Probabilistic approach in High Energy Physics

discussions, examples, applications

Lorenzo Bellagamba

lorenzo.bellagamba@bo.infn.it

Istituto Nazionale di Fisica Nucleare
Bologna

Hadron Collider School Gottingen, Germany, 7-19 July 2013

Outline

- Confidence interval in classical statistics
- Inference, subjective probability and Bayes theorem
- Applications to discrete and continuos distributions
- Treatment of systematic uncertainties
- Montecarlo methods

Confidence interval (1)

$X \sim N(\mu,\sigma)$ random variable with unknown μ and known σ

n realization of X: $x_1,....x_n$

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

μ gaussian parameter

$\bar{x}$ mean estimation

Meaning of confidence level

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

$$m_{top} = (173.2 \pm 1)\ \mathrm{GeV} \xrightarrow{?} P(172.2GeV < m_{top} < 174.2GeV) = 0.68$$

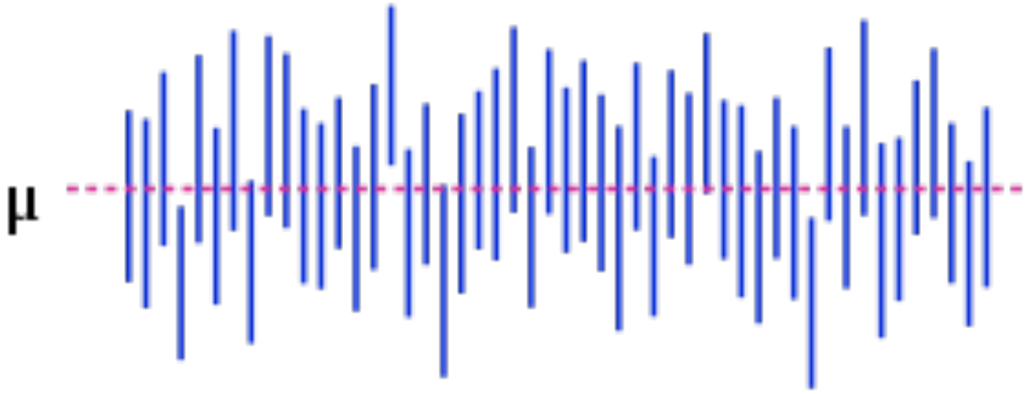
$$m_{Higgs} > 114GeV(95\%C.L.) \xrightarrow{?} P(m_{Higgs} > 114GeV) = 0.95$$

Confidence interval (2)

In classical statistics (frequentist) m_{top}, m_{Higgs}, have certain values even if unknown and it is not allowed to talk of probability for such quantities.

The limits of the confidence interval depend on the current estimation, but does not mean probability to find the true value within the interval. Such probability will be indeed 0 o 1 (the true value will be within or outside such interval).

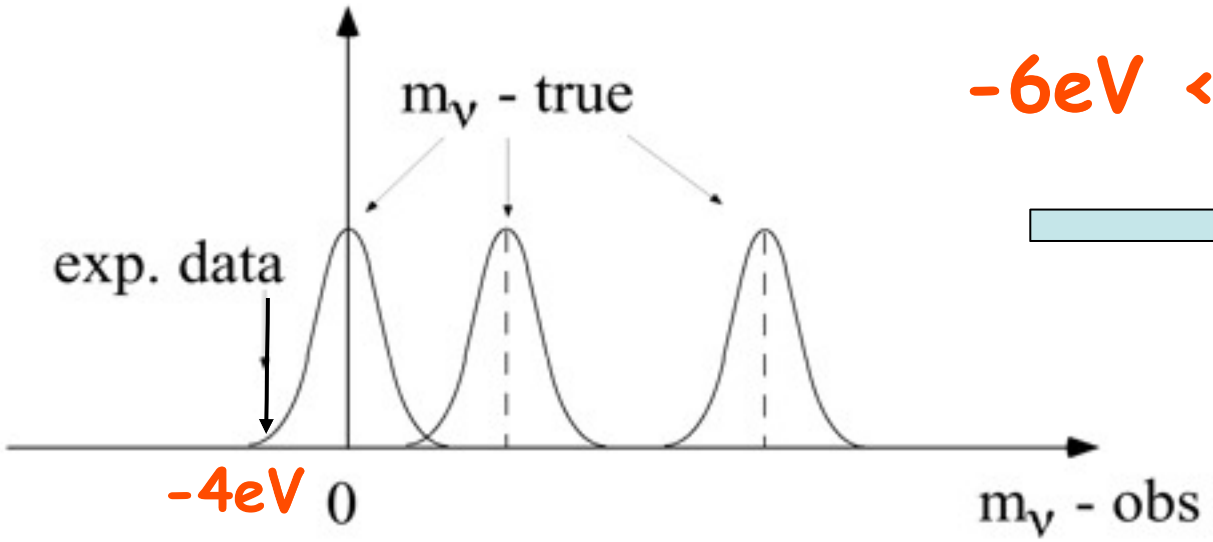
What we can say is that performing a long run of independent measurements under the same conditions of the same quantity, we'll get a series of different x%CL intervals which will include the true value x% of the times



