Telling the Truth with Statistics

Lecture 5

Giulio D'Agostini

Università di Roma La Sapienza e INFN

Roma, Italy

Overview of the contents

1st part Review of the process of learning from data

Mainly based on

- *“From observations to hypotheses: Probabilistic reasoning versus falsificationism and its statistical variations”* (Vulcano 2004, physics/0412148)
- Chapter 1 of *“Bayesian reasoning in high energy physics. Principles and applications”* (CERN Yellow Report 99-03)

Overview of the contents

1st part Review of the process of learning from data
Mainly based on

- *"From observations to hypotheses: Probabilistic reasoning versus falsificationism and its statistical variations"* (Vulcano 2004, physics/0412148)
- Chapter 1 of *"Bayesian reasoning in high energy physics. Principles and applications"* (CERN Yellow Report 99-03)

2nd part Review of the probability and 'direct probability' problems, including 'propagation of uncertainties.
Partially covered in

- First 3 sections of Chapter 3 of YR 99-03
- Chapter 4 of YR 99-03
- *"Asymmetric uncertainties: sources, treatment and possible dangers"* (physics/0403086)

Overview of the contents

3th part Probabilistic inference and applications to HEP

Much material and references in my web page. In particular, I recommend a quite concise review

- *"Bayesian inference in processing experimental data: principles and basic applications"*, Rep.Progr.Phys. 66 (2003)1383 [physics/0304102]

For a more extensive treatment:,

- *“Bayesian reasoning in data analysis – A critical introduction”*, World Scientific Publishing, 2003

  (CERN Yellow Report 99-03 updated and ≈ doubled in contents)

Status report from previous lecture

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

Status report from previous lecture

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

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:

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

And their sequential use

The posterior becomes the prior of the next inference

For conditionally independent E_i:

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

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

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:

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

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 $P(A,\,B\,C,\ldots) = P(A)\cdot P(B\,|\,A)\cdot P(C\,|\,A,\,B)\ldots$*)*

Today

- More on model comparison
- Parametric inference
- Some applications

→ Goal is to to allow you to read more technical literature understanding the basis of the reasoning and, most of all, without being afraid of the terms ‘subjective’ or ’Bayesian’

Solution of the AIDS test problem

$$\begin{aligned} P(\text{Pos}\,|\,\text{HIV}) &= 100\% \\ P(\text{Pos}\,|\,\overline{\text{HIV}}) &= 0.2\% \\ P(\text{Neg}\,|\,\overline{\text{HIV}}) &= 99.8\% \end{aligned}$$

We miss something: $P_\circ(\text{HIV})$ and $P(\overline{\text{HIV}})$: Yes! We need some input from our best knowledge of the problem. Let us take $P_\circ(\text{HIV}) = 1/600$ and $P(\overline{\text{HIV}}) \approx 1$ (the result is rather stable against *reasonable* variations of the inputs!)

$$\begin{aligned} \frac{P(\text{HIV}\,|\,\text{Pos})}{P(\overline{\text{HIV}}\,|\,\text{Pos})} &= \frac{P(\text{Pos}\,|\,\text{HIV})}{P(\text{Pos}\,|\,\overline{\text{HIV}})} \cdot \frac{P_\circ(\text{HIV})}{P(\overline{\text{HIV}})} \\ &= \frac{\approx 1}{0.002} \times \frac{0.1/60}{\approx 1} = 500 \times \frac{1}{600} = \frac{1}{1.2} \end{aligned}$$

Odd ratios and Bayes factor

$$
\begin{aligned}
\frac{P(\mathsf{HIV}\,|\,\mathsf{Pos})}{P(\overline{\mathsf{HIV}}\,|\,\mathsf{Pos})} &= \frac{P(\mathsf{Pos}\,|\,\mathsf{HIV})}{P(\mathsf{Pos}\,|\,\overline{\mathsf{HIV}})} \cdot \frac{P_\circ(\mathsf{HIV})}{P(\overline{\mathsf{HIV}})} \\
&= \frac{\approx 1}{0.002} \times \frac{0.1/60}{\approx 1} = 500 \times \frac{1}{600} = \frac{1}{1.2} \\
\Rightarrow P(\mathsf{HIV}\,|\,\mathsf{Pos}) &= 45.5\%\,.
\end{aligned}
$$

Odd ratios and Bayes factor

$$
\begin{aligned}
\frac{P(\text{HIV}\,|\,\text{Pos})}{P(\overline{\text{HIV}}\,|\,\text{Pos})} &= \frac{P(\text{Pos}\,|\,\text{HIV})}{P(\text{Pos}\,|\,\overline{\text{HIV}})}\cdot\frac{P_\circ(\text{HIV})}{P(\overline{\text{HIV}})} \\
&= \frac{\approx 1}{0.002}\times\frac{0.1/60}{\approx 1} = 500\times\frac{1}{600} = \frac{1}{1.2} \\
\Rightarrow P(\text{HIV}\,|\,\text{Pos}) &= 45.5\%\,.
\end{aligned}
$$

There are some advantages in expressing Bayes theorem in terms of odd ratios:

- There is no need to consider all possible hypotheses (how can we be sure?)
  We just make a comparison of any couple of hypotheses!

Odd ratios and Bayes factor

$$\begin{aligned}
\frac{P(\text{HIV}\,|\,\text{Pos})}{P(\overline{\text{HIV}}\,|\,\text{Pos})} &= \frac{P(\text{Pos}\,|\,\text{HIV})}{P(\text{Pos}\,|\,\overline{\text{HIV}})}\cdot\frac{P_\circ(\text{HIV})}{P(\overline{\text{HIV}})}\\
&= \frac{\approx 1}{0.002}\times\frac{0.1/60}{\approx 1} = 500\times\frac{1}{600} = \frac{1}{1.2}\\
\Rightarrow P(\text{HIV}\,|\,\text{Pos}) &= 45.5\%\,.
\end{aligned}$$

There are some advantages in expressing Bayes theorem in terms of odd ratios:

- There is no need to consider all possible hypotheses (how can we be sure?)
  We just make a comparison of any couple of hypotheses!
- Bayes factor is usually much more inter-subjective, and it is often considered an 'objective' way to report how much the data favor each hypothesis.

The hidden uniform

What was the mistake of people saying $P(\overline{\text{HIV}}\,|\,\text{Pos}) = 0.2$?

We can easily check that this is due to have set $\frac{P_\circ(\text{HIV})}{P(\overline{\text{HIV}})} = 1$, that, hopefully, does not apply for a randomly selected Italian.

- This is typical in arbitrary inversions, and often also in frequentistic prescriptions that are used by the practitioners to form their confidence on something:

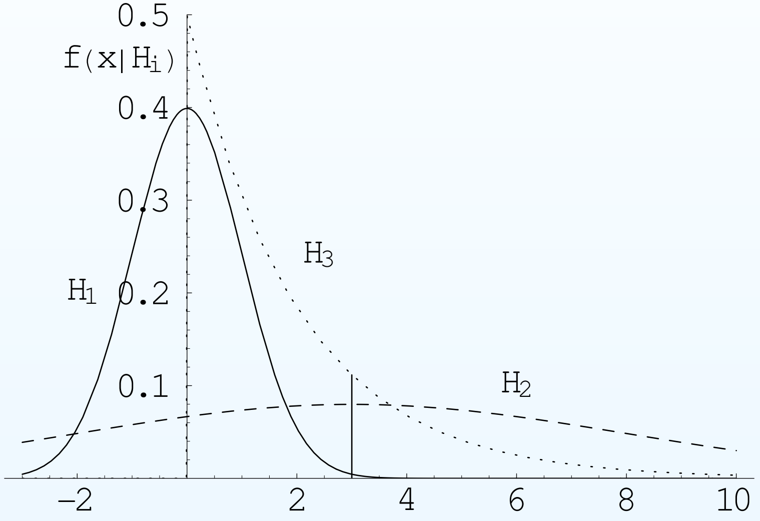
→ “absence of priors” means in most times uniform priors over the all possible hypotheses

- but they criticize the Bayesian approach because it takes into account priors explicitly !

Better methods based on ‘sand’ than methods based on nothing!

