S)}{P_\circ(N)} \\
\frac{P(\mu\,|\,T)}{P(\pi\,|\,T)} &= 47.5 \times 0.111 = 5.3
\end{aligned}
$$

Signal-to-noise ratio

Rewrite S/N as Bayes factor times prior odds:

$$\begin{aligned} S/N &= \frac{P(S\,|\,E)}{P(N\,|\,E)} = \frac{P(E\,|\,S)}{P(E\,|\,N)} \cdot \frac{P_\circ(S)}{P_\circ(N)} \\ \frac{P(\mu\,|\,T)}{P(\pi\,|\,T)} &= 47.5 \times 0.111 = 5.3 \end{aligned}$$

This formula explicitly shows that when there are noisy conditions,

$$P_\circ(S) \ll P_\circ(N)\,,$$

the experiment must be very selective,

$$P(E\,|\,S) \gg P(E\,|\,N)\,,$$

in order to have a decent S/N ratio ($\longrightarrow$ AIDS problem).

Signal-to-noise ratio

Rewrite S/N as Bayes factor times prior odds:

$$\begin{aligned} S/N &= \frac{P(S\,|\,E)}{P(N\,|\,E)} = \frac{P(E\,|\,S)}{P(E\,|\,N)} \cdot \frac{P_\circ(S)}{P_\circ(N)} \\ \frac{P(\mu\,|\,T)}{P(\pi\,|\,T)} &= 47.5 \times 0.111 = 5.3 \end{aligned}$$

This formula explicitly shows that when there are noisy conditions,

$$P_\circ(S) \ll P_\circ(N)\,,$$

the experiment must be very selective,

$$P(E\,|\,S) \gg P(E\,|\,N)\,,$$

in order to have a decent S/N ratio ($\longrightarrow$ AIDS problem).
(Follow up: new S/N if two indep. detectors are used.)

Why do frequentistic tests often work?

$\longrightarrow$ See slides:

Uncertainties in measurements

Having to perform a measurement:

Uncertainties in measurements

Having to perform a measurement:

Which numbers shall come out from our device?

Having performed a measurement:

What have we learned about the value of the quantity of interest?

How to quantify these kinds of uncertainty?

Uncertainties in measurements

Having to perform a measurement:

Which numbers shall come out from our device?

Having performed a measurement:

What have we learned about the value of the quantity of interest?

How to quantify these kinds of uncertainty?

Under well controlled conditions (calibration) we can make use of past frequencies to evaluate 'somehow' the detector response $f(x \mid \mu)$.

Uncertainties in measurements

Having to perform a measurement:

Which numbers shall come out from our device?

Having performed a measurement:

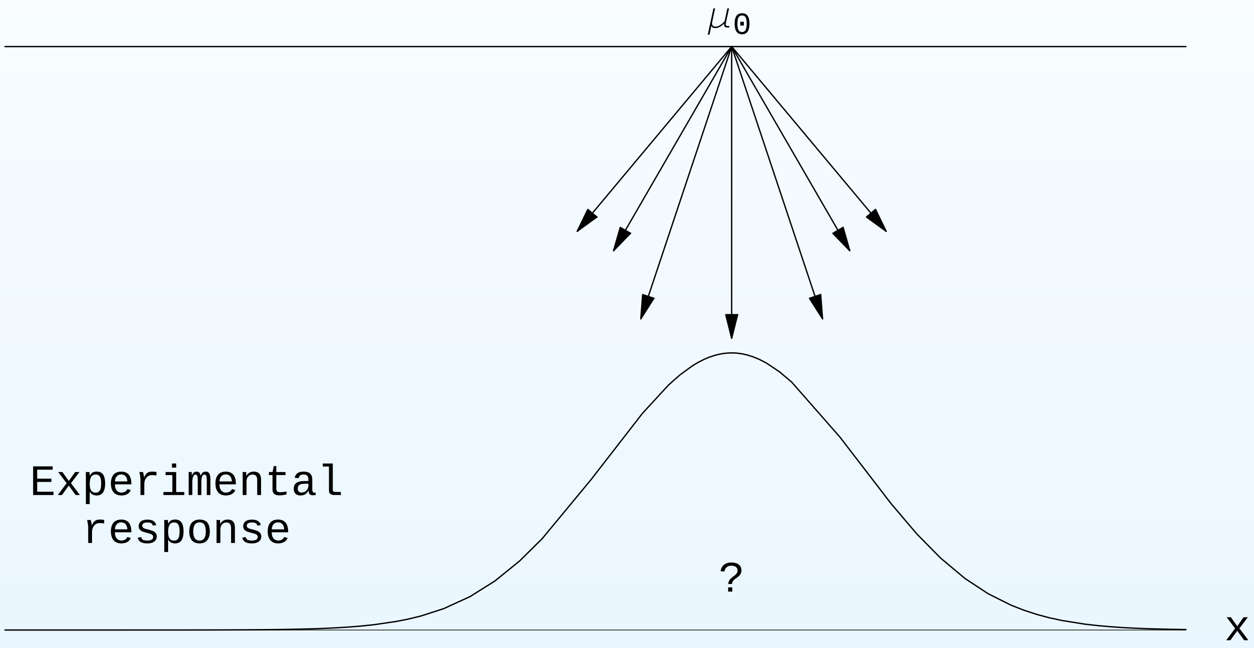
What have we learned about the value of the quantity of interest?

How to quantify these kinds of uncertainty?

Under well controlled conditions (calibration) we can make use of past frequencies to evaluate 'somehow' the detector response $f(x \mid \mu)$.

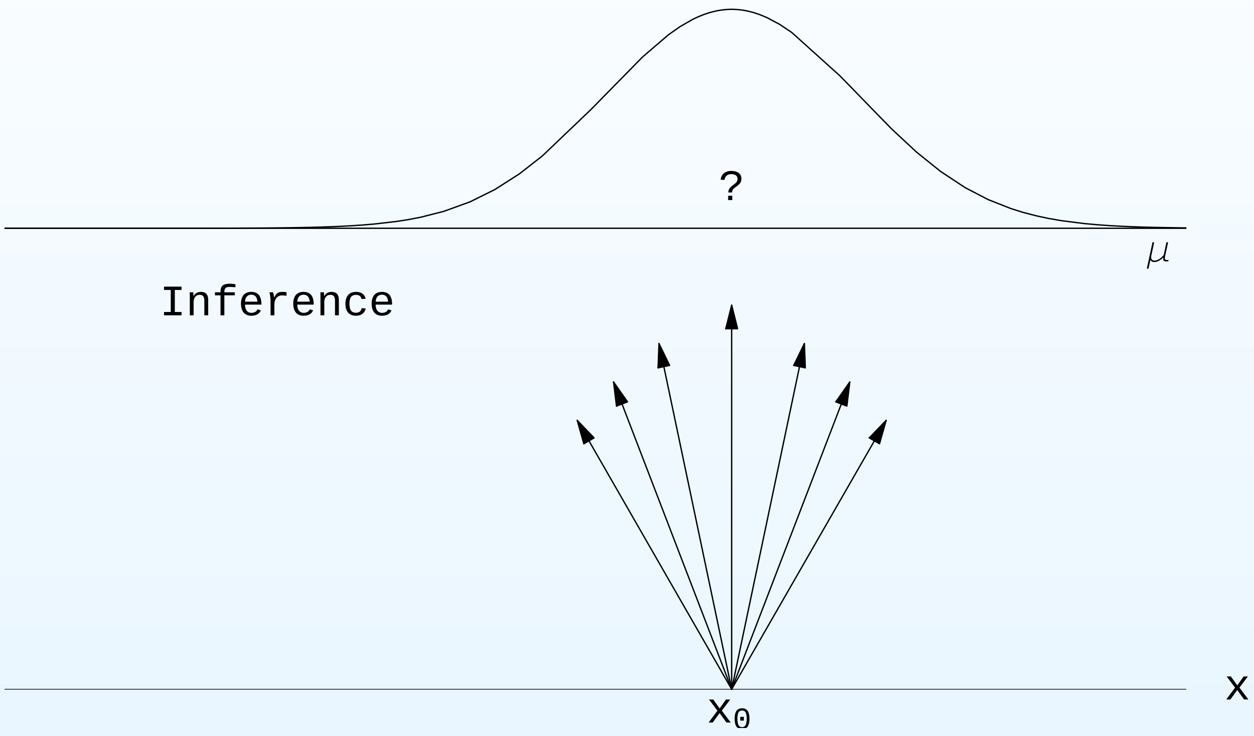
There is (in most cases) no way to get *directly* hints about $f(\mu \mid x)$.

Uncertainties in measurements



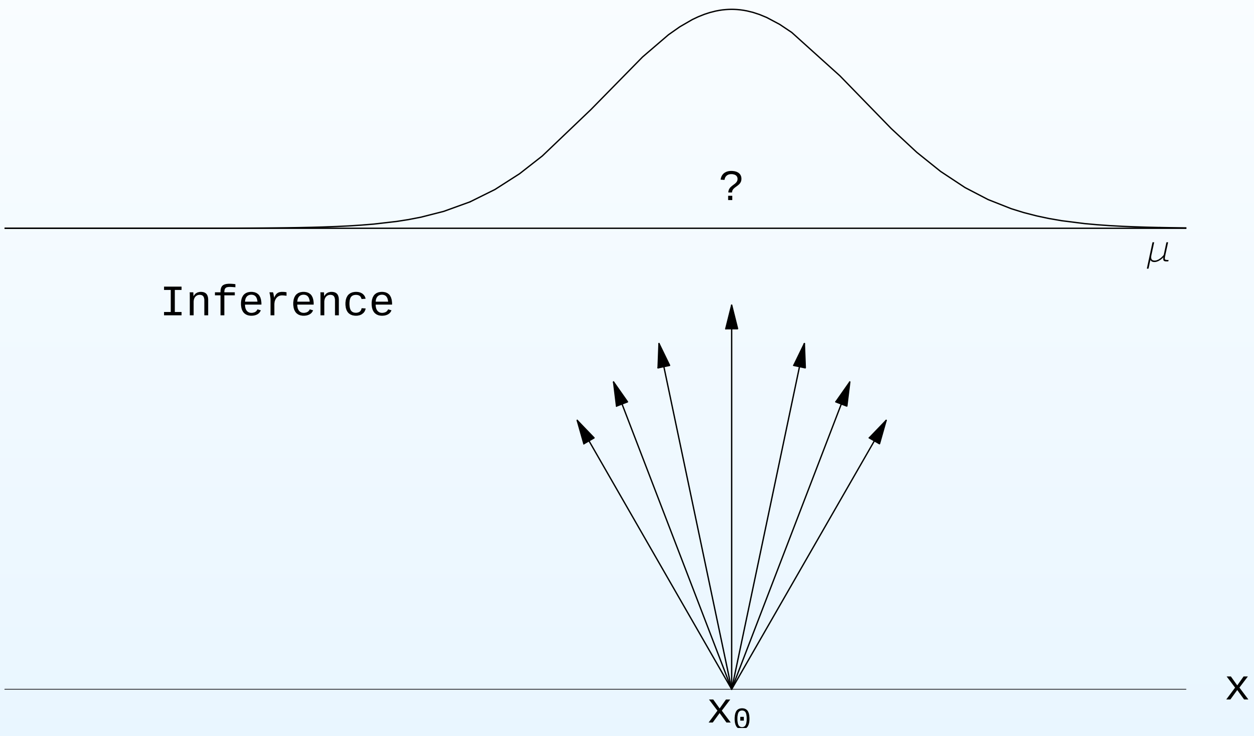
$f(x \mid \mu)$ experimentally accessible (though 'model filtered')

Uncertainties in measurements



$f(\mu \,|\, x)$ experimentally inaccessible

Uncertainties in measurements

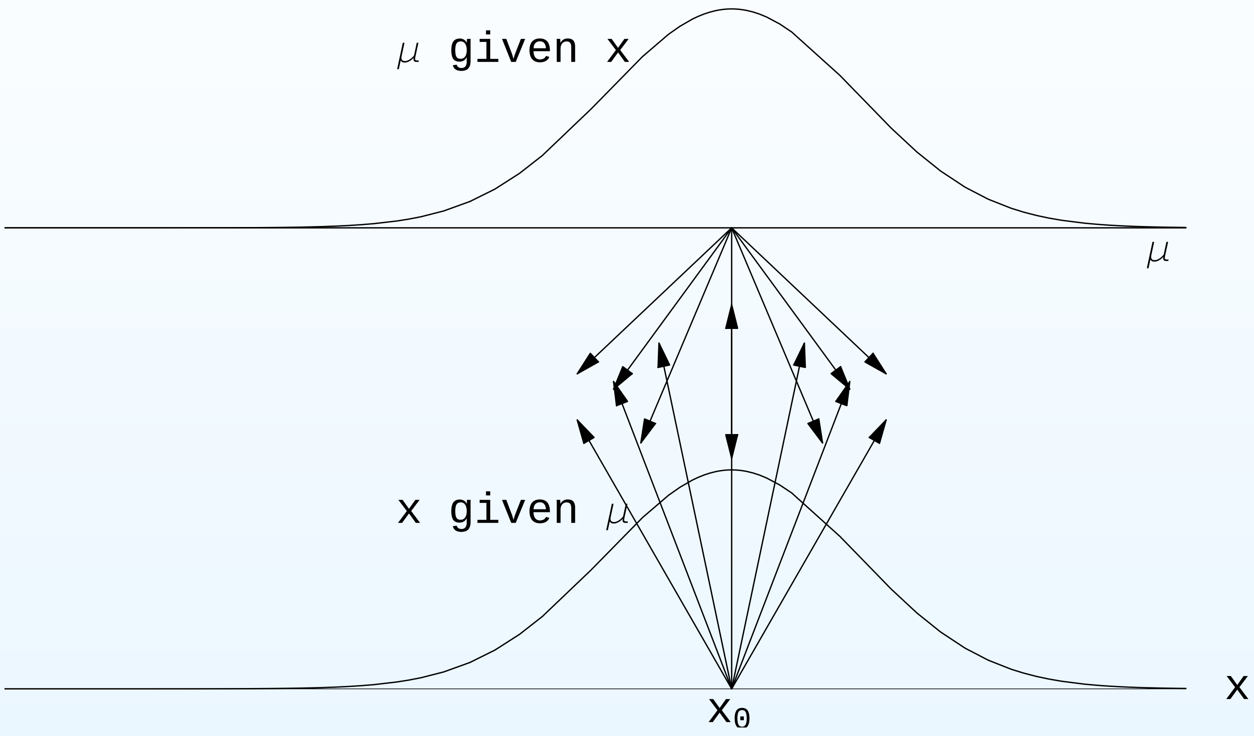


$f(\mu\,|\,x)$ experimentally inaccessible

but logically accessible!