Some interesting cases

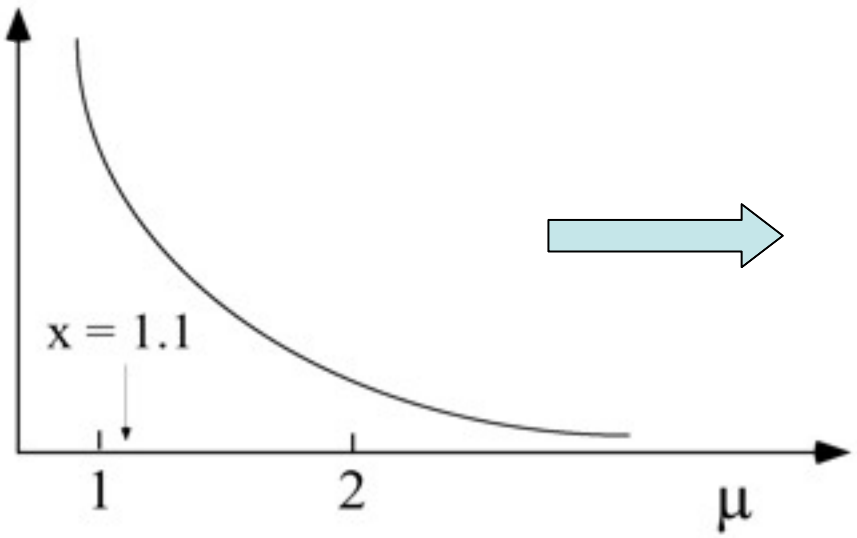
Negative mass ?

σ=2eV



$-6\text{eV} < m_\nu < -2\text{eV}$ 68% ??

$m_\nu < 0$ 98% ??



$P(\mu > 1.1)=P(\mu < 1.1)$??

Not uniform distribution of a physics variable

Probability of a hypothesis - Inference

$P(H|D)$

Probability of the hypothesis H given the observation D.
It is the quantity of interest.

$P(D|H)$

Probability of the observation D assuming correct the hypothesis H.
This is the quantity we can evaluate and it is a measurement of the “likelihood” of the observation in the framework described by the hypothesis H.

$$\boxed{P(H|D) \quad \not\leftrightarrow \quad P(D|H)}$$

Example test HIV

$P(\text{positive}|\ HIV) = 1.$ $\quad P(\text{positive}|\overline{HIV})=0.002$

Test on a randomly chosen person from the european population: infected ~1/600

$P(\overline{HIV}|\text{positive}) = ?$

Standard definition of probability

1- Ratio between the number of positive cases and all the possible cases (combinatorial)

2- Ratio between the number of times an event happens in a run and the total numbers of trials (frequentist)

Satisfactory definition?

Both incomplete:
1- It assumes implicitly that all possible cases have same probability (reuse of the probability concept in its definition)
2- The number of trials should be large (go to infinity), furthermore assumes implicitly that the process happened in the past and will happen in the future with the same probability.

They can be a good operative way to assign a value to a probability once the above conditions are satisfied, but as definitions they are not really satisfactory.

Subjective probability

Probability: Measurement of the degree of belief that an event happen

It could seem a vague and useless definition respect to the previous ones that at least give useful hints to evaluate it.
To better understand the definition let's consider a parallel with the bets and try to make the definition operative:

the larger the degree of belief that an event occur, more will be the quantity of money A that a better would pay to get back a money B in case of win

➡ Estimating the probability of an event means evaluating **p**=A/B such that it makes no difference for a rational better to bet in favor or against the event (coherent bet)

Of course the above considerations set constraints on **p**, since no rational people would bet A>B ➡ $0 \leq p \leq 1$.