The three models example

Choose among H_1, H_2 and H_3 having observed $x = 3$:



In case of 'likelihoods' given by pdf's, the same formulae apply: "$P(\text{data}\,|\,H_j)$" $\longleftrightarrow$ "$f(\text{data}\,|\,H_j)$".

$$BF_{j,k} = \frac{f(x\!=\!3\,|\,H_j)}{f(x\!=\!3\,|\,H_k)}$$

$BF_{2,1} = 18$, $BF_{3,1} = 25$ and $BF_{3,2} = 1.4$ → data favor model H_3 (as we can see from figure!), but if we want to state how much we believe to each model we need to 'filter' them with priors.

Assuming the three models initially equally likely, we get final probabilities of 2.3%, 41% and 57% for the three models.

Comparing ‘complex’ hypotheses

In the case of ‘simple hypotheses’, i.e. hypotheses that do not contain free parameters, that was all!

Complex hypotheses require some more thinking:

Let us consider, for example, models $\mathcal{M}_A$ characterized by n_A parameters $\boldsymbol{\alpha}$, and $\mathcal{M}_B$ with n_B parameters $\boldsymbol{\beta}$.

Bayes factor $\frac{P(\textit{Data}\,|\,\mathcal{M}_A, I)}{P(\textit{Data}\,|\,\mathcal{M}_B, I)}$, but which set of parameters do we have to choose for the comparison?

- The ‘best fit’ one? NO! This would be correct if the models came with that fixed set of parameters!
- We have to take into account all possible sets of parameters that are that each model can take a priori. And probability theory teaches us how to do it:

$$P(\text{Data}\,|\,\mathcal{M}_A, I) = \int P(\text{Data}\,|\,\mathcal{M}_A, \boldsymbol{\alpha}, I)\, f_0(\boldsymbol{\alpha}\,|\,I)\, d\boldsymbol{\alpha}$$

(*An extension of the ‘decomposition rule’*)

Model dependence based on the integrated likelihood

Complex model Bayes factor:

$$\frac{P(\text{Data}\,|\,\mathcal{M}_A, I)}{P(\text{Data}\,|\,\mathcal{M}_B, I)} = \frac{\int P(\text{Data}\,|\,\mathcal{M}_A, \boldsymbol{\alpha}, I)\, f_0(\boldsymbol{\alpha}\,|\,I)\, d\boldsymbol{\alpha}}{\int P(\text{Data}\,|\,\mathcal{M}_B, \boldsymbol{\beta}, I)\, f_0(\boldsymbol{\beta}\,|\,I)\, d\boldsymbol{\beta}}$$

$$= \frac{\int \mathcal{L}_A(\boldsymbol{\alpha};\ \text{Data})\, f_0(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}}{\int \mathcal{L}_B(\boldsymbol{\beta};\ \text{Data})\, f_0(\boldsymbol{\beta})\, d\boldsymbol{\beta}}\,,$$

where $f_0(\boldsymbol{\alpha}\,|\,I)$ and $f_0(\boldsymbol{\beta}\,|\,I)$ are the parameter priors.

Model dependence based on the integrated likelihood

Complex model Bayes factor:

$$\frac{P(\text{Data}\,|\,\mathcal{M}_A, I)}{P(\text{Data}\,|\,\mathcal{M}_B, I)} = \frac{\int P(\text{Data}\,|\,\mathcal{M}_A, \boldsymbol{\alpha}, I)\, f_0(\boldsymbol{\alpha}\,|\,I)\, d\boldsymbol{\alpha}}{\int P(\text{Data}\,|\,\mathcal{M}_B, \boldsymbol{\beta}, I)\, f_0(\boldsymbol{\beta}\,|\,I)\, d\boldsymbol{\beta}} = \frac{\int \mathcal{L}_A(\boldsymbol{\alpha};\, \text{Data})\, f_0(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}}{\int \mathcal{L}_B(\boldsymbol{\beta};\, \text{Data})\, f_0(\boldsymbol{\beta})\, d\boldsymbol{\beta}}\,,$$

where $f_0(\boldsymbol{\alpha}\,|\,I)$ and $f_0(\boldsymbol{\beta}\,|\,I)$ are the parameter priors. The inference depends, then, on the *integrated likelihood* ("evidence")

$$\int \mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta};\, \text{Data})\, f_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,,$$

where $\mathcal{M}$ and $\boldsymbol{\theta}$ stand for the generic model and its parameters.

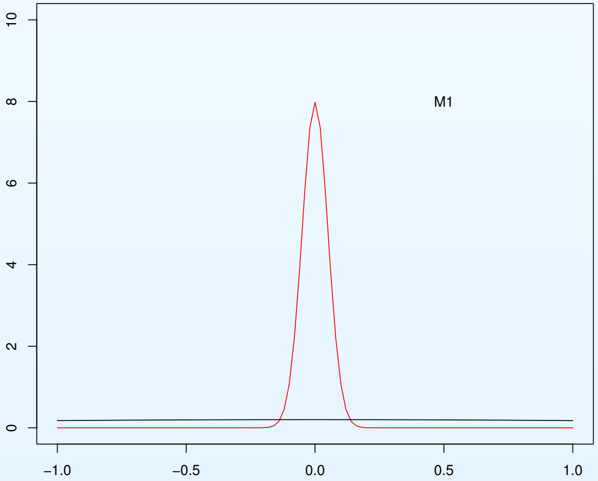
Automatic ‘Ockham Razor’

What enters the Bayes factor is the integrated likelihood:

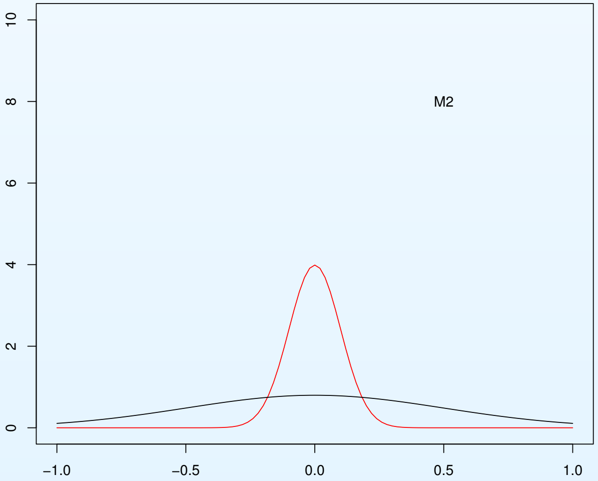
$$\int \mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta};\ \text{Data})\ f_0(\boldsymbol{\theta})\ d\boldsymbol{\theta}$$

Convolution of likelihood and parameter prior

Note: $\mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta};\ \text{Data})$ has, by definition, a maximum around the maximum likelihood point $\boldsymbol{\theta}_{ML}$,



Higher ML (→ ‘smaller χ^2’)



smaller ML (→ ‘larger χ^2’)

M_2 is ‘preferred from data’!

Automatic ‘Ockham razor’: Simpler models are preferred.

Further references

For an example of application of complex model comparison and resulting ‘Automatic Ockham Razor’, see e.g.

- P. Astone, S. D’Antonio and GdA, *“Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data”*, Class. Quant. Grav. 20 (2003) 769 [gr-qc/0304096].

Further references

For an example of application of complex model comparison and resulting ‘Automatic Ockham Razor’, see e.g.

- P. Astone, S. D’Antonio and GdA, *“Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data”*, Class. Quant. Grav. 20 (2003) 769 [gr-qc/0304096].

Or Google Bayesian Ockham razor . . .