$\rightarrow$ probability inversion $\rightarrow$ Bayes

Uncertainties in measurements



- How measurement uncertainties are currently treated?
- How to treat them logically using probability theory?

Usual way to deal with measurement uncertainties

Uncertainties due to statistical errors are currently treated using the frequentistic concept of ‘confidence interval’,

Usual way to deal with measurement uncertainties

Uncertainties due to statistical errors are currently treated using the frequentistic concept of ‘confidence interval’, although

- there are well-know cases — of great relevance in frontier physics — in which the approach is not applicable (e.g. small number of observed events, or measurement close to the edge of the physical region);

Usual way to deal with measurement uncertainties

Uncertainties due to statistical errors are currently treated using the frequentistic concept of ‘confidence interval’, although

- there are well-know cases — of great relevance in frontier physics — in which the approach is not applicable (e.g. small number of observed events, or measurement close to the edge of the physical region);
- the procedure is rather unnatural, and in fact the interpretation of the results is unconsciously (intuitively) probabilistic (see later).
  $\rightarrow$ Intuitive reasoning $\Longleftrightarrow$ statistics education

Usual way to deal with measurement uncertainties

Uncertainties due to statistical errors are currently treated using the frequentistic concept of ‘confidence interval’, although

- there are well-know cases — of great relevance in frontier physics — in which the approach is not applicable (e.g. small number of observed events, or measurement close to the edge of the physical region);
- the procedure is rather unnatural, and in fact the interpretation of the results is unconsciously (intuitively) probabilistic (see later).
  $\rightarrow$ Intuitive reasoning $\Longleftrightarrow$ statistics education

Usual way to deal with measurement uncertainties

There is no satisfactory theory or model to treat uncertainties due to systematic errors:

- *“my supervisor says . . . ”*
- *“add them linearly”;*
- *“add them linearly if . . . , else add them quadratically”*;
- *“don’t add them at all”.*

Usual way to deal with measurement uncertainties

There is no satisfactory theory or model to treat uncertainties due to systematic errors:

- *"my supervisor says . . . "*
- *"add them linearly";*
- *"add them linearly if . . . , else add them quadratically"*;
- *"don't add them at all".*

The modern *fashion*: add them quadratically if they are considered to be independent, or build a covariance matrix of statistical and systematic contributions in the general case.

Usual way to deal with measurement uncertainties

There is no satisfactory theory or model to treat uncertainties due to systematic errors:

- *“my supervisor says . . . ”*
- *“add them linearly”;*
- *“add them linearly if . . . , else add them quadratically”*;
- *“don’t add them at all”.*

The modern *fashion*: add them quadratically if they are considered to be independent, or build a covariance matrix of statistical and systematic contributions in the general case.
In my opinion, simply due to reluctance to combine linearly 10, 20 or more contributions to a global uncertainty, as the (out of fashion) ‘theory’ of maximum bounds would require.

Usual way to deal with measurement uncertainties

There is no satisfactory theory or model to treat uncertainties due to systematic errors:

- *“my supervisor says . . . ”*
- *“add them linearly”;*
- *“add them linearly if . . . , else add them quadratically”*;
- *“don’t add them at all”.*

The modern *fashion*: add them quadratically if they are considered to be independent, or build a covariance matrix of statistical and systematic contributions in the general case.
In my opinion, simply due to reluctance to combine linearly 10, 20 or more contributions to a global uncertainty, as the (out of fashion) ‘theory’ of maximum bounds would require.
$\rightarrow$ Right in most cases!
$\rightarrow$ Good sense of physicists $\Longleftrightarrow$ cultural background

A simple case

n independent measurements of the same quantity μ (with n large enough and no systematic effects, to avoid, for the moment, extra complications).

Evaluate $\overline{x}$ and σ from the data

report result: $\rightarrow \mu = \overline{x} \pm \sigma/\sqrt{n}$

- what does it mean?

A simple case

n independent measurements of the same quantity μ (with n large enough and no systematic effects, to avoid, for the moment, extra complications).

Evaluate $\overline{x}$ and σ from the data

report result: $\rightarrow \mu = \overline{x} \pm \sigma/\sqrt{n}$

- what does it mean?

1. For the large majority of physicists $P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$

A simple case

n independent measurements of the same quantity μ (with n large enough and no systematic effects, to avoid, for the moment, extra complications).

Evaluate $\overline{x}$ and σ from the data

report result: $\rightarrow \mu = \overline{x} \pm \sigma/\sqrt{n}$

- what does it mean?

1 For the large majority of physicists $P(\overline{x} - \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$

2 And many explain (also to students!) that *"this means that, if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval* $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$*."*

A simple case

n independent measurements of the same quantity μ (with n large enough and no systematic effects, to avoid, for the moment, extra complications).

Evaluate $\overline{x}$ and σ from the data

report result: $\rightarrow \mu = \overline{x} \pm \sigma/\sqrt{n}$

- what does it mean?

1. For the large majority of physicists
   $P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$
2. And many explain (also to students!) that *"this means that, if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$."*
3. Statistics experts tell that the interval $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$ covers the true μ in 68% of cases

A simple case

n independent measurements of the same quantity μ (with n large enough and no systematic effects, to avoid, for the moment, extra complications).

Evaluate $\overline{x}$ and σ from the data

report result: $\rightarrow \mu = \overline{x} \pm \sigma/\sqrt{n}$

- what does it mean? **Objections?**

1 For the large majority of physicists
$P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$

2 And many explain (also to students!) that *"this means that, if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval* $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$*."*

3 Statistics experts tell that the interval $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$ covers the true μ in 68% of cases

Meaning of $\mu = \overline{x} \pm \sigma/\sqrt{n}$

1 $P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$

OK to me, and perhaps no objections by many of you

- But it depends on what we mean by probability
- If probability is the "limit of the frequency", this statement is meaningless, because the 'frequency based' probability theory only speak about

$$P(\mu - \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\%\,,$$

(that is a probabilistic statement about $\overline{X}$: probabilistic statements about μ are not allowed by the theory).

Meaning of $\mu = \overline{x} \pm \sigma/\sqrt{n}$

2 *"if I repeat the experiment a great number of times, then I will find that in roughly 68% of the cases the observed average will be in the interval* $[\overline{x} - \sigma/\sqrt{n},\ \overline{x} + \sigma/\sqrt{n}]$*."*

- Nothing wrong in principle (in my opinion)
- but a $\sqrt{2}$ mistake in the width of the interval

$\rightarrow$ $P(\overline{x} - \sigma/\sqrt{n} \ \leq\ \overline{x}_f \ \leq\ \overline{x} + \sigma/\sqrt{n}) = 52\%$,
where $\overline{x}_f$ stands for future averages;

or $P(\overline{x} - \sqrt{2}\,\sigma/\sqrt{n} \ \leq\ \overline{x}_f \ \leq\ \overline{x} + \sqrt{2}\,\sigma/\sqrt{n}) = 68\%$,

as we shall see later ($\rightarrow$ 'predictive distributions').

Meaning of $\mu = \overline{x} \pm \sigma/\sqrt{n}$

3 Frequentistic coverage $\longrightarrow$ “several problems”

Meaning of $\mu = \overline{x} \pm \sigma/\sqrt{n}$

3 Frequentistic coverage $\longrightarrow$ “several problems”

- ‘Trivial’ interpretation problem: $\longrightarrow$ taken by most users as if it were a probability interval (not just semantic!)

Meaning of $\mu = \overline{x} \pm \sigma/\sqrt{n}$

3 Frequentistic coverage $\longrightarrow$ “several problems”

- ‘Trivial’ interpretation problem: $\longrightarrow$ taken by most users as if it were a probability interval (not just semantic!)
- It fails in frontier cases
  - ’technically’ [see e.g. G. Zech, *Frequentistic and Bayesian confidence limits*, EPJdirect C12 (2002) 1]
  - ‘in terms of performance’ $\longrightarrow$ ‘very strange’ that no quantities show in ‘other side’ of a 95% C.L. bound !

Arbitrary probability inversions

As with hypotheses tests, problem arises from arbitrary probability inversions.

How do we turn, just 'intuitively'

$$P(\mu - \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\%$$

into

$$P(\overline{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%\,?$$

Arbitrary probability inversions

As with hypotheses tests, problem arises from arbitrary probability inversions.

How do we turn, just 'intuitively'

$$P(\mu - \frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + \frac{\sigma}{\sqrt{n}}) = 68\%$$

into

$$P(\overline{x} - \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + \frac{\sigma}{\sqrt{n}}) = 68\%\,?$$

We can paraphrase as

"the dog and the hunter"

The dog and the hunter

We know that a dog has a 50% probability of being 100 m from the hunter

$\Rightarrow$ if we observe the dog, what can we say about the hunter?

The dog and the hunter

We know that a dog has a 50% probability of being 100 m from the hunter

$\Rightarrow$ if we observe the dog, what can we say about the hunter?

The terms of the analogy are clear:

$$\begin{aligned} \text{hunter} \quad &\leftrightarrow \quad \text{true value} \\ \text{dog} \quad &\leftrightarrow \quad \text{observable}\,. \end{aligned}$$

The dog and the hunter

We know that a dog has a 50% probability of being 100 m from the hunter

$\Rightarrow$ if we observe the dog, what can we say about the hunter?

The terms of the analogy are clear:

$$\begin{aligned} \text{hunter} \quad &\leftrightarrow \quad \text{true value} \\ \text{dog} \quad &\leftrightarrow \quad \text{observable}\,. \end{aligned}$$

Intuitive and reasonable answer:

> *"The hunter is, with 50% probability, within 100 m of the position of the dog."*

The dog and the hunter

We know that a dog has a 50% probability of being 100 m from the hunter

$\Rightarrow$ if we observe the dog, what can we say about the hunter?

The terms of the analogy are clear:

$$\begin{aligned} \text{hunter} \quad &\leftrightarrow \quad \text{true value} \\ \text{dog} \quad &\leftrightarrow \quad \text{observable}\,. \end{aligned}$$

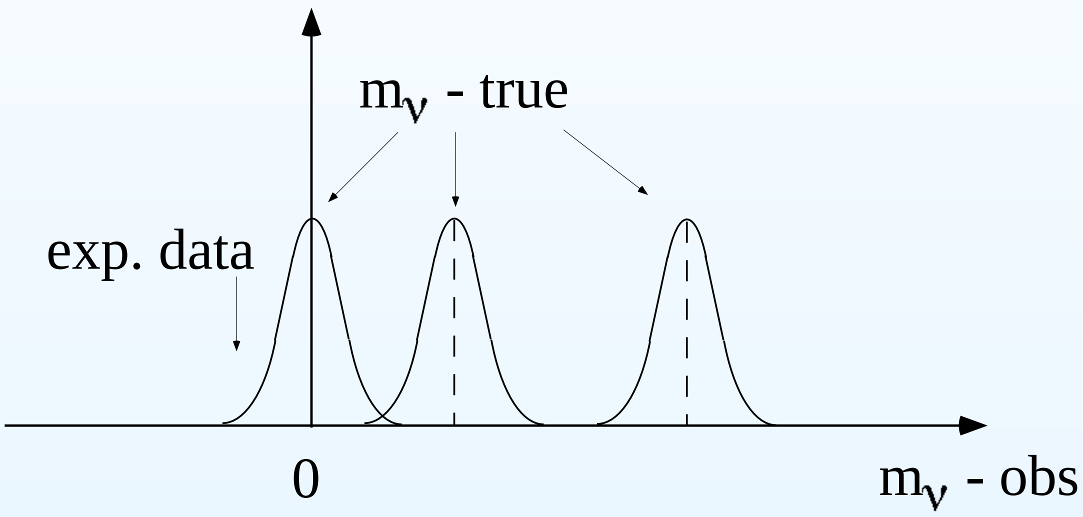
Easy to understand that this conclusion is based on some tacit assumptions:

- the hunter can be anywhere around the dog
- the dog has no preferred direction of arrival at the point where we observe him.

$\longrightarrow$ not always valid!

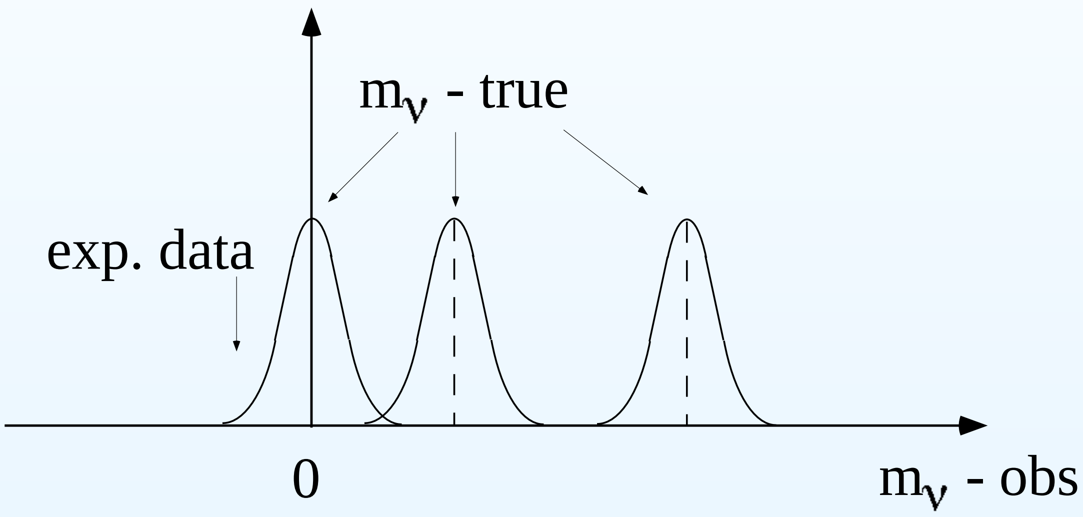
Measurement at the edge of a physical region

Electron-neutrino experiment, mass resolution $\sigma = 2\,$eV, independent of m_ν.