The subjective probability definition, together with the coherence condition leads to the usual probability laws.

Inference and Bayes theorem

Let’s suppose to consider all possible different hypothesis H_i which can originate the event E. In this case the problem is the following: which is the probability of H_i assuming to have observed E?

$$P(H_i\,|\,E)P(E) = P(E\,|\,H_i)P(H_i)$$

→ $$\boxed{P(H_i\,|\,E) = \frac{P(E\,|\,H_i)P(H_i)}{P(E)}}$$

$\boxed{P(H_i)}$ → probability of H_i before the measurement (prior) → $P_0(H_i)$

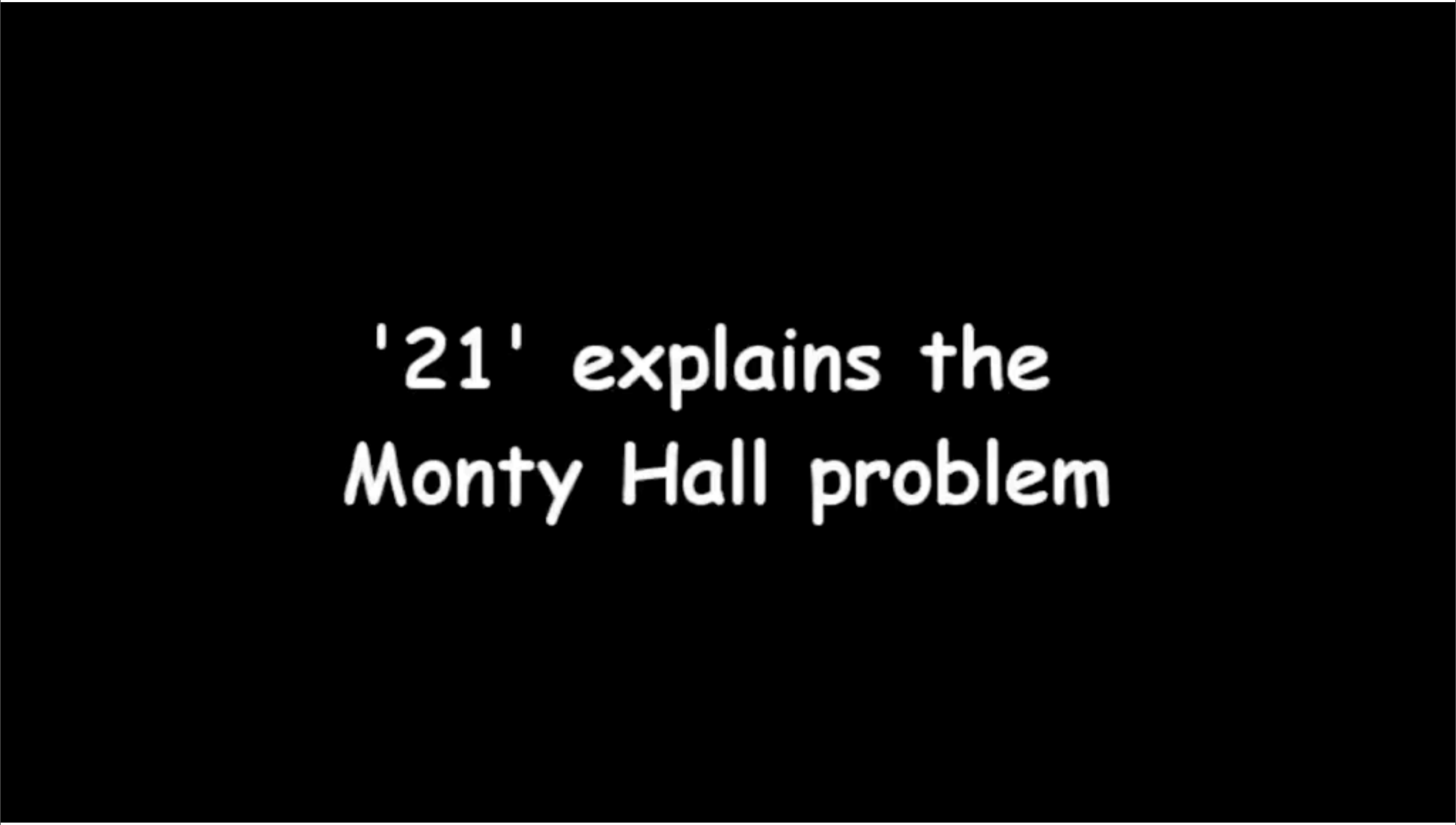
$\boxed{P(E\,|\,H_i)}$ → likelihood, new information gained by the observation of the event E