Further references

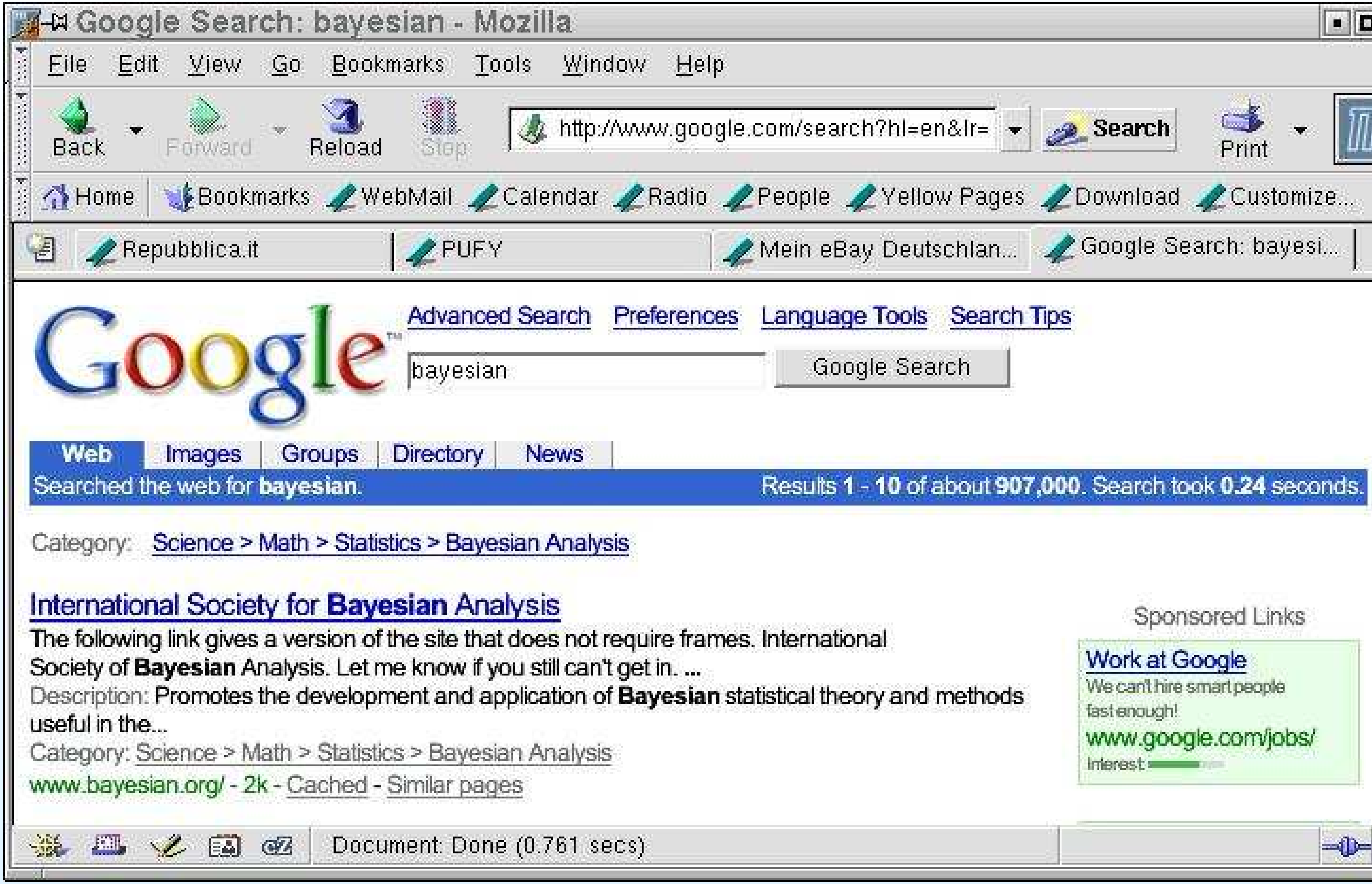
For an example of application of complex model comparison and resulting ‘Automatic Ockham Razor’, see e.g.

- P. Astone, S. D’Antonio and GdA, *“Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data”*, Class. Quant. Grav. 20 (2003) 769 [gr-qc/0304096].

Or Google Bayesian Ockham razor . . .

And, by the way, since you have already your browser open on Google, you might want to search for “Bayesian”, and find out that, contrary to most physicists, they are not much afraid of this word. . .

Are Bayesians ‘smart’ and ‘brilliant’?



Searched the web for **bayesian**. Results **1** - **10** of about **907,000**. Search took **0.24** seconds.

Category: Science > Math > Statistics > Bayesian Analysis

International Society for **Bayesian** Analysis

The following link gives a version of the site that does not require frames. International Society of **Bayesian** Analysis. Let me know if you still can't get in. **...**

Description: Promotes the development and application of **Bayesian** statistical theory and methods useful in the...

Category: Science > Math > Statistics > Bayesian Analysis

www.bayesian.org/ - 2k - Cached - Similar pages

Sponsored Links

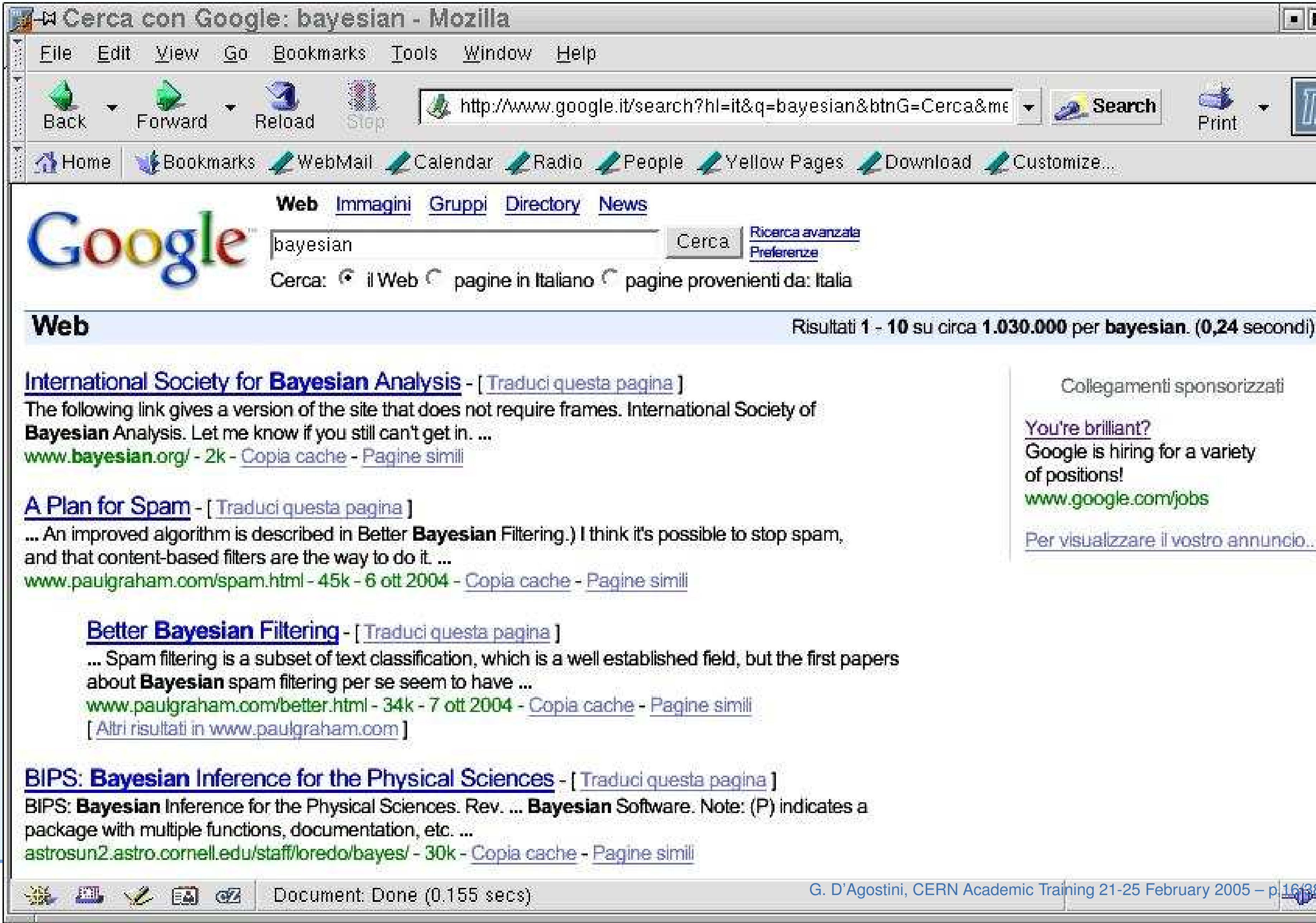
Work at Google

We can't hire smart people fast enough!

www.google.com/jobs/

Interest:

Are Bayesians ‘smart’ and ‘brilliant’?