Measurement at the edge of a physical region

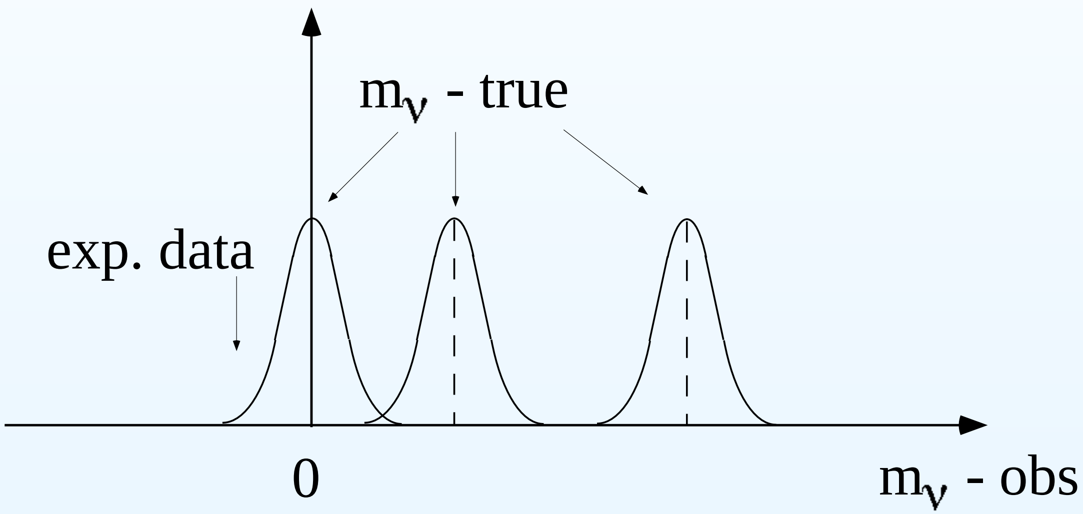
Electron-neutrino experiment, mass resolution $\sigma = 2$ eV, independent of m_ν.



Observation: -4 eV.
What can we tell about m_ν?

Measurement at the edge of a physical region

Electron-neutrino experiment, mass resolution $\sigma = 2$ eV, independent of m_ν.

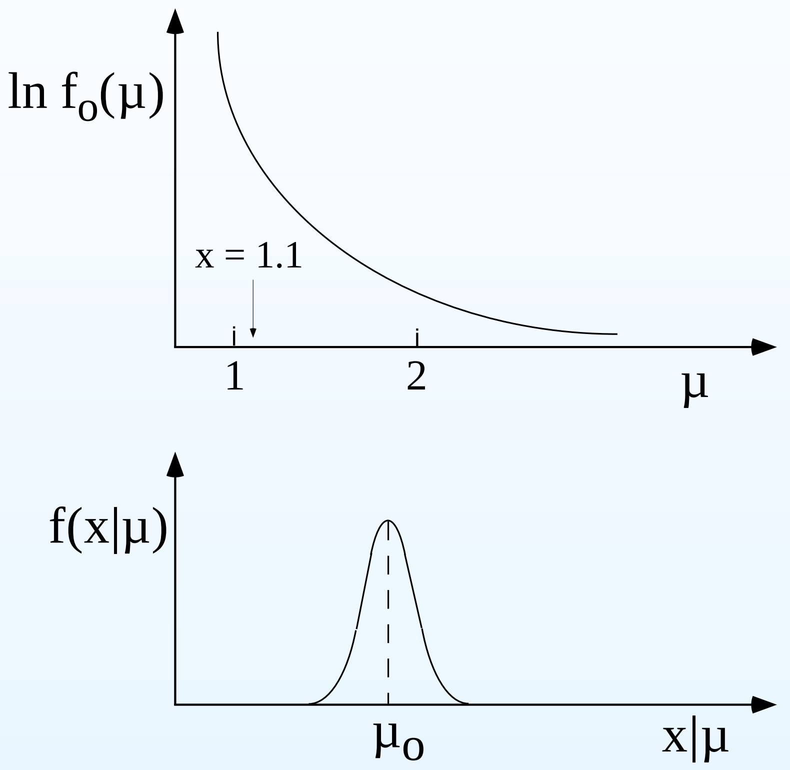


Observation: -4 eV.
What can we tell about m_ν?

$m_\nu = -4 \pm 2$ eV ?
$P(-6 \leq m_\nu/\text{eV} \leq -2) = 68\%$?
$P(m_\nu \leq 0 \text{ eV}) = 98\%$?

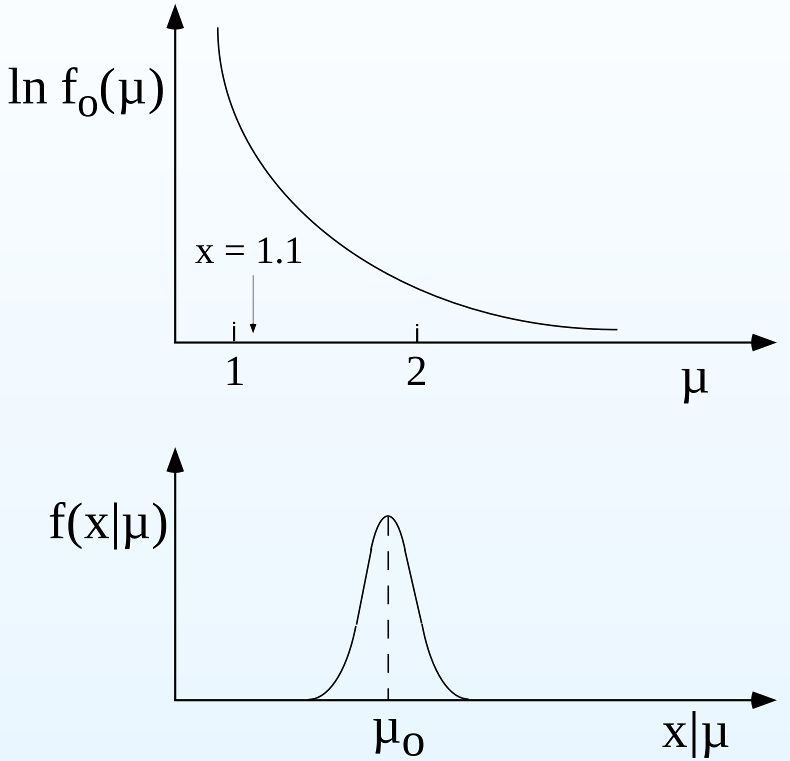
Non-flat distribution of a physical quantity

Imagine a cosmic ray particle or a bremsstrahlung γ.



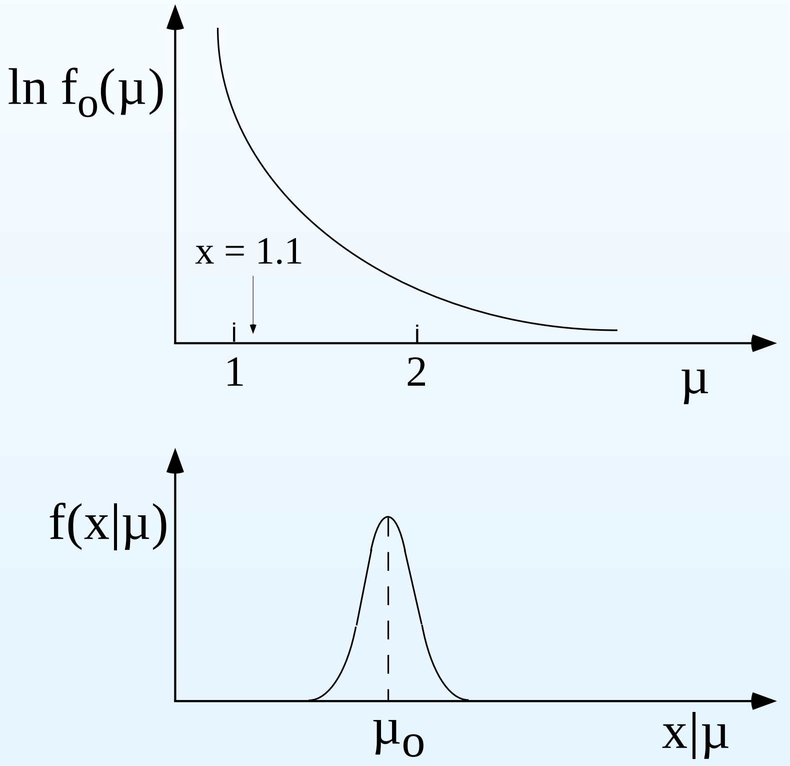
Non-flat distribution of a physical quantity

Imagine a cosmic ray particle or a bremsstrahlung γ.
Observed $x = 1.1$.
What can we say about the true value μ that has caused this observation?



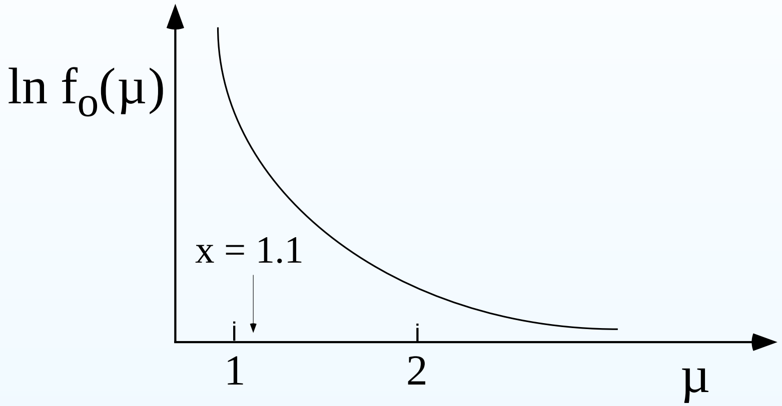
Non-flat distribution of a physical quantity

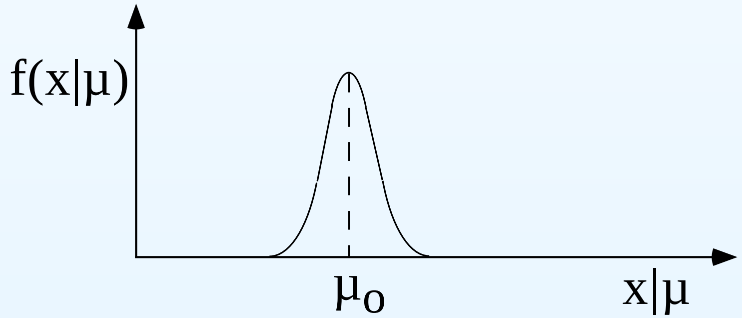
Imagine a cosmic ray particle or a bremsstrahlung γ.
Observed $x = 1.1$.
What can we say about the true value μ that has caused this observation?
Also in this case the formal definition of the confidence interval does not work.
Intuitively, we feel that there is more chance that μ is on the left of 1.1 than on the right. In the jargon of the experimentalists, *"there are more migrations from left to right than from right to left".*



Non-flat distribution of a physical quantity

These two examples deviates from the dog-hunter picture only because of an asymmetric possible position of the 'hunter', i.e our expectation about μ is not uniform. But there are also interesting cases in which the response of the apparatus $f(x \mid \mu)$ is not symmetric around μ, e.g. the reconstructed momentum in a magnetic spectrometer.



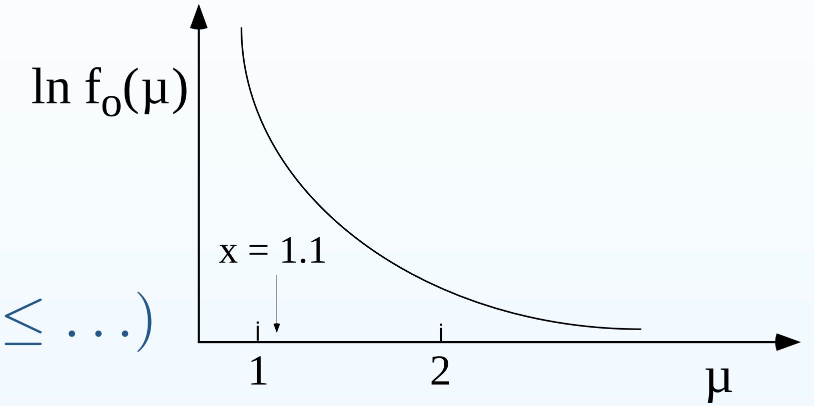


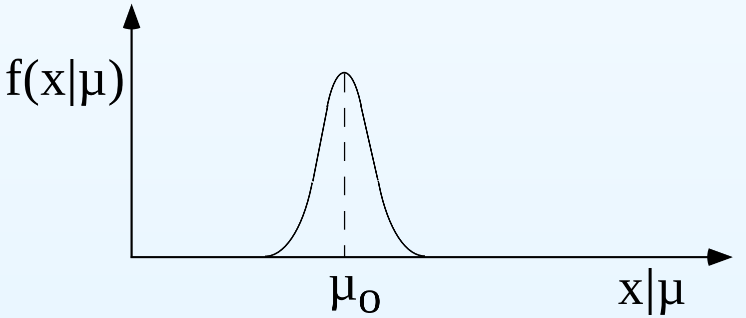
Non-flat distribution of a physical quantity

Summing up:
“the intuitive inversion of probability

$$P(\ldots \leq \overline{X} \leq \ldots) \Longrightarrow P(\ldots \leq \mu \leq \ldots)$$

besides being theoretically unjustifiable, yields results which are numerically correct only in the case of symmetric problems.”





Summary about standard methods

Situation is not satisfactory in the critical situations that often occur in HEP, both in

- hypotheses tests
- confidence intervals

Summary about standard methods

Situation is not satisfactory in the critical situations that often occur in HEP, both in

- hypotheses tests
- confidence intervals

Plus there are issues not easy to treat in that frame
[and I smile at the heroic effort to get some result :-)]

- systematic errors
- background

Parametric inference

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

$$\boxed{f(\theta \,|\, \text{data}) \propto f(\text{data} \,|\, \theta)\, f_0(\theta)}$$

Parametric inference

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

$$\boxed{f(\theta\,|\,\text{data}) \propto f(\text{data}\,|\,\theta)\, f_0(\theta)}$$

Remark: the probabilistic result is the pdf $f(\theta\,|\,\text{data})$,

Parametric inference

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

$$\boxed{f(\theta\,|\,\text{data}) \propto f(\text{data}\,|\,\theta)\, f_0(\theta)}$$

Remark: the probabilistic result is the pdf $f(\theta\,|\,\text{data})$,

- $f(\theta\,|\,\text{data})$ can eventually be summarised with average ('expected value'), standard deviation ('standard uncertainty'), value of highest probability (mode), probability intervals, etc. (any statement consistent with $f(\theta\,|\,\text{data})$ is virtually valid).

Parametric inference

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

$$\boxed{f(\theta\,|\,\text{data}) \propto f(\text{data}\,|\,\theta)\, f_0(\theta)}$$

Remark: the probabilistic result is the pdf $f(\theta\,|\,\text{data})$,

- $f(\theta\,|\,\text{data})$ can eventually be summarised with average ('expected value'), standard deviation ('standard uncertainty'), value of highest probability (mode), probability intervals, etc. (any statement consistent with $f(\theta\,|\,\text{data})$ is virtually valid).
- $E[\theta]$ and $\sigma(\theta)$ are particulary convenient for further propagations, thanks to general theorem that apply to them, but not to mode, median or intervals!