Mutual exclusivity $H_i \cap H_j = \emptyset,\ \forall i, j$ Exhaustivity $\bigcup_i H_i = \Omega$ →

$$\boxed{P(E)} \longrightarrow \begin{aligned} P(E\cap\Omega)] &= P\left(\bigcup_i (E\cap H_i)\right) \\ &= \sum_i P(E\cap H_i) \\ &= \sum_i P(E\,|\,H_i)P(H_i) \end{aligned}$$

$$\boxed{P(H_i|E) = \frac{P(E|H_i)P_0(H_i)}{\sum_j P(E|H_j)P_0(H_j)}}$$

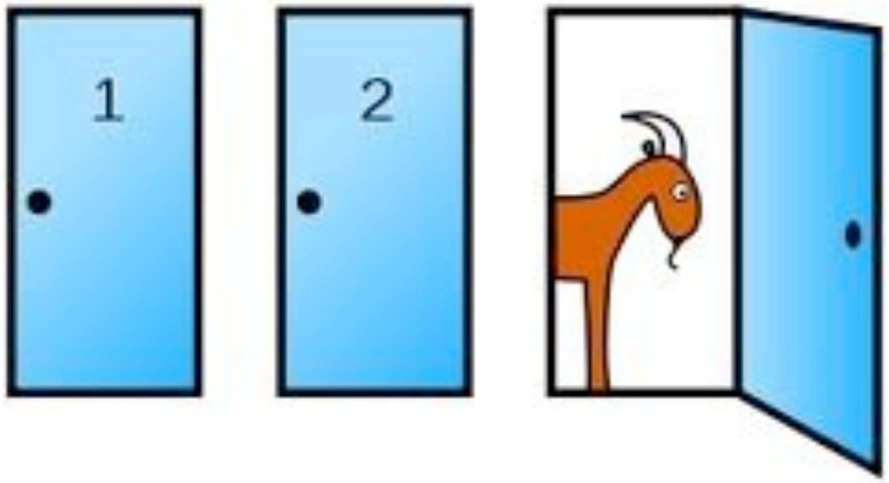
go back to the HIV problem →



'21' explains the Monty Hall problem

The Monty Hall problem

Brief history:
based on the American television game show “Let’s Make a Deal” (started in the sixties) and named after the shows original host.
A well-known statement of the problem was published in Marylin vos Savant's "Ask Marilyn" column in Parade magazine in 1990.
Many readers refused to believe that switching is beneficial. After the Monty Hall problem appeared in Parade, approximately 10000 readers, including nearly 1000 with PhDs, wrote to the magazine claiming that vos Savant was wrong.