Cerca con Google: bayesian - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
http://www.google.it/search?hl=it&q=bayesian&btnG=Cerca&me
Search
Print
Home
Bookmarks
WebMail
Calendar
Radio
People
Yellow Pages
Download
Customize...
Google
Web Immagini Gruppi Directory News
Cerca
Ricerca avanzata
Preferenze
Document: Done (0.155 secs)

bayesian

Cerca: ● il Web ○ pagine in Italiano ○ pagine provenienti da: Italia

Web Risultati **1** - **10** su circa **1.030.000** per **bayesian**. (**0,24** secondi)

International Society for **Bayesian** Analysis - [Traduci questa pagina]
The following link gives a version of the site that does not require frames. International Society of **Bayesian** Analysis. Let me know if you still can't get in. ...
www.**bayesian**.org/ - 2k - Copia cache - Pagine simili

A Plan for Spam - [Traduci questa pagina]
... An improved algorithm is described in Better **Bayesian** Filtering.) I think it's possible to stop spam, and that content-based filters are the way to do it. ...
www.paulgraham.com/spam.html - 45k - 6 ott 2004 - Copia cache - Pagine simili

Better **Bayesian** Filtering - [Traduci questa pagina]
... Spam filtering is a subset of text classification, which is a well established field, but the first papers about **Bayesian** spam filtering per se seem to have ...
www.paulgraham.com/better.html - 34k - 7 ott 2004 - Copia cache - Pagine simili
[Altri risultati in www.paulgraham.com]

BIPS: **Bayesian** Inference for the Physical Sciences - [Traduci questa pagina]
BIPS: **Bayesian** Inference for the Physical Sciences. Rev. ... **Bayesian** Software. Note: (P) indicates a package with multiple functions, documentation, etc. ...
astrosun2.astro.cornell.edu/staff/loredo/bayes/ - 30k - Copia cache - Pagine simili

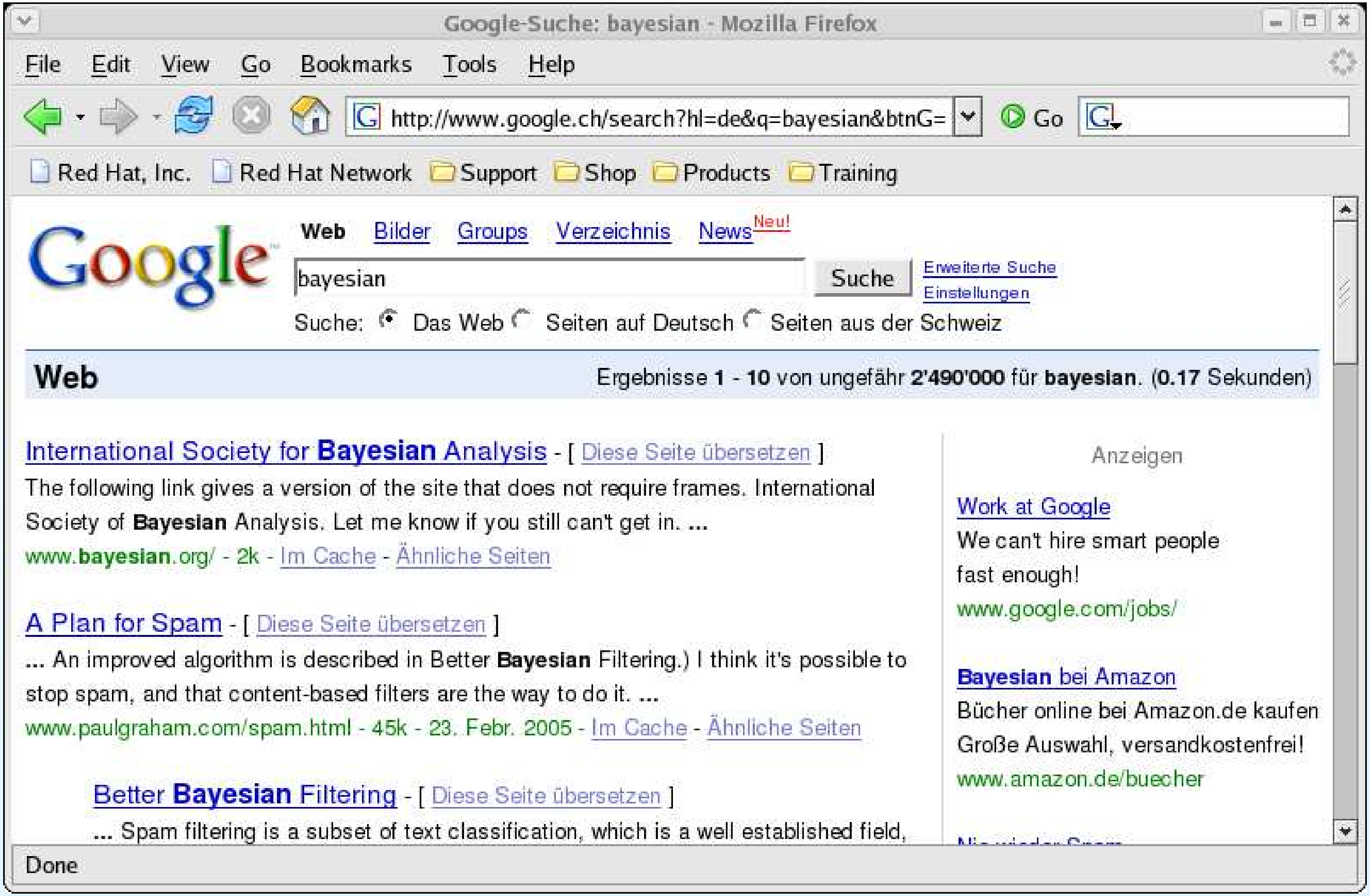
Collegamenti sponsorizzati

You're brilliant?
Google is hiring for a variety of positions!
www.google.com/jobs

Per visualizzare il vostro annuncio...

Are Bayesians ‘smart’ and ‘brilliant’?

Slide addedd after the lecture: search of 25 Feb 05, 13:01



Web Ergebnisse **1** - **10** von ungefähr **2'490'000** für **bayesian**. (**0.17** Sekunden)

International Society for **Bayesian** Analysis - [Diese Seite übersetzen]
The following link gives a version of the site that does not require frames. International Society of **Bayesian** Analysis. Let me know if you still can't get in. ...
www.**bayesian**.org/ - 2k - Im Cache - Ähnliche Seiten

A Plan for Spam - [Diese Seite übersetzen]
... An improved algorithm is described in Better **Bayesian** Filtering.) I think it's possible to stop spam, and that content-based filters are the way to do it. ...
www.paulgraham.com/spam.html - 45k - 23. Febr. 2005 - Im Cache - Ähnliche Seiten

Better **Bayesian** Filtering - [Diese Seite übersetzen]
... Spam filtering is a subset of text classification, which is a well established field,

Anzeigen

Work at Google
We can't hire smart people fast enough!
www.google.com/jobs/

Bayesian bei Amazon
Bücher online bei Amazon.de kaufen
Große Auswahl, versandkostenfrei!
www.amazon.de/buecher

Nie wieder Spam

A last remark

A last remark on model comparisons

- for a ‘serious’ probabilistic model comparisons, **at least two well defined models are needed**
- p-values (e.g. ‘χ^2 tests) have to be considered very useful starting points to understand if further investigation is worth [Yes, I also use χ^2 to get an idea of the “distance” between a model and the experimental data – but not more than that].
- But until you don’t have an alternative and credible model to explain the data, there is little to say about the “chance that the data come from the model”, unless the data are really impossible.
- Why do frequentistic test often work? → Slides

Parametric inference

→ Choose a model and infer its parameter(s).