Parametric inference

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

$$\boxed{f(\theta\,|\,\text{data}) \propto f(\text{data}\,|\,\theta)\, f_0(\theta)}$$

Remark: the probabilistic result is the pdf $f(\theta\,|\,\text{data})$,

- $f(\theta\,|\,\text{data})$ can eventually be summarised with average ('expected value'), standard deviation ('standard uncertainty'), value of highest probability (mode), probability intervals, etc. (any statement consistent with $f(\theta\,|\,\text{data})$ is virtually valid).
- $E[\theta]$ and $\sigma(\theta)$ are particulary convenient for further propagations, thanks to general theorem that apply to them, but not to mode, median or intervals!
- but the full answer is $f(\theta\,|\,\text{data})$!

Inferring the Binomial p

$\longrightarrow$ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

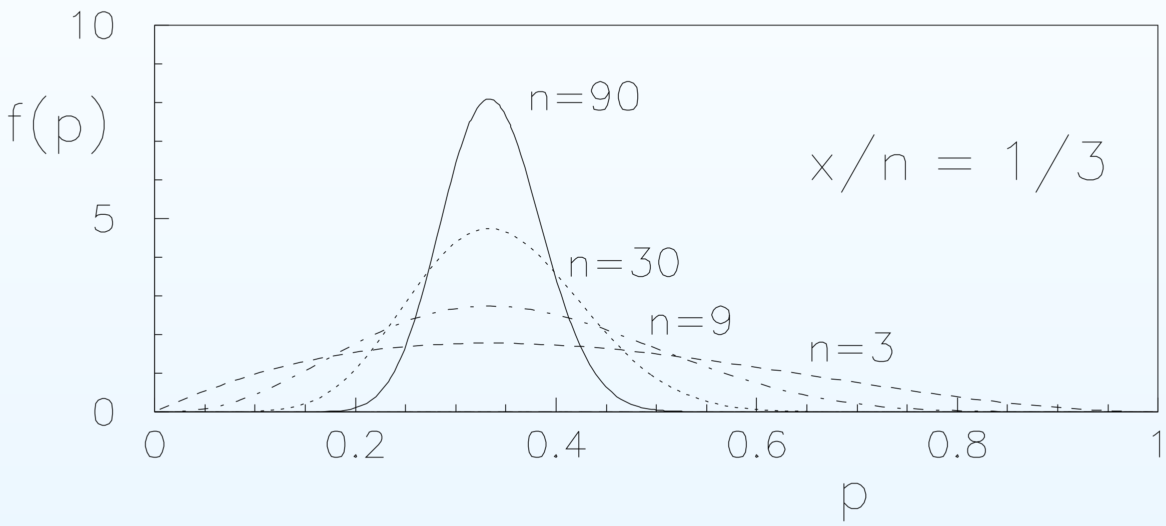
$$\boxed{f(\theta \,|\, \text{data}) \propto f(\text{data} \,|\, \theta)\, f_0(\theta)}$$

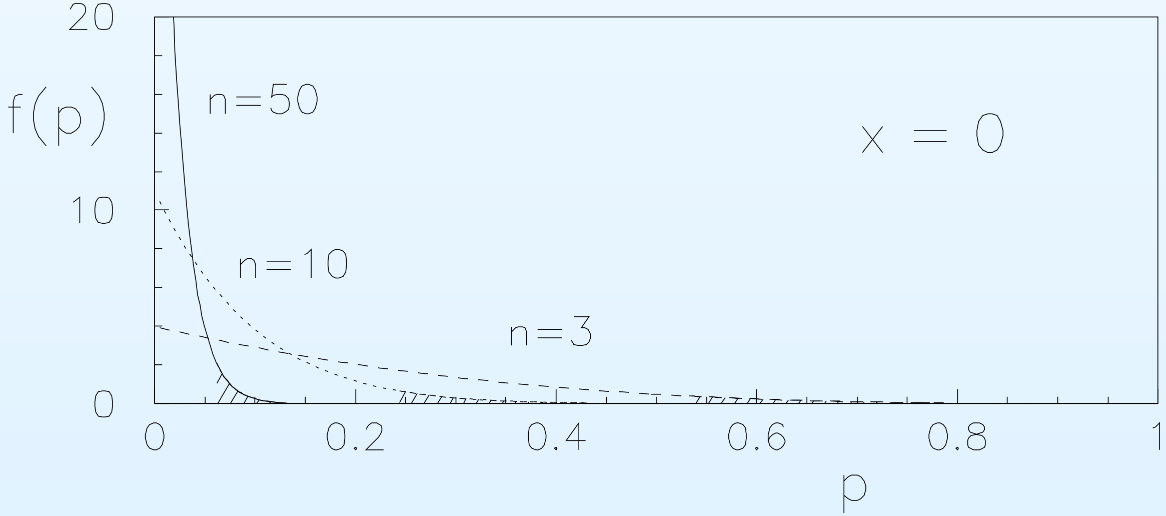
First application: inferring Bernoulli p from n trials with x successes (taking a uniform prior for p)

$$\begin{aligned}
f(p \,|\, x, n, \mathcal{B}) &= \frac{f(x \,|\, \mathcal{B}_{n,p})\, f_\circ(p)}{\int_0^1 f(x \,|\, \mathcal{B}_{n,p})\, f_\circ(p)\, dp} \\
&= \frac{\frac{n!}{(n-x)!\,x!}\, p^x\,(1-p)^{n-x}\, f_\circ(p)}{\int_0^1 \frac{n!}{(n-x)!\,x!}\, p^x\,(1-p)^{n-x}\, f_\circ(p)\, dp} \\
&= \frac{p^x\,(1-p)^{n-x}}{\int_0^1 p^x\,(1-p)^{n-x}\, dp}\,,
\end{aligned}$$

$f(p \,|\, x, n, \mathcal{B})$, $\mathrm{E}(p)$, $\sigma(p)$

$$f(p \,|\, x, n, \mathcal{B}) = \frac{(n+1)!}{x!\,(n-x)!}\, p^x\,(1-p)^{n-x}\,,$$





$f(p \mid x, n, \mathcal{B})$, E$(p)$, $\sigma(p)$

$$f(p \mid x, n, \mathcal{B}) = \frac{(n+1)!}{x!\,(n-x)!}\, p^x\, (1-p)^{n-x}\,,$$

$$\begin{aligned}
\mathrm{E}(p) &= \frac{x+1}{n+2} \qquad \boxed{\text{Laplace's rule of successions}} \\
\mathrm{Var}(p) &= \frac{(x+1)(n-x+1)}{(n+3)(n+2)^2} \\
&= \mathrm{E}(p)\,\left(1-\mathrm{E}(p)\right)\,\frac{1}{n+3} \\
\sigma(p) &= \sqrt{\mathrm{Var}(p)}\,.
\end{aligned}$$

Interpretation of E(p)

Interpretation of E(p). Imagine any future event $E_{i>n}$, thinking that, if we were sure of p then our confidence on $E_{i>n}$ will be exactly p, i.e. $P(E_i \,|\, p) = p$.

Interpretation of E(p)

Interpretation of E(p). Imagine any future event $E_{i>n}$, thinking that, if we were sure of p then our confidence on $E_{i>n}$ will be exactly p, i.e. $P(E_i \,|\, p) = p$.

But we are uncertain about p.
How much should we believe $E_{i>n}$?.

Interpretation of $\mathsf{E}(p)$

Interpretation of $\mathsf{E}(p)$. Imagine any future event $E_{i>n}$, thinking that, if we were sure of p then our confidence on $E_{i>n}$ will be exactly p, i.e. $P(E_i \,|\, p) = p$.

But we are uncertain about p.
How much should we believe $E_{i>n}$?.

$$\begin{aligned}
P(E_{i>n} \,|\, x, n, \mathcal{B}) &= \int_0^1 P(E_i \,|\, p)\, f(p \,|\, x, n, \mathcal{B})\, \mathrm{d}p \\
&= \int_0^1 p\, f(p \,|\, x, n, \mathcal{B})\, \mathrm{d}p \\
&= \mathsf{E}(p) \\
&= \frac{x+1}{n+2} \qquad \text{(for uniform prior)}\,.
\end{aligned}$$

From relative frequencies to probabilities

$$\begin{aligned} \mathsf{E}(p) &= \frac{x+1}{n+2} \\ \mathsf{Var}(p) &= \mathsf{E}(p)\,(1-\mathsf{E}(p))\,\frac{1}{n+3}. \end{aligned}$$

Laplace's rule of successions

For 'large' n, x and $n-x$ (in practice $\geq \mathcal{O}(10)$ is enough for many practical purposes), asymptotic behaviors of $f(p)$:

$$\begin{aligned} \mathsf{E}(p) &\approx p_m = \frac{x}{n} \quad \text{[with } p_m \text{ mode of } f(p)\text{]} \\ \sigma_p &\approx \sqrt{\frac{p_m\,(1-p_m)}{n}} \xrightarrow[n\to\infty]{} 0 \\ p &\sim \mathcal{N}(p_m, \sigma_p)\,. \end{aligned}$$

Under these conditions the frequentistic "definition" (evaluation rule!) of probability (x/n) is recovered.

Estimating Poisson λ

It becomes now an exercise, at least using a uniform prior on λ (not appropriate when searching for rare processes!)

$$f(\lambda\,|\,x,\mathcal{P}) = \frac{\frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)}{\int_0^\infty \frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)\,\mathrm{d}\lambda}\,.$$

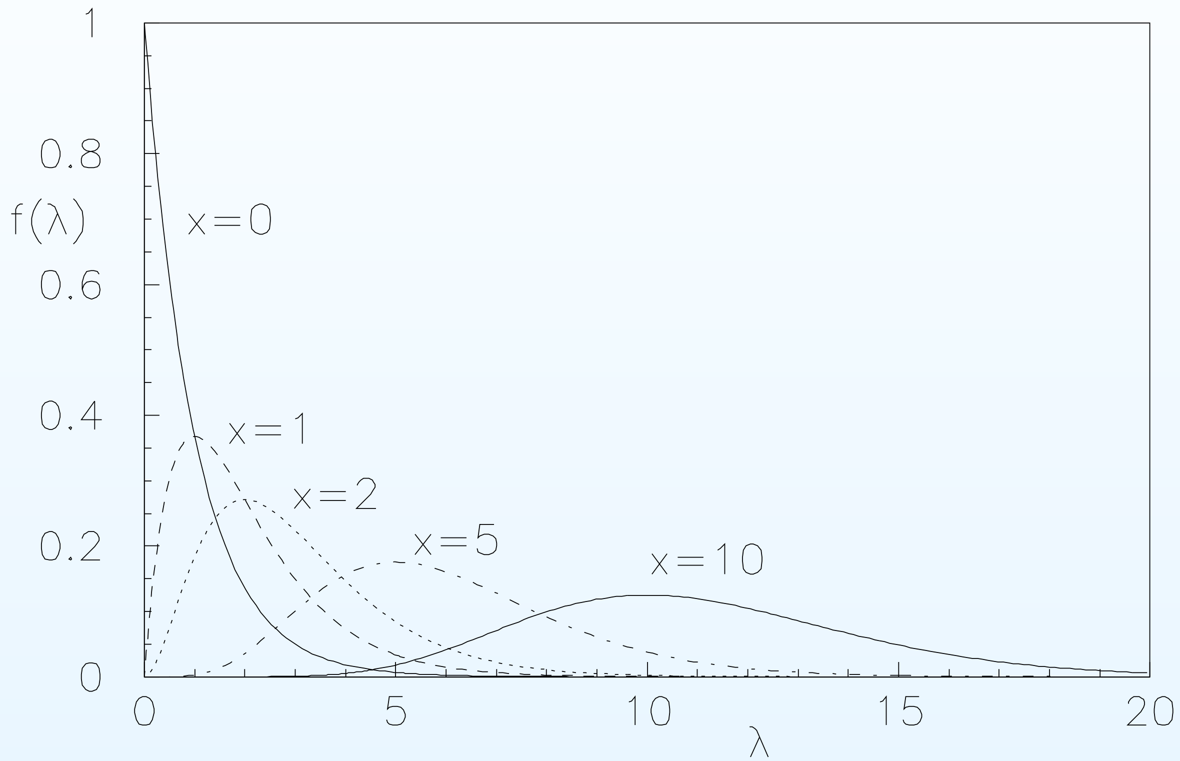
$$f(\lambda\,|\,x,\mathcal{P}) = \frac{\lambda^x\,e^{-\lambda}}{x!}$$

$$F(\lambda\,|\,x,\mathcal{P}) = 1 - e^{-\lambda}\left(\sum_{n=0}^{x}\frac{\lambda^n}{n!}\right),$$

Expected value, variance and mode of the probability distribution are

$$\begin{aligned} \mathrm{E}(\lambda) &= x+1, \\ \mathrm{Var}(\lambda) &= x+1, \\ \lambda_m &= x\,. \end{aligned}$$

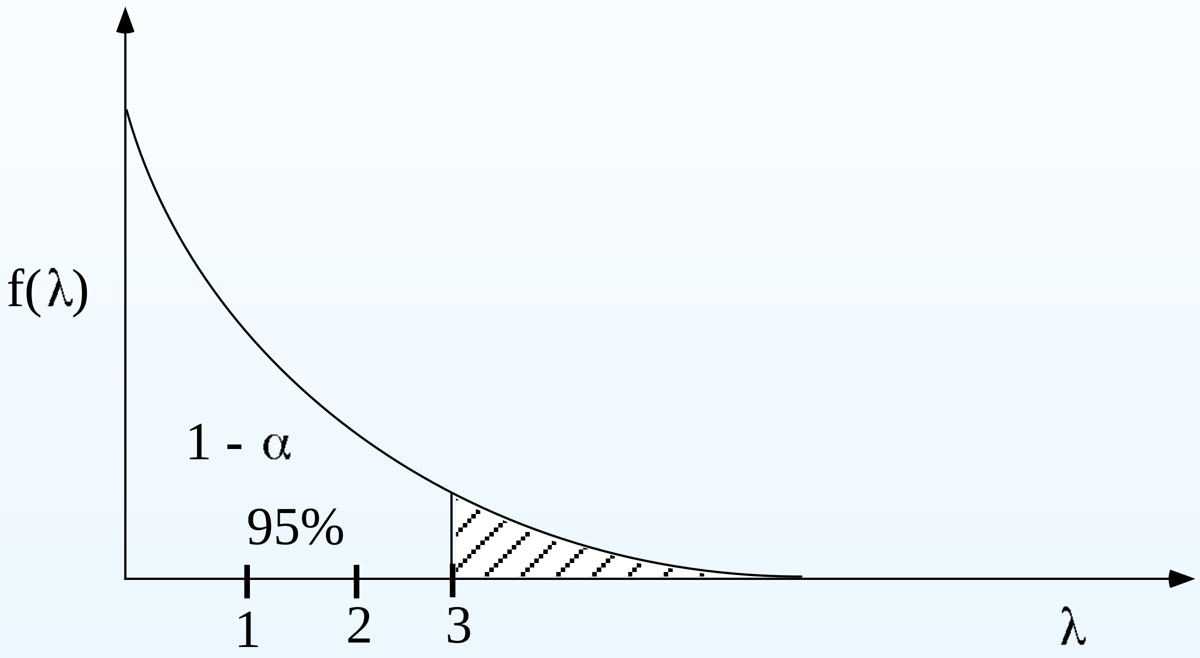
Some examples of $f(\lambda)$



For 'large' x $f(\lambda)$ becomes Gaussian with expected value x and standard deviation $\sqrt{x}$.