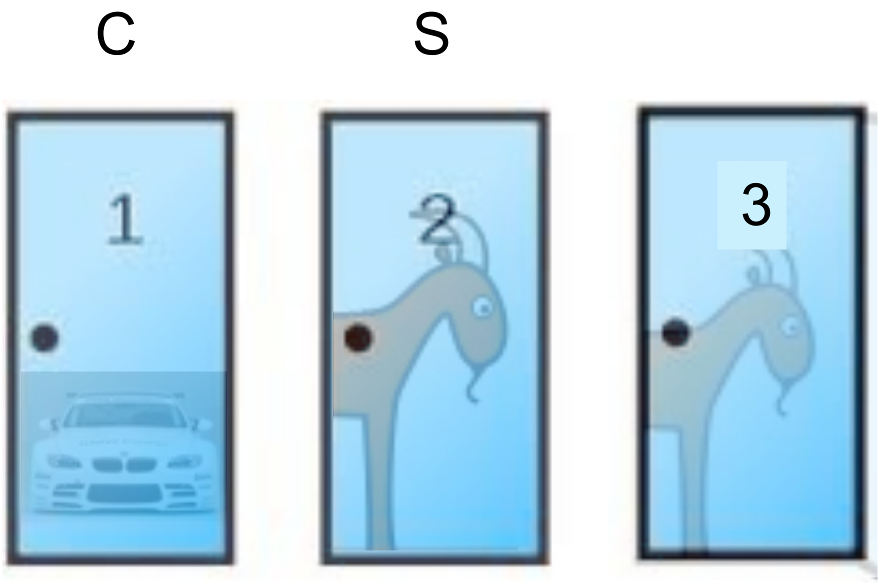


The bayesian approach to the Monty Hall problem

C: door hiding the car

S: door selected by the player

H: door opened by the host



$$P(C=c)=1/3 \quad c=\{1,2,3\}$$

$$P(C=c|S=s)=P(C=c) \quad c=\{1,2,3\}\ s=\{1,2,3\}$$

Car position independent of the choice of the player

Version 1: MH knows and the contestant knows he knows

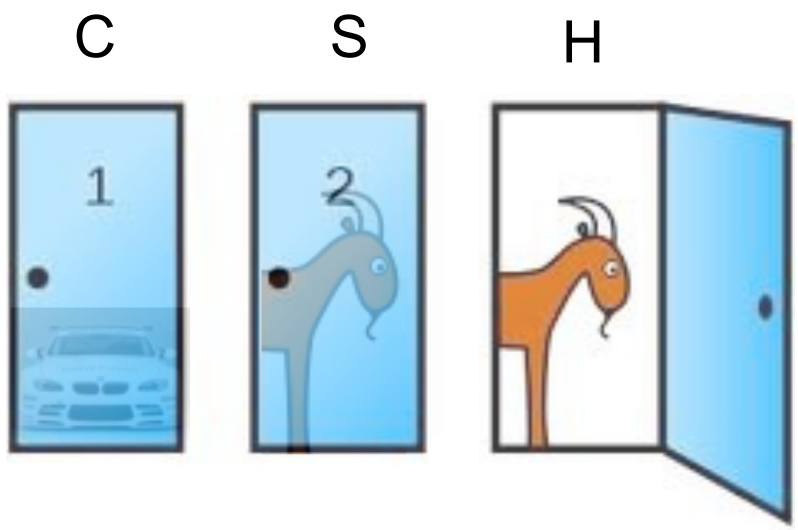
new information ⟶ the host opens one door with a gout

$$P(C=c|H=h,S=s)=\frac{P(H=h|C=c,S=s)P(C=c|S=s)}{P(H=h|S=s)}$$

$$P(H=h|S=s)=\sum_{c=1}^{3}P(H=h,C=c|S=s)=$$
$$\sum_{c=1}^{3}P(H=h|C=c,S=s)P(C=c|S=s)$$

$$P(H=h|C=c,S=s)=\begin{cases}0 & h=s\text{ (the host cannot open the door picked by the player)}\\ 0 & h=c\text{ (the host cannot open the door with the car)}\\ 1/2 & s=c\ h\neq s\\ 1 & s\neq c\ h\neq s\ h\neq c\end{cases}$$

Thus if the player initially chose door n. 2 and the host opens door n.3, the probability to win switching is:



$$P(C=1|H=3,S=2)=\frac{1\times\frac{1}{3}}{1\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}}=\frac{2}{3}$$

Version 2: MH doesn't know and the contestant knows he doesn't know

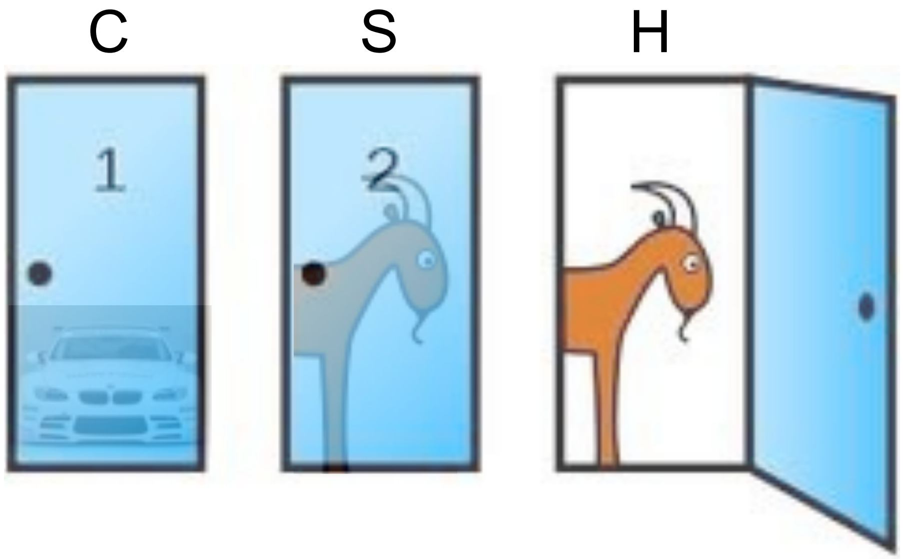
new information → the host opens one door with a gout