Bayes theorem for continuous variables has following structure

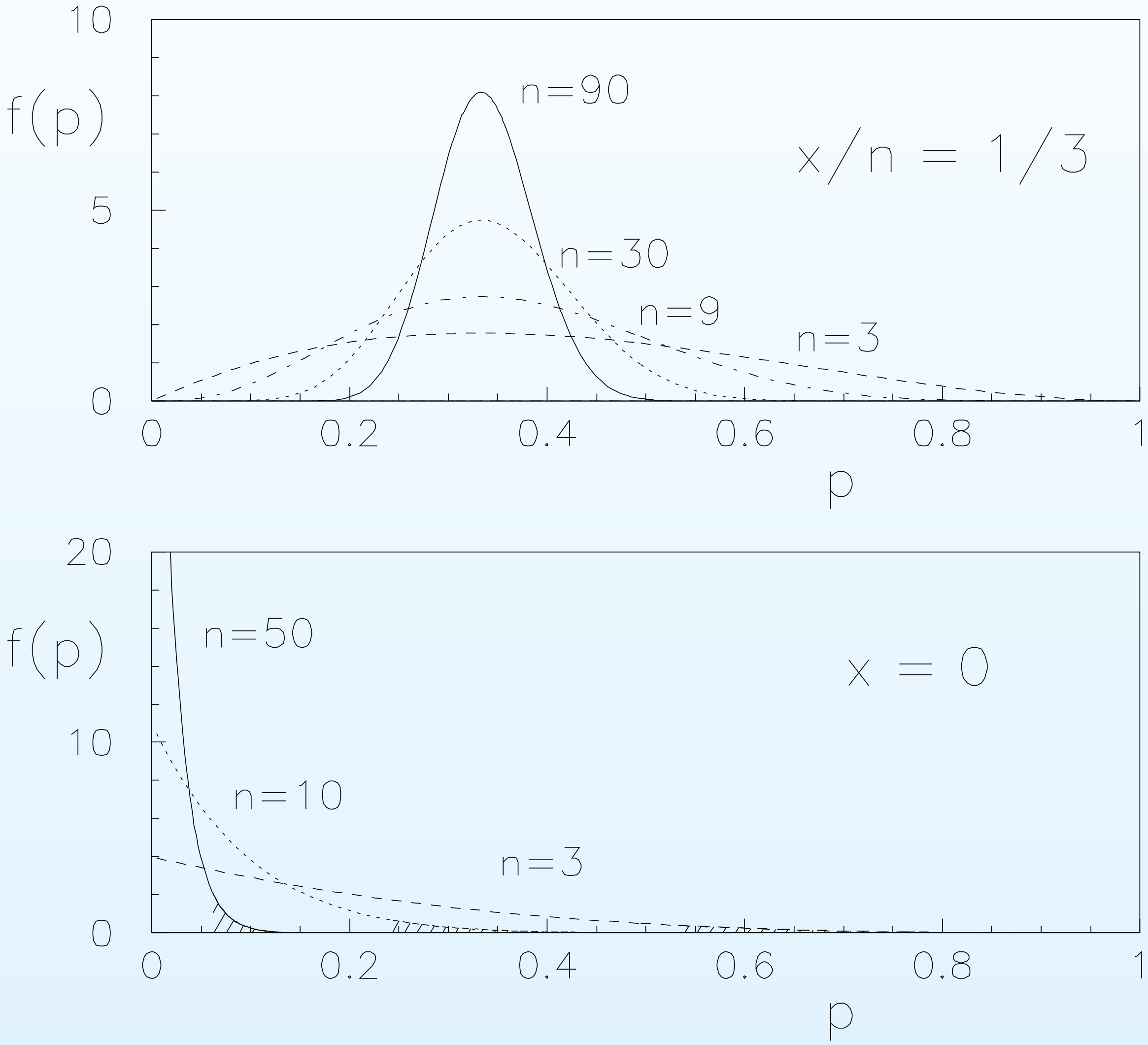
$$\boxed{f(\theta\,|\,\text{data}) \propto f(\text{data}\,|\,\theta)\, f_0(\theta)}$$

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

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

Inferring the Binomial p

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



Inferring the Binomial p

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

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

Laplace's rule of successions

Interpretation of E(p)

Interpretation of E(p). Think at 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. $P(E_i\,|\,p) = p$. (see comments on “physical probability” in lecture 3)

Interpretation of E(p)

Interpretation of E(p). Think at 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. $P(E_i \,|\, p) = p$. (see comments on “physical probability” in lecture 3)

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

Interpretation of E(p)

Interpretation of E(p). Think at 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. $P(E_i\,|\,p) = p$. (see comments on “physical probability” in lecture 3)

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

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

From relative frequencies to probabilities

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

$\boxed{\text{Laplace's rule of successions}}$

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

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

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

Further info about inferring p

$\rightarrow$ *"Inferring the success parameter p of a binomial model from small samples affected by background"*, physics/0412069.

Estimating Poisson λ

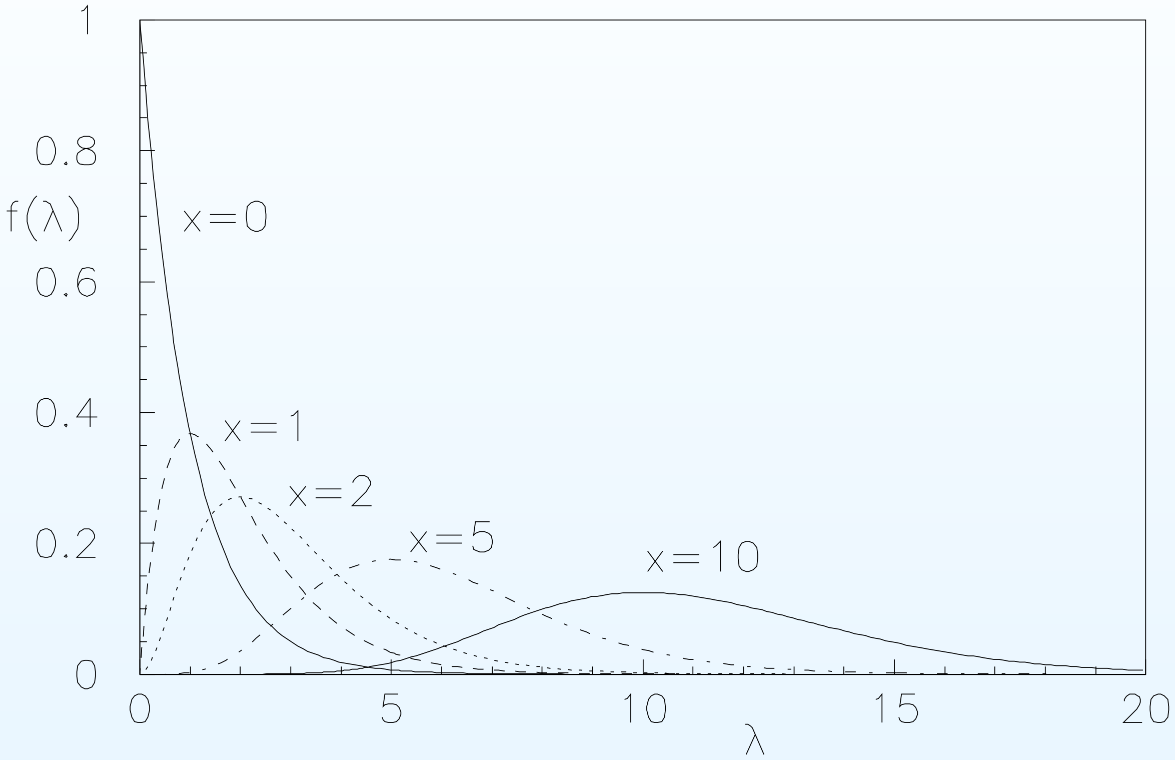
It becomes now an exercise, at least using a uniform prior on λ (not appropriate when searching for rare processes!)

$$
\begin{aligned}
f(\lambda\,|\,x,\mathcal{P}) &= \frac{\frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)}{\int_0^\infty \frac{\lambda^x\,e^{-\lambda}}{x!}\,f_\circ(\lambda)\,\mathrm{d}\lambda}\,. \\
f(\lambda\,|\,x,\mathcal{P}) &= \frac{\lambda^x\,e^{-\lambda}}{x!} \\
F(\lambda\,|\,x,\mathcal{P}) &= 1-e^{-\lambda}\left(\sum_{n=0}^{x}\frac{\lambda^n}{n!}\right)\,,
\end{aligned}
$$