The difference between most probable λ and its expected value for small x is due to the asymmetry of $f(\lambda)$.

case of observed $x = 0$



$$
\begin{aligned}
f(\lambda \,|\, x=0, \mathcal{P}) &= e^{-\lambda}, \\
F(\lambda \,|\, x=0, \mathcal{P}) &= 1 - e^{-\lambda}, \\
\lambda &< 3 \text{ at } 95\,\% \text{ probability}\,.
\end{aligned}
$$

But not just because $f(x=0 \,|\, \mathcal{P}_{\lambda=3})$ *= 0.05! In this case it works by chance*

Frequentistic upper/lower limits

Only in the Poisson case we have that, *assuming a flat prior*

$$f(x=0\,|\,\mathcal{P}_3) = \int_3^\infty f(\lambda\,|\,x=0,\mathcal{P})\,\mathrm{d}\lambda\,.$$

Not true in general!

although this is the (somehow) way frequentistic upper/lower limits are calculated.

Frequentistic upper/lower limits

Only in the Poisson case we have that, *assuming a flat prior*

$$f(x=0\,|\,\mathcal{P}_3) = \int_3^\infty f(\lambda\,|\,x=0,\mathcal{P})\mathrm{d}\lambda\,.$$

Not true in general!

although this is the (somehow) way frequentistic upper/lower limits are calculated.

$\longrightarrow$ This is the reason why the lower bound on Higgs mass does not mean that M_H is above that limit with 95% probability!

Frequentistic upper/lower limits

Only in the Poisson case we have that, *assuming a flat prior*

$$f(x=0\,|\,\mathcal{P}_3) = \int_3^\infty f(\lambda\,|\,x=0,\mathcal{P})\mathrm{d}\lambda\,.$$

Not true in general!

although this is the (somehow) way frequentistic upper/lower limits are calculated.

$\longrightarrow$ This is the reason why the lower bound on Higgs mass does not mean that M_H is above that limit with 95% probability!

That is simply the mass value such that there is 5% probability to observe a number of events equal or less thasn the observed number

Frequentistic upper/lower limits

Only in the Poisson case we have that, *assuming a flat prior*

$$f(x=0\,|\,\mathcal{P}_3) = \int_3^\infty f(\lambda\,|\,x=0,\mathcal{P})\mathrm{d}\lambda\,.$$

Not true in general!

although this is the (somehow) way frequentistic upper/lower limits are calculated.

$\rightarrow$ Instead, just 'by chance', the upper M_H value can be interpreted in a probabilistic way, because it comes from a different likelihood (Gaussian in $\log M_H$, due to radiative corrections).

Isn't it ridiculous?

Adding background of expected intensity

Two independent Poisson processes, the signal one of intensity r_S and the background one of r_B:
$r = r_S + r_B \rightarrow \lambda = \lambda_S + \lambda_B$.

If λ_B is somehow known (though uncertain) we can infer λ_S from the observed numbers of events x:

$$f(\lambda_S \,|\, x, \lambda_{B_\circ}) = \frac{e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)}{\int_0^\infty e^{-(\lambda_{B_\circ}+\lambda_S)}\,(\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)\,d\lambda_S}\,.$$

$$f(\lambda_S \,|\, x, \lambda_{B_\circ}) = \frac{e^{-\lambda_S}\,(\lambda_{B_\circ}+\lambda_S)^x}{x!\,\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}\,,$$

$$F(\lambda_S \,|\, x, \lambda_{B_\circ}) = 1 - \frac{e^{-\lambda_S}\,\sum_{n=0}^{x}\frac{(\lambda_{B_\circ}+\lambda_S)^n}{n!}}{\sum_{n=0}^{x}\frac{\lambda_{B_\circ}^n}{n!}}\,.$$

(If we are uncertain *about the background we* model the uncertainty with $f(\lambda_B)$*, and apply once more probability rules, as we shall see later)*

Uncertainty on the expected value of background

What happens if λ_B is not exactly know?

Uncertainty on the expected value of background

What happens if λ_B is not exactly know?
No problem (withing the probabilistic approach):

- uncertain λ_B: $\longrightarrow f(\lambda_B)$;
- use probability theory:

$$f(\lambda_S \,|\, x) \;\;=\;\; \int_0^\infty f(\lambda_S \,|\, x, \lambda_{B_\circ}) \cdot f(\lambda_B)\, d\lambda_B$$

Uncertainty on the expected value of background

What happens if λ_B is not exactly know?
No problem (withing the probabilistic approach):

- uncertain λ_B: $\rightarrow f(\lambda_B)$;
- use probability theory:

$$f(\lambda_S\,|\,x) \;=\; \int_0^{\infty} f(\lambda_S\,|\,x, \lambda_{B_\circ}) \cdot f(\lambda_B)\, d\lambda_B$$

This is the general way to treat systematics

- $f(\boldsymbol{\theta}\,|\,\text{data}) \;\rightarrow\; f(\boldsymbol{\theta}\,|\,\text{data}, \boldsymbol{h})$

$\Rightarrow\; f(\boldsymbol{\theta}\,|\,\text{data}) = \int f(\boldsymbol{\theta}\,|\,\text{data}, \boldsymbol{h}) \cdot f(\boldsymbol{h}\; d\boldsymbol{h}$

(This integral can be done by MC)

The Gaussian model

Gaussian case left on purpose at the end, because I find that it can be dis-educative

- tendency to believe that everything must be so nicely bell-shaped
- methods only valid for Gaussian are sometime acritically used elsewhere
- (I have even found teachers explaining that the standard deviation is ‘the 68% thing‘. . .)

The Gaussian model

Gaussian case left on purpose at the end, because I find that it can be dis-educative

- tendency to believe that everything must be so nicely bell-shaped
- methods only valid for Gaussian are sometime acritically used elsewhere
- (I have even found teachers explaining that the standard deviation <u>is</u> ‘the 68% thing‘...)

$\longrightarrow$ See slides:

- simple inference with very vague prior
- inference with ’narrow’ prior: $\longrightarrow$ combinations
- predictive distributions
- measuring at the edge of the physical region
- more on systematics

General probabilistic inference $\longrightarrow$ simple fit formulae

How several ‘standard’ methods can be recovered under well defined assumptions:

$\longrightarrow$ Slides

But be careful: simplified methods fail in case of not trivial χ^2 curves, etc.

- For a detailed example, see Chapter 8 of book “Bayesian Reasoning in Data Analysis”, (World Scientific, 2003)
- containing also the rigorous treatment of linear fit with errors on both axes (and hints for non-linear fit).

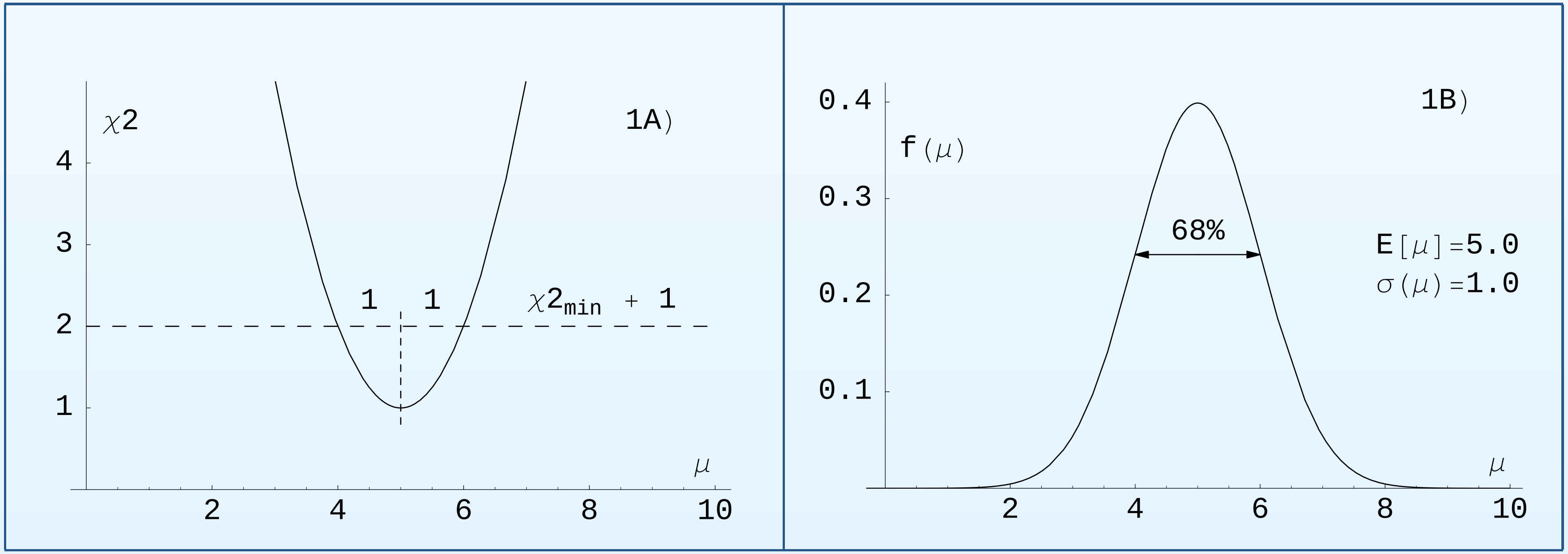
General probabilistic inference $\longrightarrow$ simple fit formulae

How several ‘standard’ methods can be recovered under well defined assumptions, as also known to Fermi, I have found out recently:

> *“In my thesis I had to find the best 3-parameter fit to my data and the errors of those parameters in order to get the 3 phase shifts and their errors. Fermi showed me a simple analytic method. At the same time other physicists were using and publishing other cumbersome methods. Also Fermi taught me a general method, which he called Bayes Theorem, where one could easily derive the best-fit parameters and their errors as a special case of the maximum-likelihood method. I remember asking Fermi how and where he learned this. I expected him to answer R.A. Fisher or some other textbook on mathematical statistics. Instead he said ‘perhaps it was Gauss’. I suspect he was embarrassed to admit that he had derived it all from his ‘Bayes Theorem’.”* (J. Orear)

Use and misuse of χ^2 fits

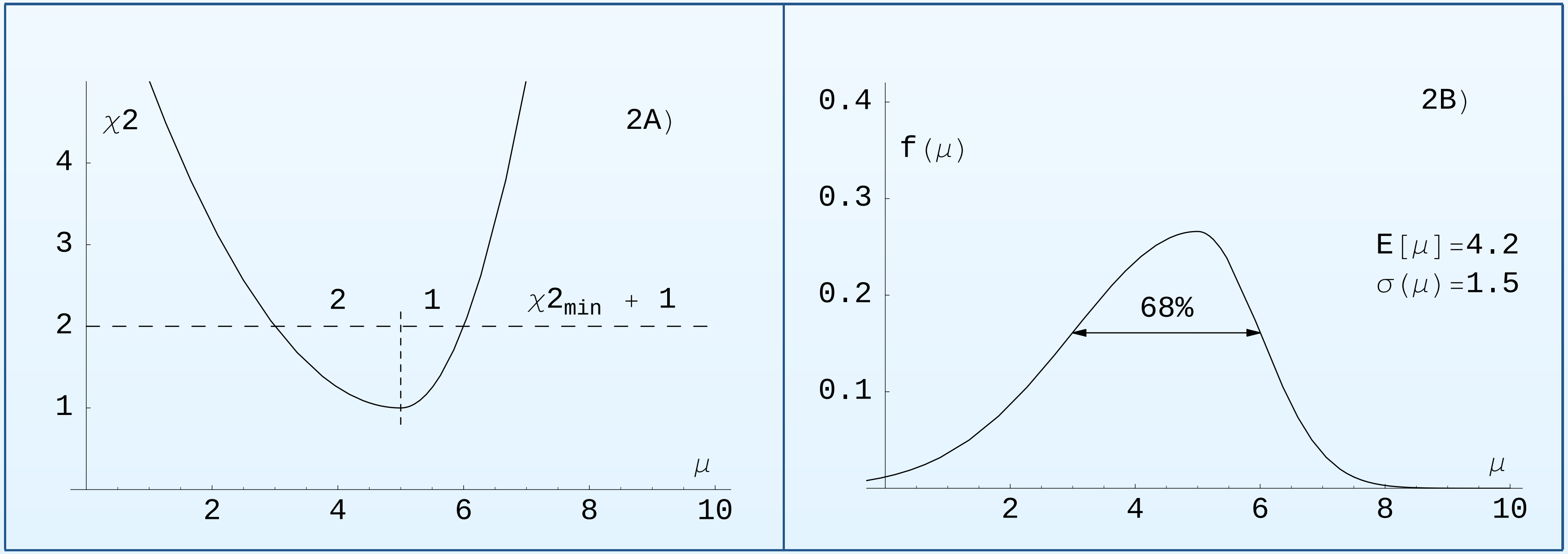
$$
\begin{aligned}
f(\mu \,|\, \text{data}) &\propto f(\text{data} \,|\, \mu) \cdot f_\circ(\mu) \\
&\propto f(\text{data} \,|\, \mu) \\
&\propto e^{-\chi^2/2}
\end{aligned}
$$