$$P(C=c|H=h,S=s)=\frac{P(H=h|C=c,S=s)P(C=c|S=s)}{P(H=h|S=s)}$$

$$P(H=h|S=s)=\sum_{c=1}^{3}P(H=h,C=c|S=s)=$$
$$\sum_{c=1}^{3}P(H=h|C=c,S=s)P(C=c|S=s)$$

$$P(H=h|C=c,S=s)=\begin{cases}0 & h=s \text{ (the host cannot open the door picked by the player)}\\ 1/3 & h=c \text{ (the host could open the door with the car)}\\ 1/2 & s=c\ h\neq s\\ 1/2 & s\neq c\ h\neq s\end{cases}$$

Thus if the player initially chose door n. 2 and the host opens door n.3, the probability to win switching is:



$$P(C=1|H=3,S=2)=\frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}}=\frac{1}{2}$$

Version 3: the contestant doesn't know if MH knows (general case)

A new variable k should be introduced as nuisance parameter: k={0,1}

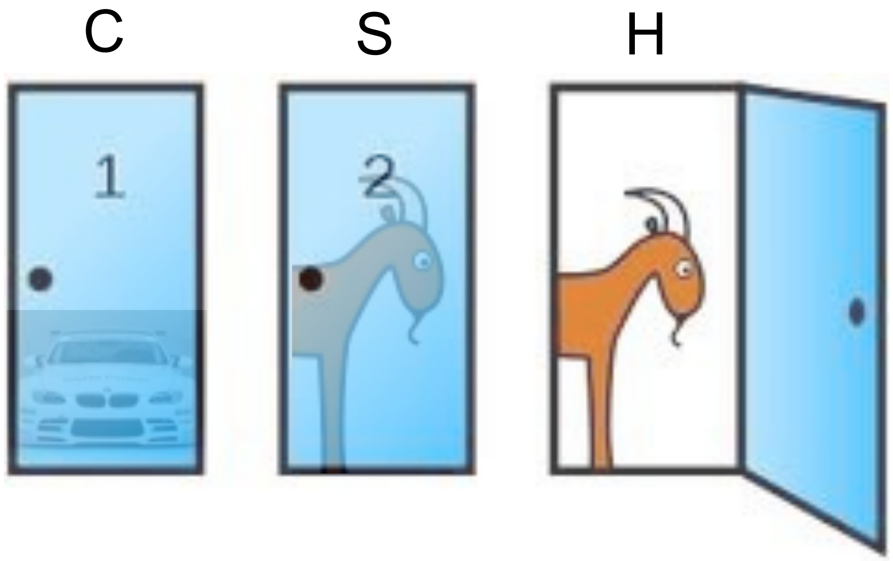
$$P(C=c|H=h,S=s)=\frac{\sum_{k=0}^{1}P(H{=}h|C{=}c,S{=}s,K{=}k)P(C{=}c|S{=}s)P(K{=}k)}{P(H{=}h|S{=}s)}$$

$$P(H=h|S=s)=\sum_{c=1}^{3}\sum_{k=0}^{1}P(H=h,C=c,K=k|S=s)=$$

$$\sum_{c=1}^{3}\sum_{k=0}^{1}P(H=h|C=c,K=k,S=s)P(C=c|S=s)P(K=k)$$

Let's use a flat prior for P(K)

$$P(K=k)=\begin{cases}1/2 & \text{k=0, MH doesn't know}\\ 1/2 & \text{k=1, MH knows}\end{cases}$$



Thus if the player initially chose door n. 2 and the host opens door n.3, the probability to win switching is:

$$P(C=1|H=3,S=2)=\frac{\frac{1}{2}\times 1\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times[1\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}]+\frac{1}{2}\times[\frac{1}{2}\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}]}=\frac{3}{5}$$

In summary

MH knows and player knows → P(switching)=2/3
MH doesn't know and player knows → P(switching)=1/2
player doesn't know if MH knows → P(switching)=3/5
(flat prior assumed)

What we learned from this "simple" problem ?