Expected value, variance and mode of the probability distribution are

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

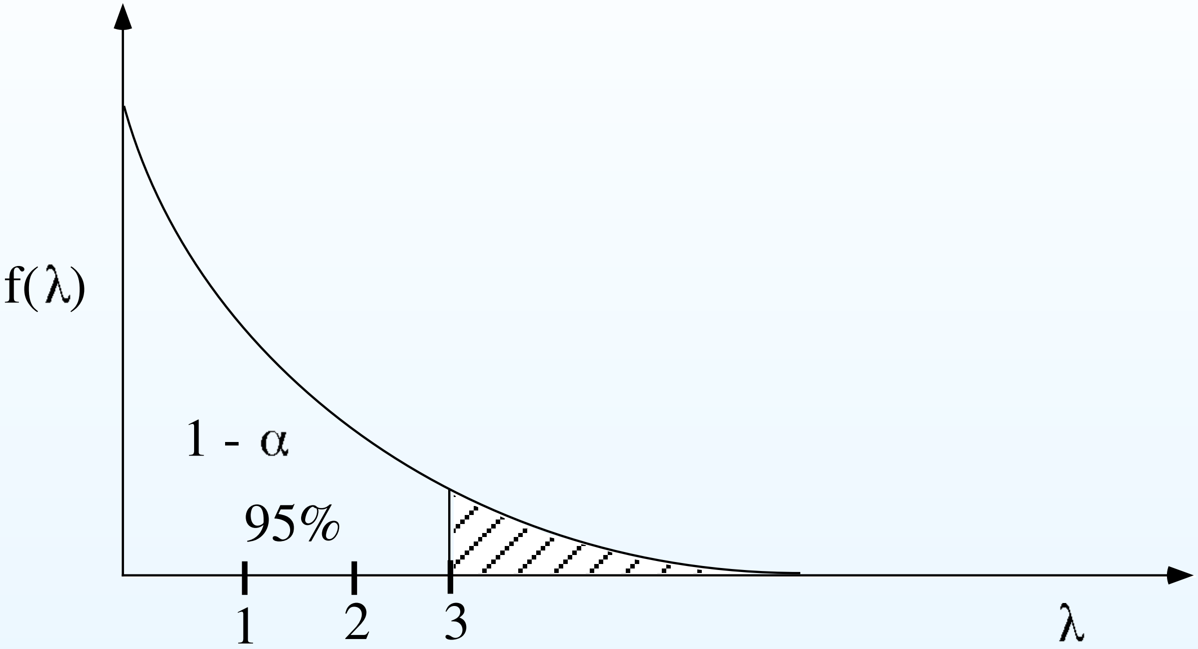
Some examples of $f(\lambda)$



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

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:

$$\begin{aligned}
f(\lambda_S \,|\, x, \lambda_{B_\circ}) &= \frac{e^{-(\lambda_{B_\circ}+\lambda_S)}\, (\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)}{\int_0^\infty e^{-(\lambda_{B_\circ}+\lambda_S)}\, (\lambda_{B_\circ}+\lambda_S)^x\, f_\circ(\lambda_S)\, d\lambda_S}\,. \\
f(\lambda_S \,|\, x, \lambda_{B_\circ}) &= \frac{e^{-\lambda_S}\, (\lambda_{B_\circ}+\lambda_S)^x}{x!\, \sum_{n=0}^{x} \frac{\lambda_{B_\circ}^n}{n!}}\,, \\
F(\lambda_S \,|\, x, \lambda_{B_\circ}) &= 1 - \frac{e^{-\lambda_S}\, \sum_{n=0}^{x} \frac{(\lambda_{B_\circ}+\lambda_S)^n}{n!}}{\sum_{n=0}^{x} \frac{\lambda_{B_\circ}^n}{n!}}\,.
\end{aligned}$$

(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)

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 is ‘the 68% thing‘. . .)

→ See slides:

- simple inference with very vague prior
- inference with ’narrow’ prior: → combinations
- predictive distributions
- measuring at the edge of the physical region
- introducing systematics

General probabilistic inference → simple fit formulae

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

→ 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 → 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)

Which prior for frontier physics?

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

→ 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~~ ⟶sensitivity bounds

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

→ 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:

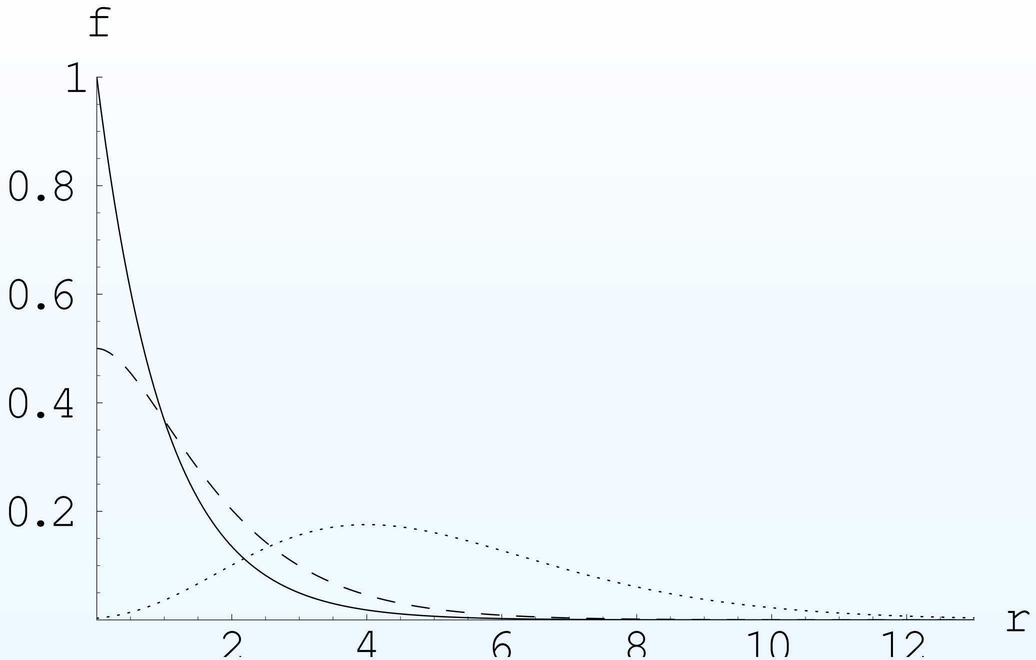
$$f(r\,|\,n_c, r_b) \propto \frac{e^{-(r+r_b)\,T}\,((r+r_b)\,T)^{n_c}}{n_c!}\,f_\circ(r)\,.$$

Uniform prior:

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

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

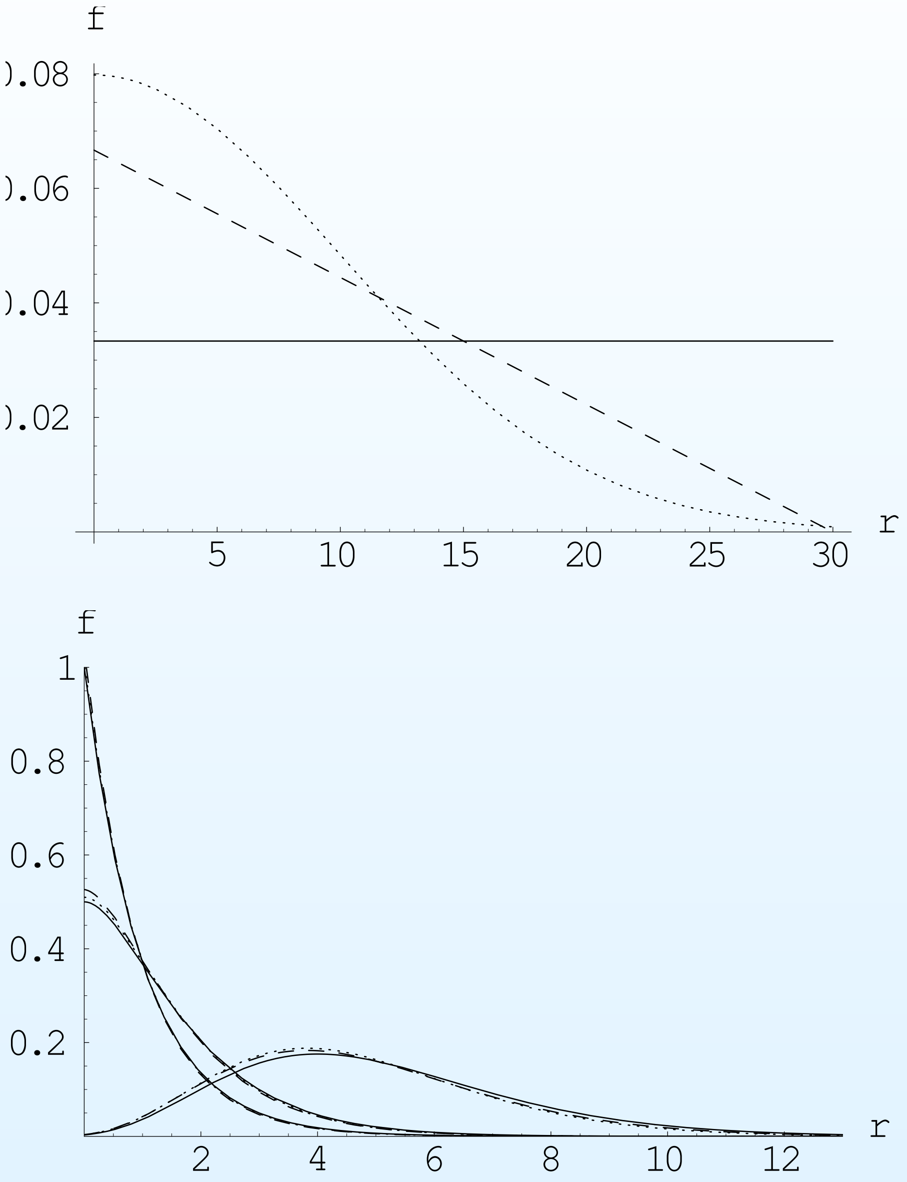
An example of inferring r



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).

→ which impression do you get? Do you see a serious problem?

Dependence for ‘optimistic priors’

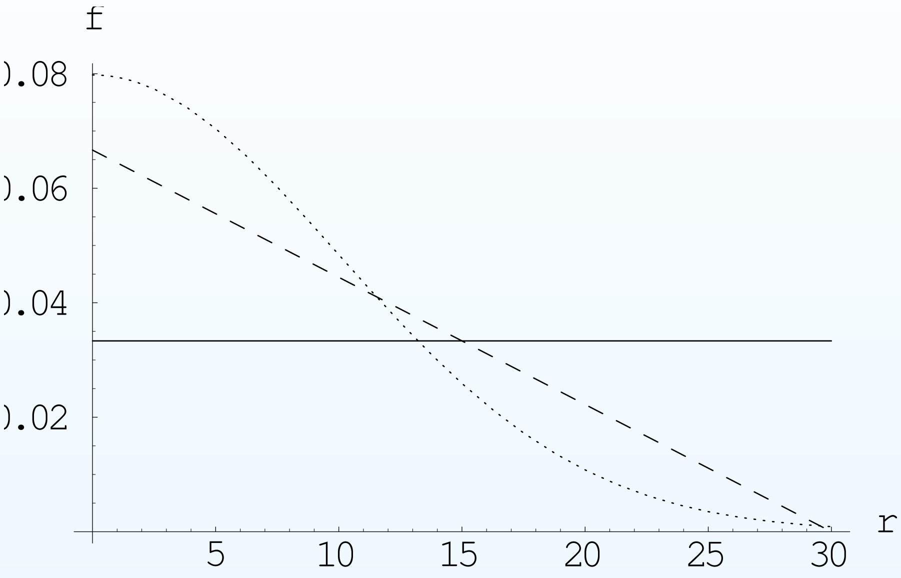


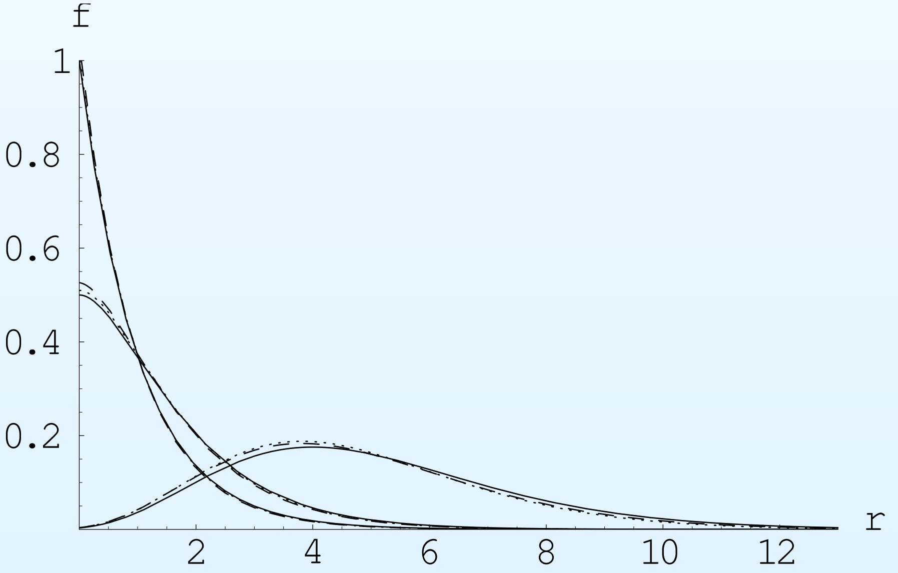
Upper plot shows some reasonable priors reflecting the positive attitude of researchers: little influence on posterior!

Dependence for ‘optimistic priors’

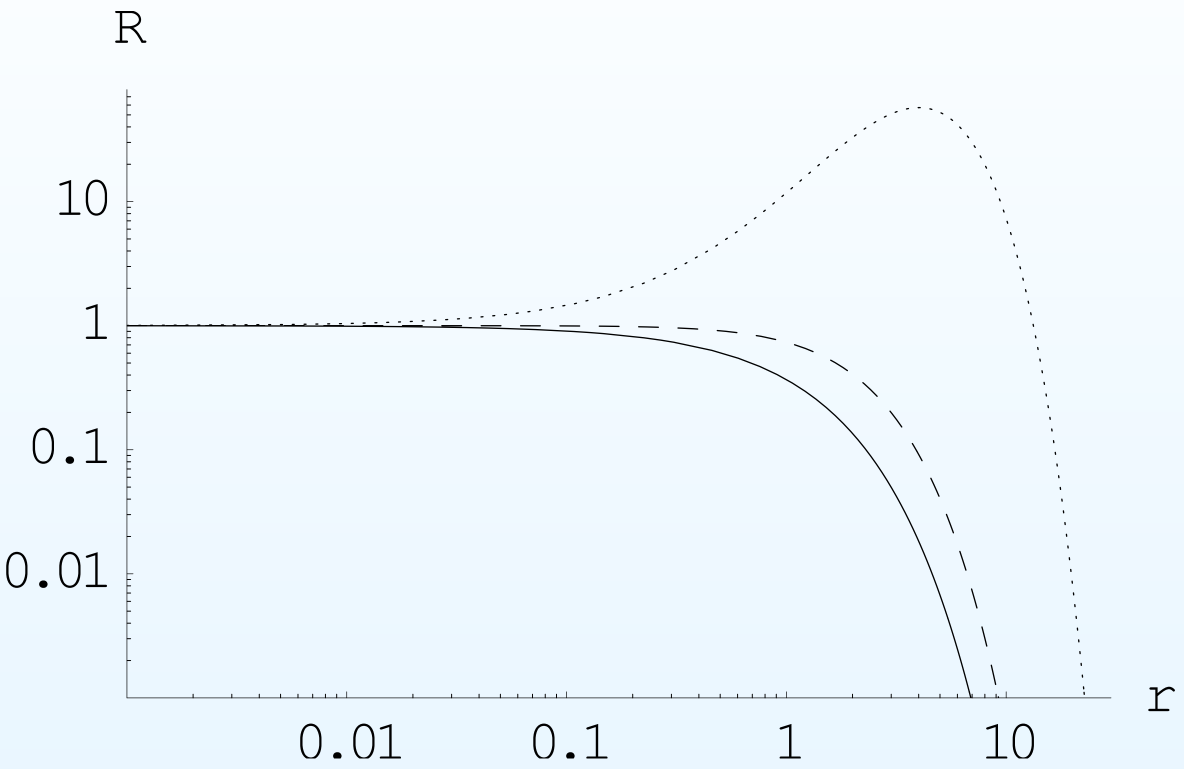
*Upper plot shows some reasonable priors reflecting the **positive attitude** of researchers: little influence on posterior!*

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!)