Parabolic χ^2: <u>OK</u> both σ and probability

Use and misuse of χ^2 fits

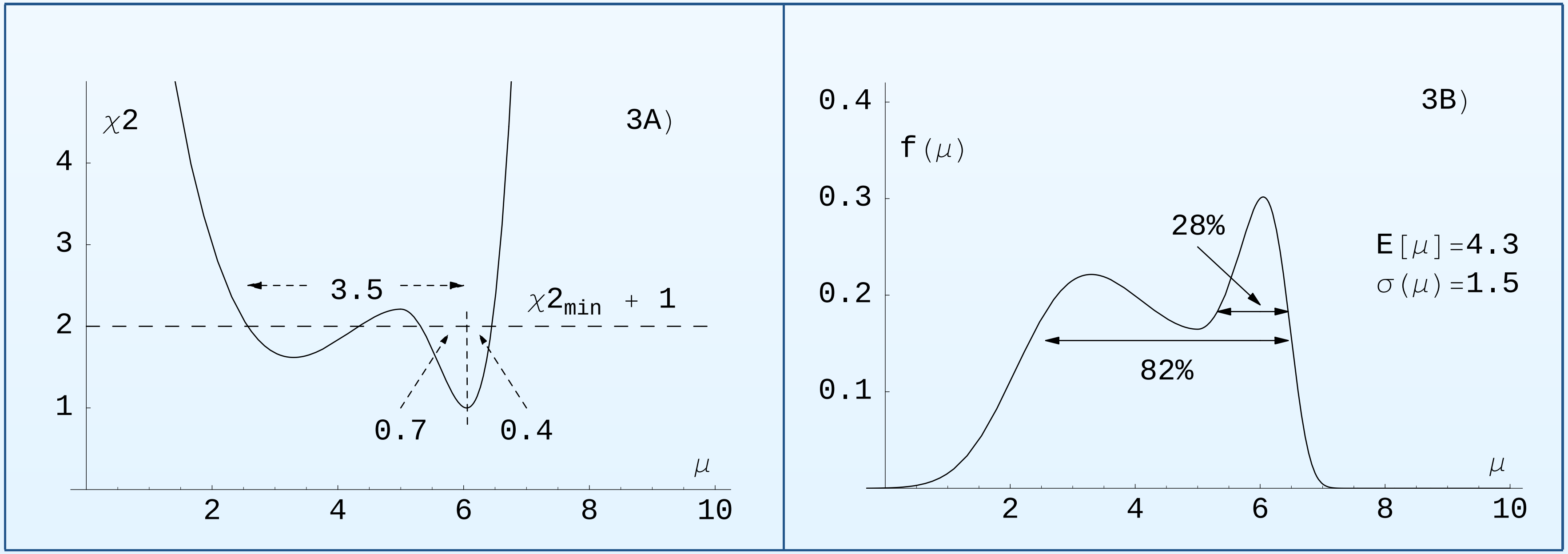
$$
\begin{aligned}
f(\mu\,|\,\text{data}) &\propto f(\text{data}\,|\,\mu)\cdot f_\circ(\mu)\\
&\propto f(\text{data}\,|\,\mu)\\
&\propto e^{-\chi^2/2}
\end{aligned}
$$



Slight asymmetry: probability OK, σ NO

Use and misuse of χ^2 fits

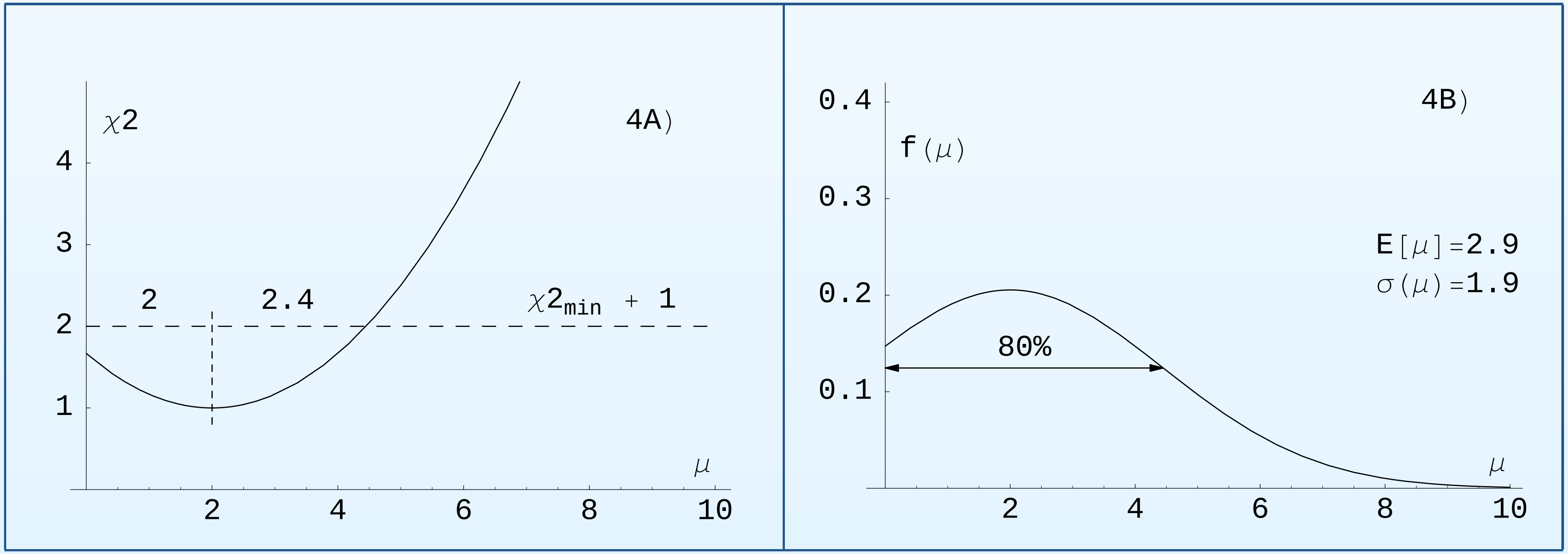
$$
\begin{aligned}
f(\mu \,|\, \text{data}) &\propto f(\text{data} \,|\, \mu) \cdot f_\circ(\mu) \\
&\propto f(\text{data} \,|\, \mu) \\
&\propto e^{-\chi^2/2}
\end{aligned}
$$



χ^2 gets crazy results!

Use and misuse of χ^2 fits

$$
\begin{aligned}
f(\mu \,|\, \text{data}) &\propto f(\text{data} \,|\, \mu) \cdot f_\circ(\mu) \\
&\propto f(\text{data} \,|\, \mu) \\
&\propto e^{-\chi^2/2}
\end{aligned}
$$



Same when χ^2 parabolic, but bounded!

Propagation of uncertainties

Easy task in the probabilistic approach:
$\Longrightarrow$ Just use probability theory

Propagation of uncertainties

Easy task in the probabilistic approach:
$\Rightarrow$ Just use probability theory

The general problem:

$$f(x_1, x_2, \ldots, x_n) \xrightarrow[Y_j = Y_j(X_1, X_2, \ldots, X_n)]{} f(y_1, y_2, \ldots, y_m)\,.$$

This calculation can be quite challenging, but it can be easily performed by Monte Carlo techniques.

General solution for discrete variables

$Y = Y(X)$, where $Y()$ stands for the mathematical function relating X and Y.

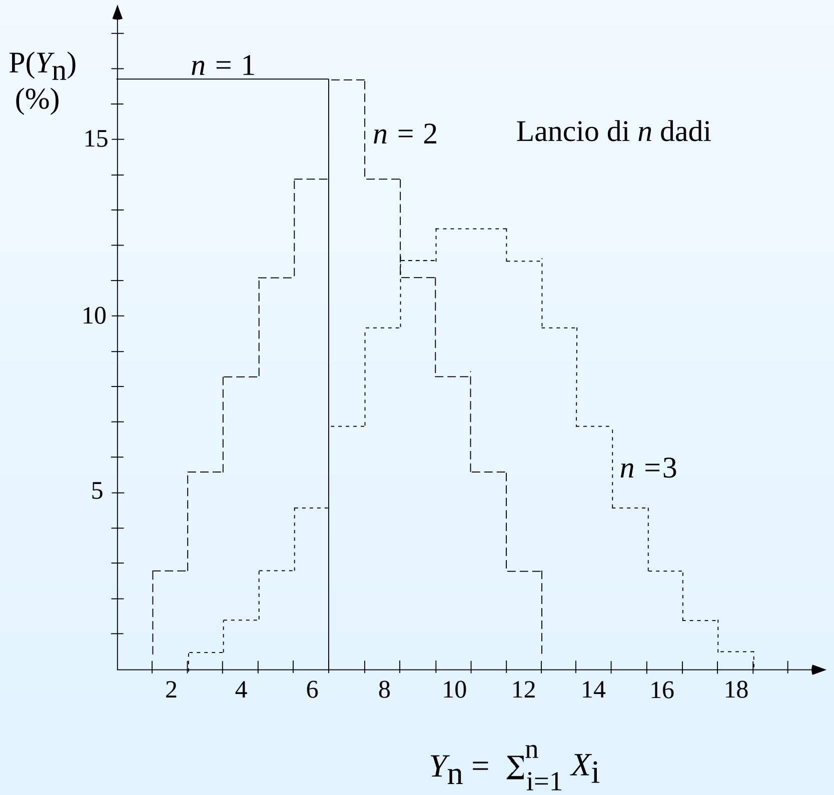
The probability of a given $Y = y$ is equal to the sum of the probability of each X such that $Y(X = x) = y$.

General solution for discrete variables

$Y = Y(X)$, where $Y()$ stands for the mathematical function relating X and Y.

The probability of a given $Y = y$ is equal to the sum of the probability of each X such that $Y(X = x) = y$.

Probability distributions of the sums of the results from n dice.



General solution for continuous variable

Just extend to the continuum the previous reasoning:

- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

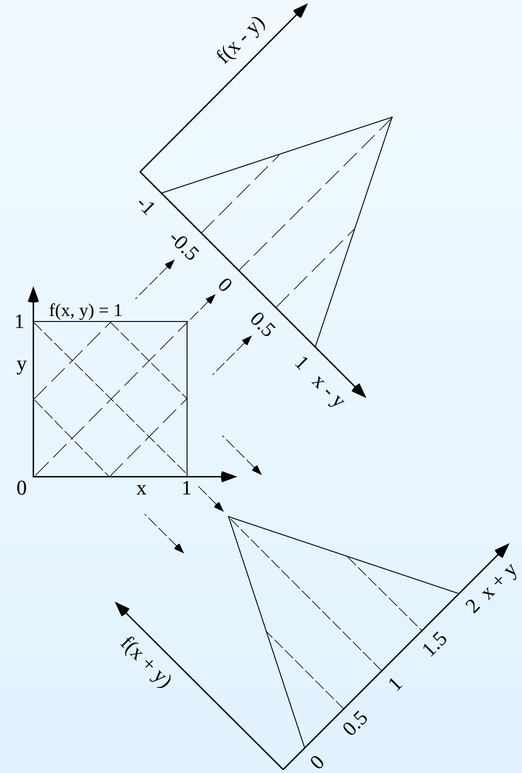
$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\, \delta(y_2 - Y_2(x_1, y_2))\, f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2\,.$$

General solution for continuous variable

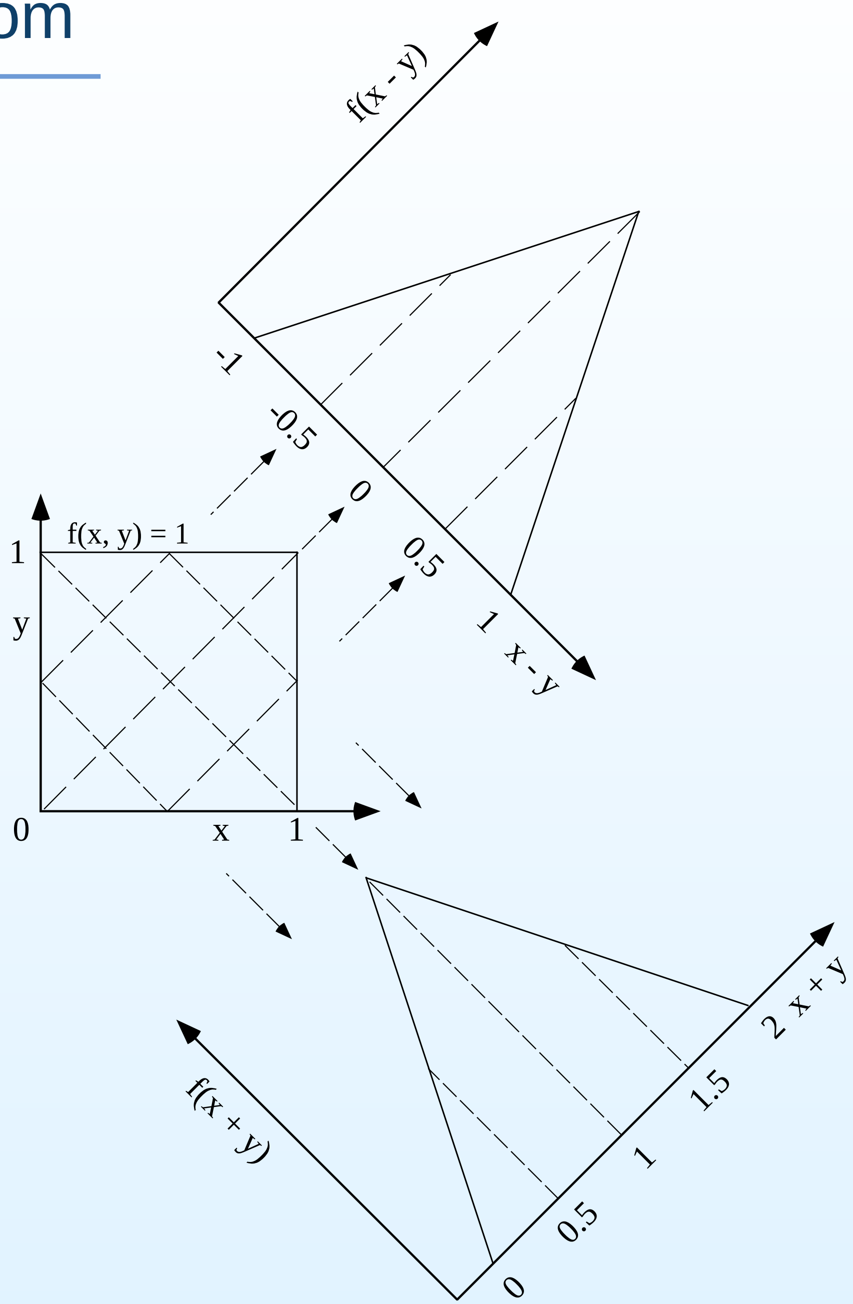
Just extend to the continuum the previous reasoning:

- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\,\delta(y_2 - Y_2(x_1, y_2))\,f(x_1, x_2)\,\mathrm{d}x_1 \mathrm{d}x_2\,.$$



Zoom



Monte Carlo implementation of the general formula

$$f(y_1,y_2) = \int \delta(y_1 - Y_1(x_1,x_2))\,\delta(y_2 - Y_2(x_1,y_2))\,f(x_1,x_2)\,\mathrm{d}x_1\mathrm{d}x_2\,.$$

Monte Carlo implementation of the general formula

- Extract a point $\{x_1,x_2\}$ according to $f(x_1,x_2)$
- Fill a table (or scatter plot) with the entry

$$\begin{aligned} y_1 &= Y_1(x_1,x_2) \\ y_2 &= Y_2(x_1,x_2) \end{aligned}$$

- Do it many times; then from the relative frequencies in each 2-D bin we can estimate the probability in each bin: $f(y_1,y_2)\,\Delta y_1 \Delta y_2$, and hence $f(y_1,y_2)$. ($\rightarrow$ examples in R)

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode (‘point of maximum belief‘) or median (‘fifty-fifty point‘)?