☑ **Also the most apparently harmless problem has its obscure sides:** ***"Statistics is subtle and even malicious."***, **but**

the Bayesian approach drives you directly to the solution, it is just needed some care to set correctly the problem.

☑ **fundamental role of the priors:**

the result depend on how the system is prepared or better from the informations on how it was prepared → priors, no way to avoid them

☑ **treatment of the nuisance parameters (systematic uncertainties)**

Example: Probability that a player be cheating after n consecutive winnings

$$P(W_n\mid H) = 2^{-n}$$

$$\begin{aligned} P(C\mid W_n) &= \frac{P(W_n\mid C)\cdot P_\circ(C)}{P(W_n\mid C)\cdot P_\circ(C) + P(W_n\mid H)\cdot P_\circ(H)} \\ &= \\ &= \frac{1\cdot P_\circ(C)}{1\cdot P_\circ(C) + 2^{-n}\cdot P_\circ(H)} \end{aligned}$$

n	$P(C\mid W_n)$ (%)	$P(H\mid W_n)$ (%)
0	5.0	95.0
1	9.5	90.5
2	17.4	82.6
3	29.4	70.6
4	45.7	54.3
5	62.7	37.3
6	77.1	22.9
...	...	...

Recursive formula: the prior is the posterior of the previous step

$$\begin{aligned} P(C\mid W_n) &= \frac{P(W\mid C)\cdot P(C\mid W_{n-1})}{P(W\mid C)\cdot P(C\mid W_{n-1}) + P(W\mid H)\cdot P(H\mid W_{n-1})} \\ &= \\ &= \frac{1\cdot P(C\mid W_{n-1})}{1\cdot P(C\mid W_{n-1}) + \frac{1}{2}\cdot P(H\mid W_{n-1})}\,, \end{aligned}$$

→ probability update as new data become available

Dependence from the prior →

$P_\circ(C)$	$P(C\mid W_n)$ (%)			
	$n=5$	$n=10$	$n=15$	$n=20$
1 %	24	91	99.7	99.99
5 %	63	98	99.94	99.998
50 %	97	99.90	99.997	99.9999

Example: Probability that a muon detector trigger be due to a true μ

- Trigger efficiency for a true μ = 0.95 → $P(T\,|\,\mu)$
- Probability of misidentification of a π = 0.02 → $P(T\,|\,\pi)$
- Sample composition: π 90% , μ 10%

$$P(\mu\,|\,T) \;=\; \frac{P(T\,|\,\mu)P_\circ(\mu)}{P(T\,|\,\mu)P_\circ(\mu)+P(T\,|\,\pi)P_\circ(\pi)}$$

Signal to noise ratio S/N:

$$\frac{S}{N} = \frac{P(S|T)}{P(N|T)} = \boxed{\frac{P(T|S)}{P(T|N)}} \times \frac{P_0(S)}{P_0(N)}$$

The likelihood ratio is the factor of improvement due to the new information

Gaussian variables

Suppose we want to estimate the parameter μ of a random variable gaussian distributed with a known standard deviation σ

Given a run of n_1 measurements, the mean value x_1 will be ~ N(μ,σ/√n_1)

$$p(x_1 \mid \mu,\sigma) : N(\mu,\sigma_1) \quad \sigma_1 = \sigma / \sqrt{n_1}$$

$$p(\mu \mid x_1, N(\mu,\sigma_1)) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)}{\int \frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\, f_\circ(\mu)\,\mathrm{d}\mu}$$