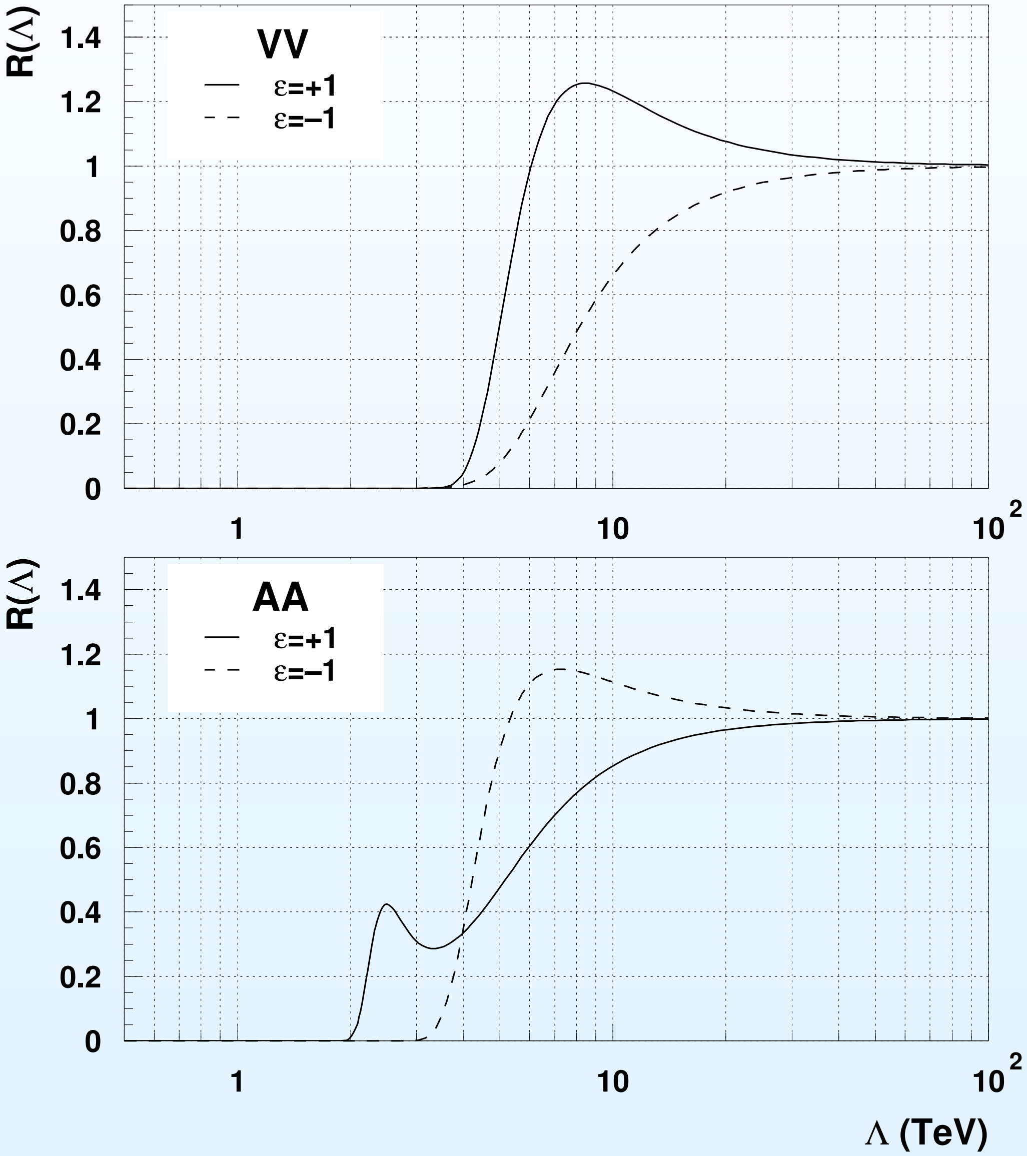
Rescaled likelihood (R function)



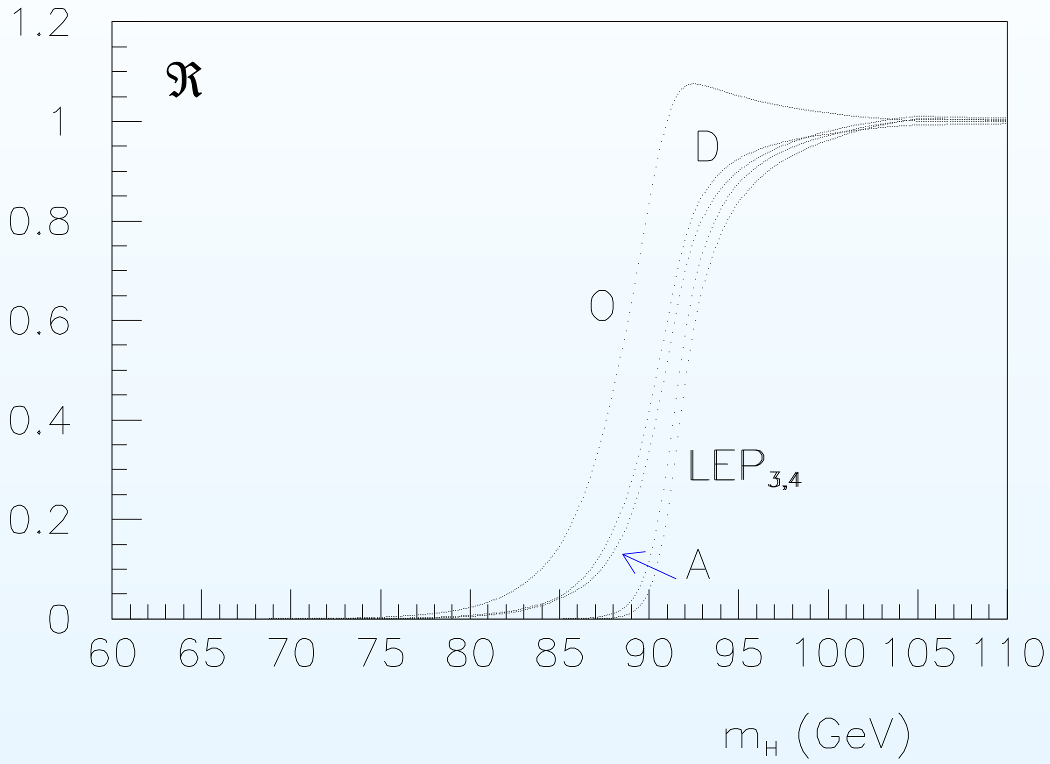
‘Relative belief updating ratio‘ $\mathcal{R}$ for the Poisson intensity parameter r for above cases. *Note log scales!*

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

→ 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~~ ⟶sensitivity bounds

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

→ PUBLISH LIKELIHOOD! (possibly in the rescaled form).

→ EASY COMBINATION OF RESULTS (independent likelihoods factorize).

My preferred conclusion

From the *ISO Guide on “the expression of uncertainty in measurement”*

“Although this Guide *provides a framework for assessing uncertainty, it cannot substitute for critical thinking, intellectual honesty, and professional skill. The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement. The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value.”*

My preferred conclusion

From the *ISO Guide on "the expression of uncertainty in measurement"*

"Although this Guide *provides a framework for assessing uncertainty, it cannot substitute for critical thinking, intellectual honesty, and professional skill. The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement. The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value."*

This is more or less how I interpret

Telling the truth with statistics

End of lecture

End of lecture 5

Transparencies written with $\LaTeX$
using fyma style of prosper class.
Thanks to the authors!