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode (‘point of maximum belief‘) or median (‘fifty-fifty point‘)?

- no ‘deep’ reason: just math,

  and this the main reason that makes expected value and variance so convenient.
- General property:

Expected value and variance of a linear combination

Why E$(Y) = $ E$(X_1) +$ E(X_2) and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned} \mathrm{E}(Y) &= \sum_i c_i \, \mathrm{E}(X_i) \\ \sigma^2_Y &= \sum_i c_i^2 \, \sigma^2(X_i) \end{aligned}$$

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode (‘point of maximum belief‘) or median (‘fifty-fifty point‘)?

- no ‘deep’ reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

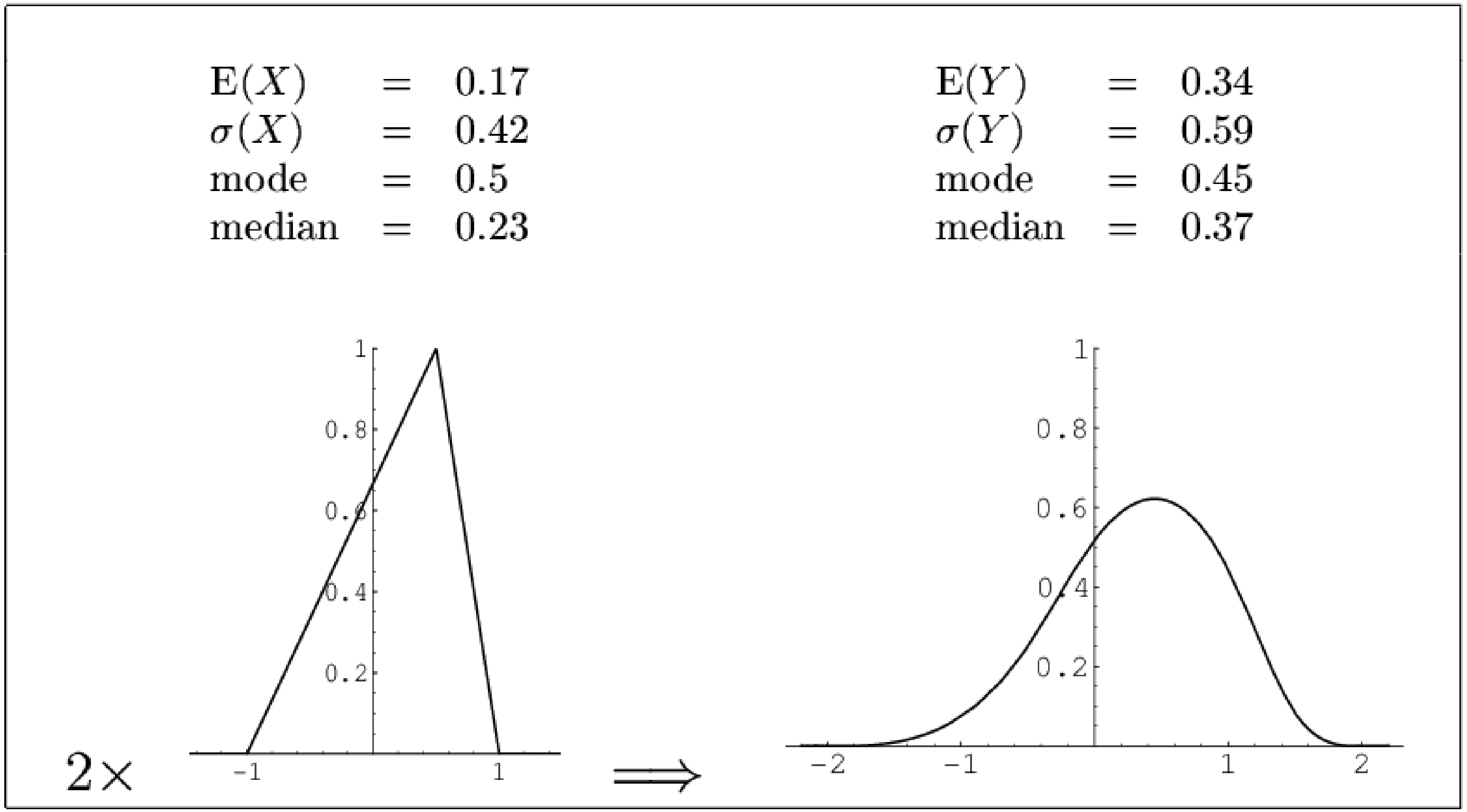
$$\begin{aligned} \mathsf{E}(Y) &= \sum_i c_i\, \mathsf{E}(X_i) \\ \sigma^2_Y &= \sum_i c_i^2\, \sigma^2(X_i) + 2 \sum_{i<j} c_i\, c_j\, \mathsf{Cov}(X_i, X_j) \end{aligned}$$

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our ‘asymmetric’ example!



No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP
  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP

  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

$\longrightarrow$ asymmetric χ^2 or log-likelihoods

$\longrightarrow$ asymmetry in – well treated! – uncertainty propagations

$\longrightarrow$ systematics (often related to non linear propagation)

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP
  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

$\longrightarrow$ asymmetric χ^2 or log-likelihoods

$\longrightarrow$ asymmetry in – well treated! – uncertainty propagations

$\longrightarrow$ systematics (often related to non linear propagation)

And remember that standard methods (χ^2 or ML fits) provide something equivalent to 'most probable values', not to E( )!

If we really have to give only two numbers. . .

. . . they should be, anyway,

- Expected value
- Standard deviation

Because this is what we need in simple propagations, using the well known formula of propagation, while – let’s repeat it – no general combination formula exists for other summaries.

If we really have to give only two numbers. . .

. . . they should be, anyway,

- Expected value
- Standard deviation

Because this is what we need in simple propagations, using the well known formula of propagation, while – let's repeat it – no general combination formula exists for other summaries.

There is also another property that make E( ) and σ very convenient:

The Central Limit Theorem

$\Rightarrow$ Result of combination is approximately Gaussian under hypotheses that 'often' hold (but always check!)

[But you can imagine that in other approaches where the expected value of a physics quantity is an absurd concept, there might be some problems. And this explains the 'prescriptions' that surrogate the luck of theoretical guidance!]

Which prior for frontier physics?

In many cases of frontier all methods can be misleading, included those based on the Bayes formula

$\longrightarrow$ Anyway, it is important to understand the probabilistic reasoning behind Bayesian methods

- In many frontier cases we just lose experimental sensitivity *around* some edge, and therefore we are unable to state our confidence that the value is before of after the edge
- ~~Confidence limits~~ $\longrightarrow$sensitivity bounds

  $\rightarrow$ see contribution at the CERN 2000 Confidence Limit Workshop, *"Confidence limits: what is the problem? Is there the solution?"*, (hep-ex/0002055)

$\longrightarrow$ PUBLISH LIKELIHOOD! (possibly in the rescaled form it will be shown).

$\longrightarrow r$ of a Poisson process in presence of bkgd

Rewriting in terms of r what we have sees before for λ:

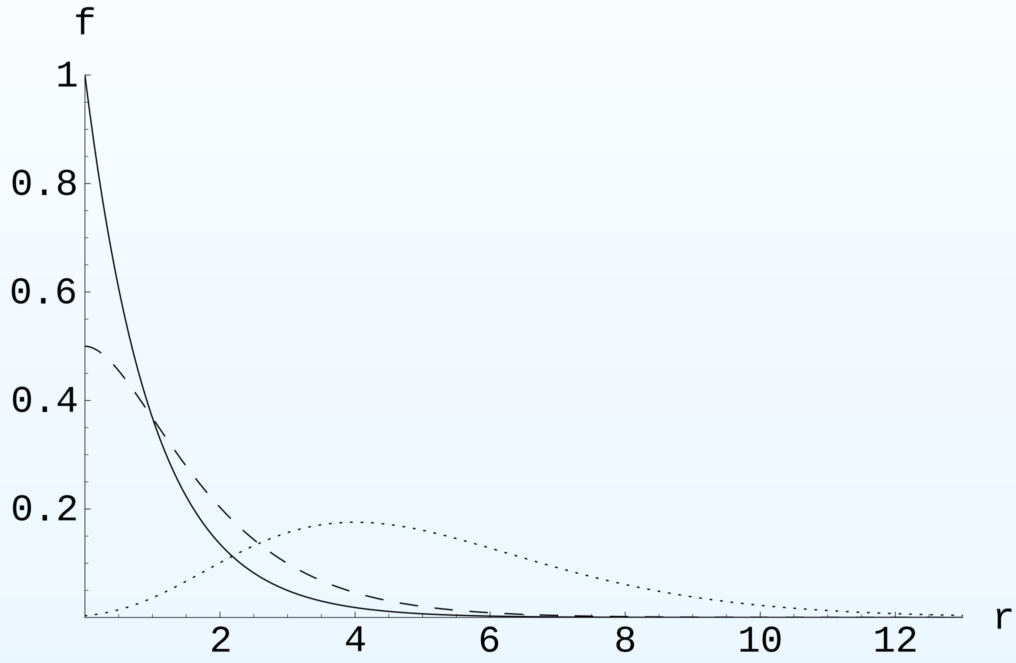
$$f(r\,|\,n_c, r_b) \propto \frac{e^{-(r+r_b)\,T}\left((r+r_b)\,T\right)^{n_c}}{n_c!}\, f_\circ(r)\,.$$

Uniform prior:

$$f(r\,|\,n_c, r_b, f_\circ(r)=k) = \frac{e^{-r\,T}\left((r+r_b)\,T\right)^{n_c}}{n_c!\,\sum_{n=0}^{n_c}\frac{(r_b\,T)^n}{n!}}\,.$$

where r_b is the expected rate of the background and n_c the observed number of counts.

An example of inferring r

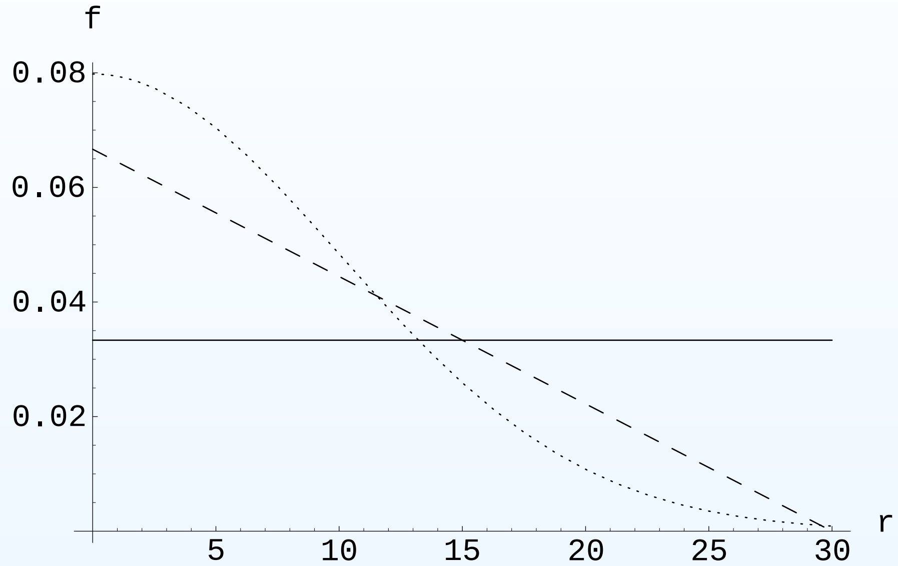


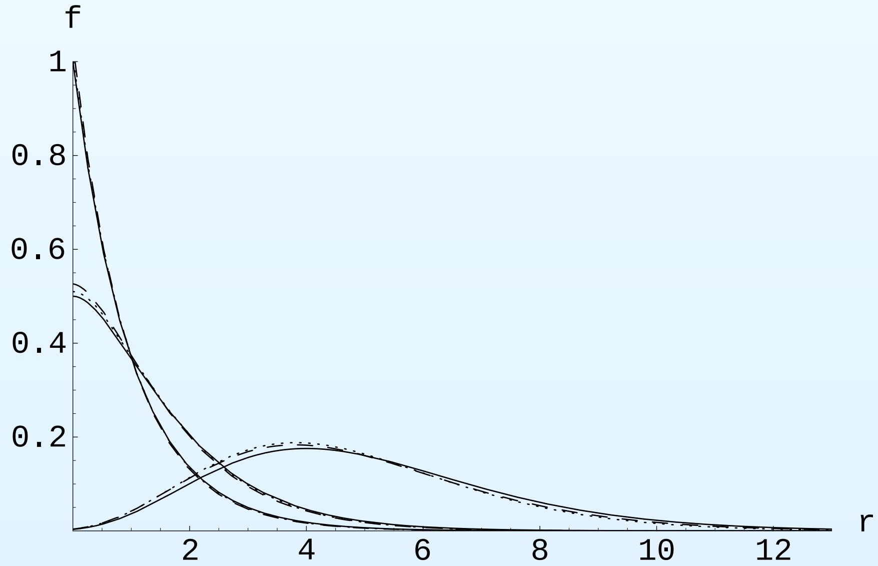
Distribution of the values of the rate r, in units of events/month, inferred from an expected rate of background events $r_b = 1$ event/month, an initial uniform distribution $f_\circ(r) = k$ and the following numbers of observed events: 0 (solid); 1 (dashed); 5 (dotted).

$\rightarrow$ which impression do you get? Do you see a serious problem?

Dependence for ‘optimistic priors’

Upper plot shows some reasonable priors reflecting the positive attitude of researchers: little influence on posterior!

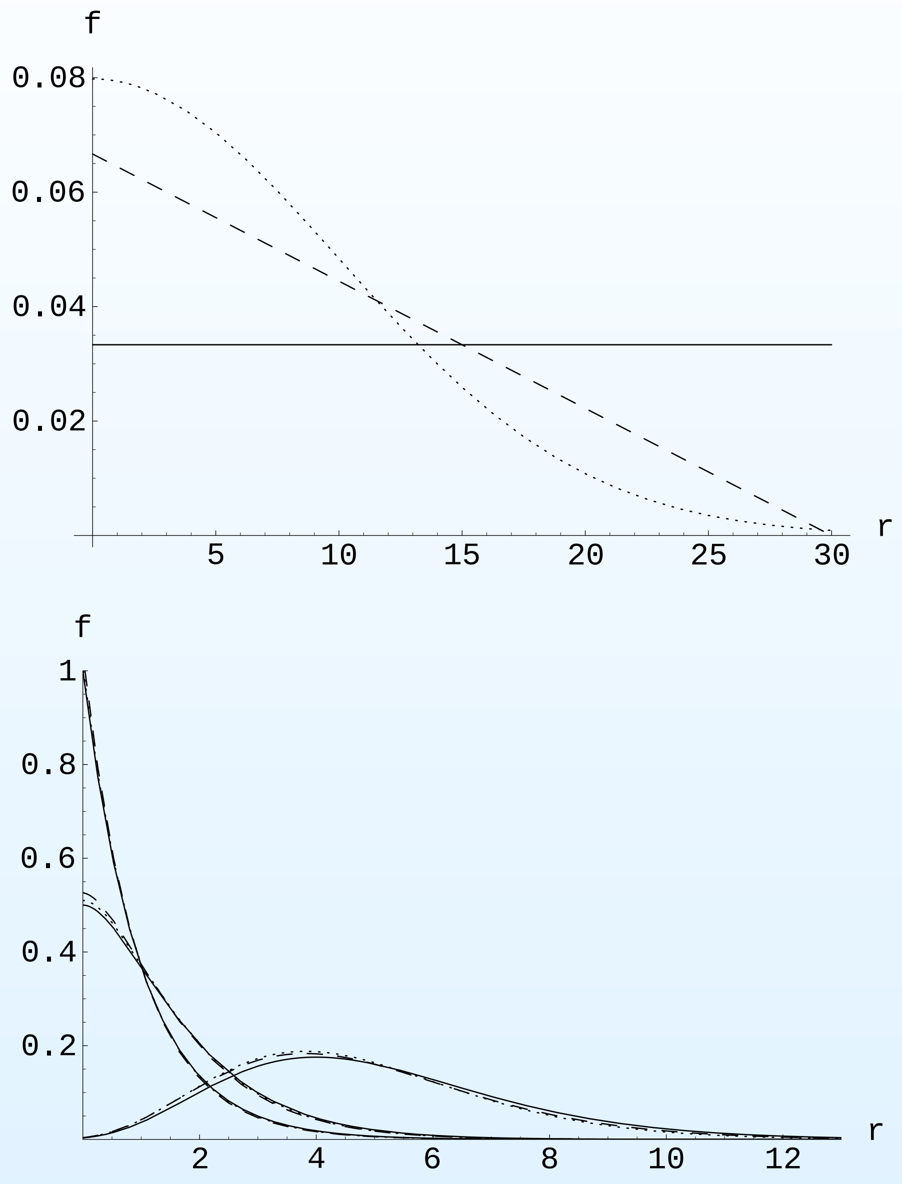




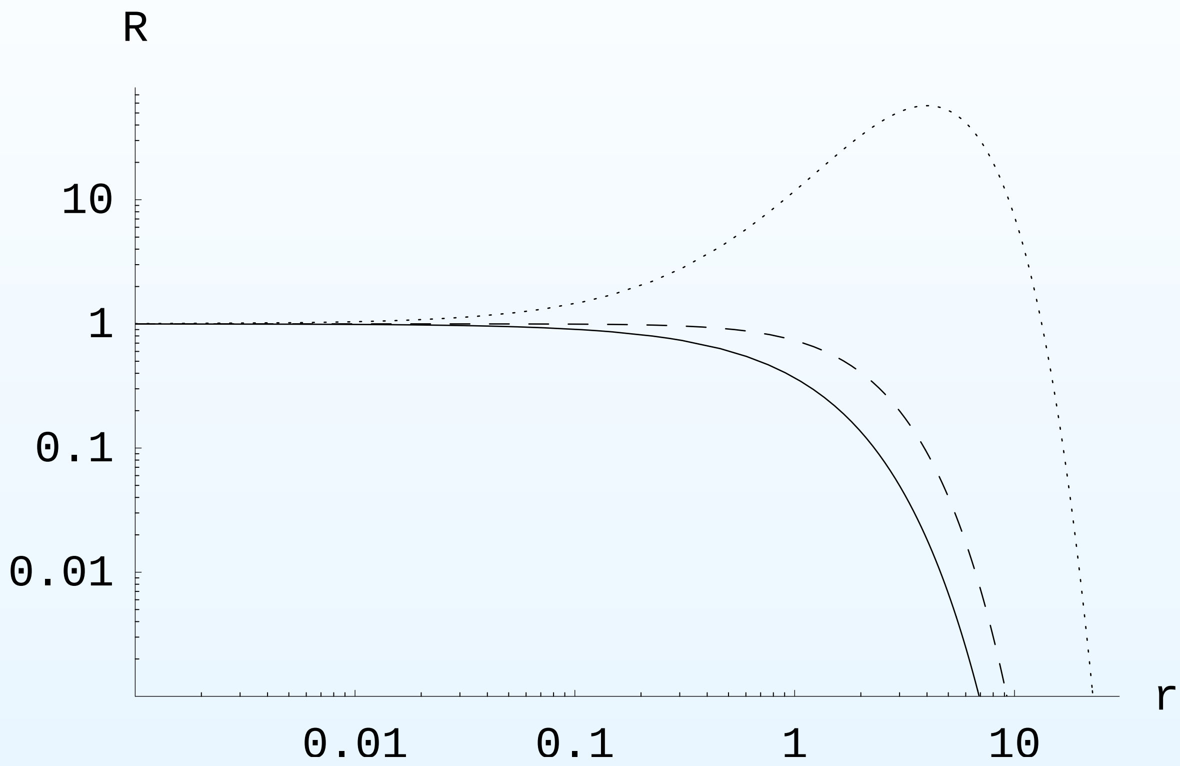
Dependence for 'optimistic priors'

Upper plot shows some reasonable priors reflecting the positive attitude of researchers: little influence on posterior!

But the priors could be concentrated at very low values of r (think e.g. gravitation wave search, or an 'exploratory' first experiment of a rare process, without real hope of finding something!)



Rescaled likelihood (R function)

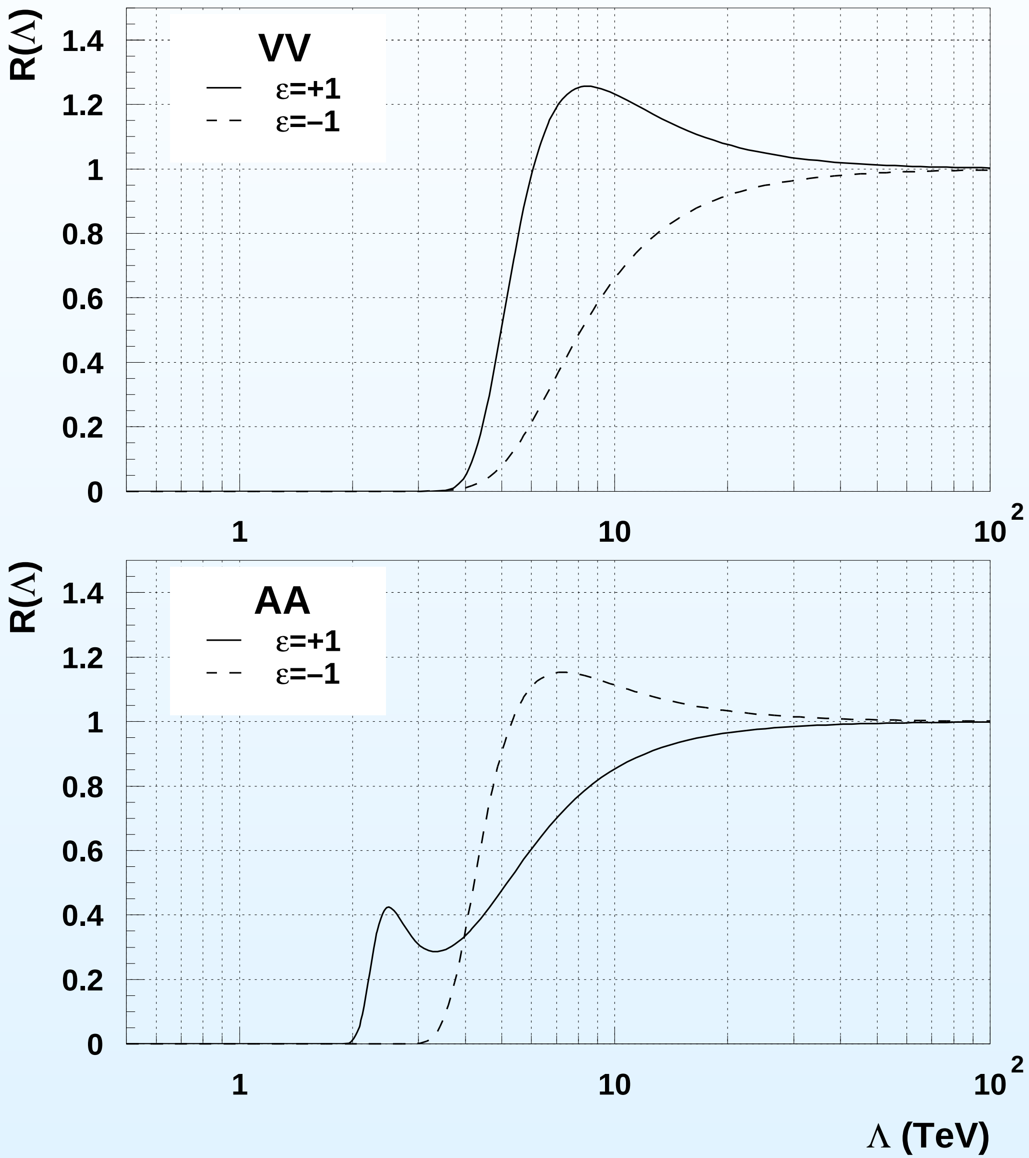


'Relative belief updating ratio' $\mathcal{R}$ *for the Poisson intensity parameter* r *for above cases.* *Note log scales!*

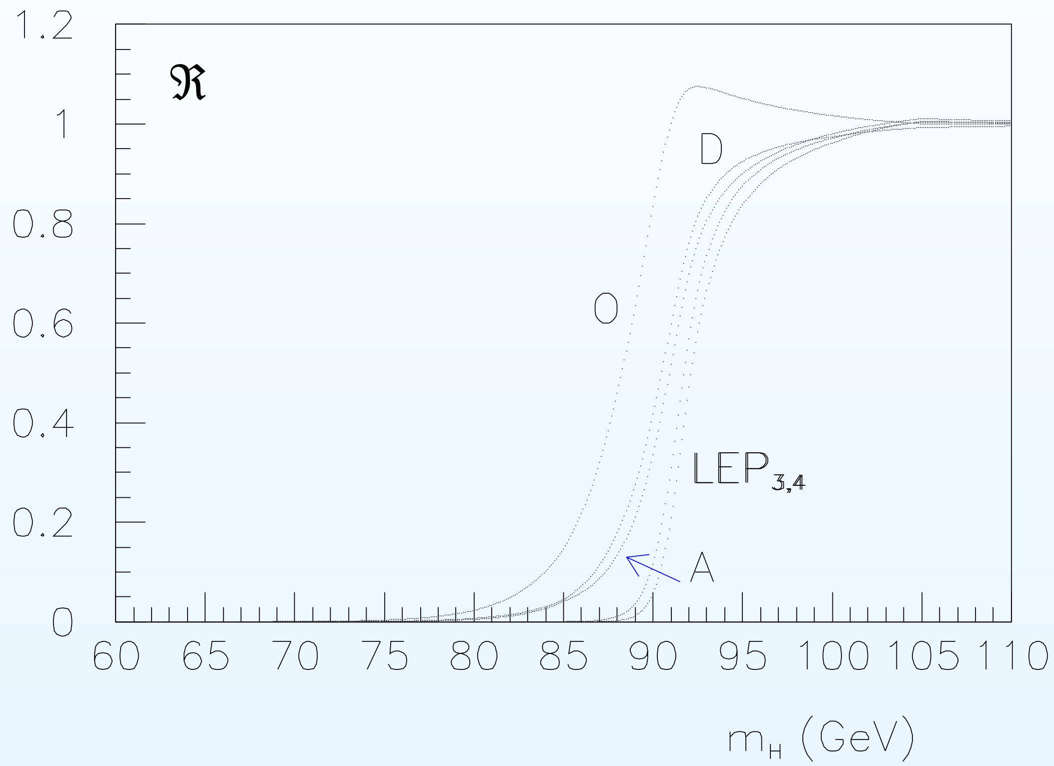
This figure gives a precise picture of what is going on!

Also clear what a sensitivity bound is, and while "C.L.'s" can be misleading

An example of R from real data (ZEUS)



Higgs mass example (≤ 1998 data)



$\mathcal{R}$-function reporting results on Higgs direct search from the reanalysis performed by GdA & Degrassi. A, D and O stand for ALEPH, DELPHI and OPAL experiments. Their combined result is indicated by LEP$_3$. The full combination (LEP$_4$) was obtained by assuming for L3 experiment a behavior equal to the average of the others experiments.

Which prior for frontier physics?

In many cases of frontier all methods can be misleading, included those based on the Bayes formula

$\longrightarrow$ Anyway, it is important to understand the probabilistic reasoning behind Bayesian methods

- In many frontier cases we just lose experimental sensitivity *around* some edge, and therefore we are unable to state our confidence that the value is before of after the edge
- ~~Confidence limits~~ $\longrightarrow$sensitivity bounds

  $\rightarrow$ see contribution at the CERN 2000 Confidence Limit Workshop, *"Confidence limits: what is the problem? Is there the solution?"*, (hep-ex/0002055)

$\longrightarrow$ PUBLISH LIKELIHOOD! (possibly in the rescaled form).

$\longrightarrow$ EASY COMBINATION OF RESULTS (independent likelihoods factorize).

Conclusions

- Subjective probability recovers intuitive idea of probability.
- It is crucial to perform 'probability inversions'. . .
- on which probabilistic inference is based.
- Very powerfull tools: do 'everything' starting from a single idea.
- 'Conventional methods' can be recovered, *if* they make sense, *when* they meke sense, *until well defined conditions*.
- Priors a logically crucial to maqke the probability inversion, but practically irrelevant if we have enough good data
- otherwise it is absolutely right that they must play a role.

Conclusions – continued

- The case in which priors can be really critical are those at the edge of the detector sensitivity, with ‘open likelihood’.
- In this case it is better to refrain from giving probabilistic result and just report likelihoods (stating clearly what one is doing) and sensitivity bounds.
- The approach is rather natural, easy for young people, harder for seniors corrupted by strange XX-th century ideologies (and with neural synapses stuck. . .).
- Anyway: It’s easy if you try!

End

FINE