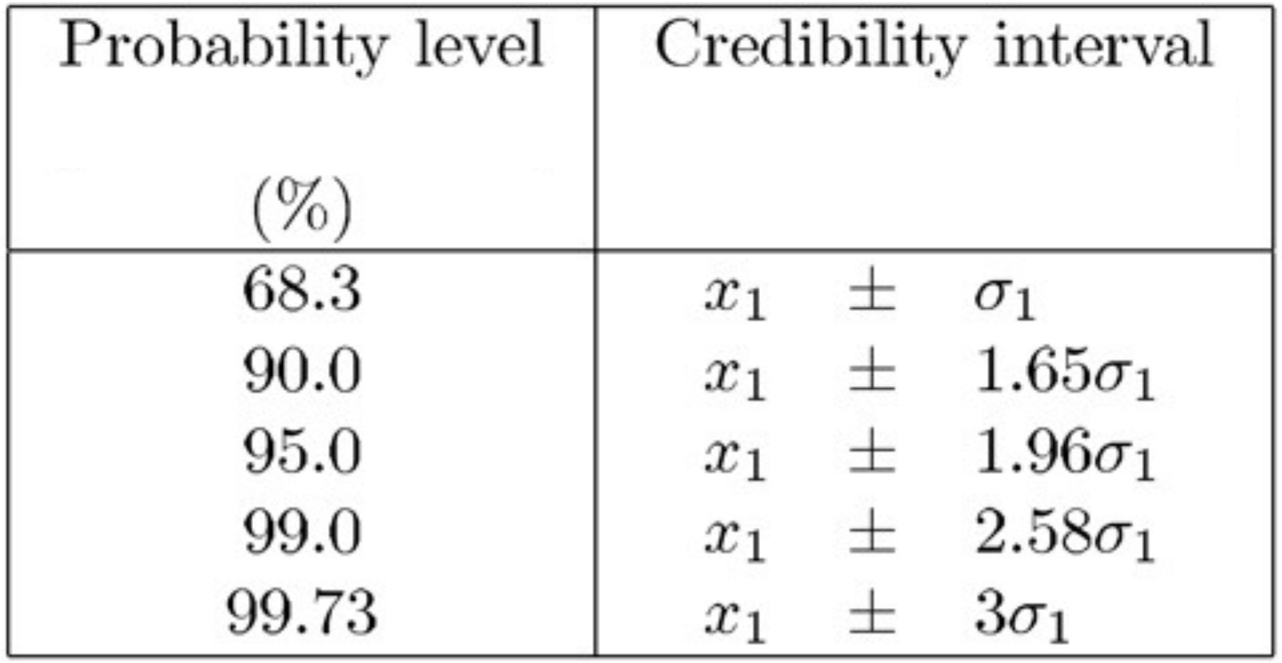
simplest choice flat prior $f_0 = \cos t$

N.B. enough to be constant for few σ_1s from x_1

$$\Rightarrow \quad p(\mu \mid x_1, N(\mu,\sigma_1)) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(x_1-\mu)^2}{2\,\sigma_1^2}}\,\mathrm{d}\mu} = \frac{1}{\sqrt{2\pi}\,\sigma_1}\, e^{-\frac{(\mu-x_1)^2}{2\,\sigma_1^2}}$$

Summarizing:

- The true value μ is gaussian distributed around x_1
- It's best estimation is $\mu = x_1$
- The "credibility" intervals are easily computable



Probability level	Credibility interval
(%)	
68.3	$x_1 \pm \sigma_1$
90.0	$x_1 \pm 1.65\sigma_1$
95.0	$x_1 \pm 1.96\sigma_1$
99.0	$x_1 \pm 2.58\sigma_1$
99.73	$x_1 \pm 3\sigma_1$

Combination of different measurements:

Suppose to analyze a second set of measurements:

Let’s use the results of the first set as prior:

$$p(x_2\,|\,\mu,\sigma):\ N(\mu,\sigma_2) \qquad \sigma_2 = \sigma/\sqrt{n_2}$$

$$f(\mu\,|\,x_1,\sigma_1,x_2,\sigma_2,\mathcal{N}) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_2}\,e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\,f(\mu\,|\,x_1,\mathcal{N}(\cdot,\sigma_1))}{\int \frac{1}{\sqrt{2\pi}\,\sigma_2}\,e^{-\frac{(x_2-\mu)^2}{2\,\sigma_2^2}}\,f(\mu\,|\,x_1,\mathcal{N}(\cdot,\sigma_1))\,\mathrm{d}\mu}$$

Final result:

$$f(\mu\,|\,x_1,\sigma_1,x_2,\sigma_2,\mathcal{N}) = \frac{1}{\sqrt{2\,\pi}\,\sigma_A}\,e^{-\frac{(\mu-x_A)^2}{2\,\sigma_A^2}}$$

$$\begin{aligned} x_A &= \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\,, \\ \frac{1}{\sigma_A^2} &= \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,. \end{aligned}$$

Where we found the usual formula of the weighted mean

Measurements close to the physics limit: neutrino mass

Experiment to measure the ν_e mass

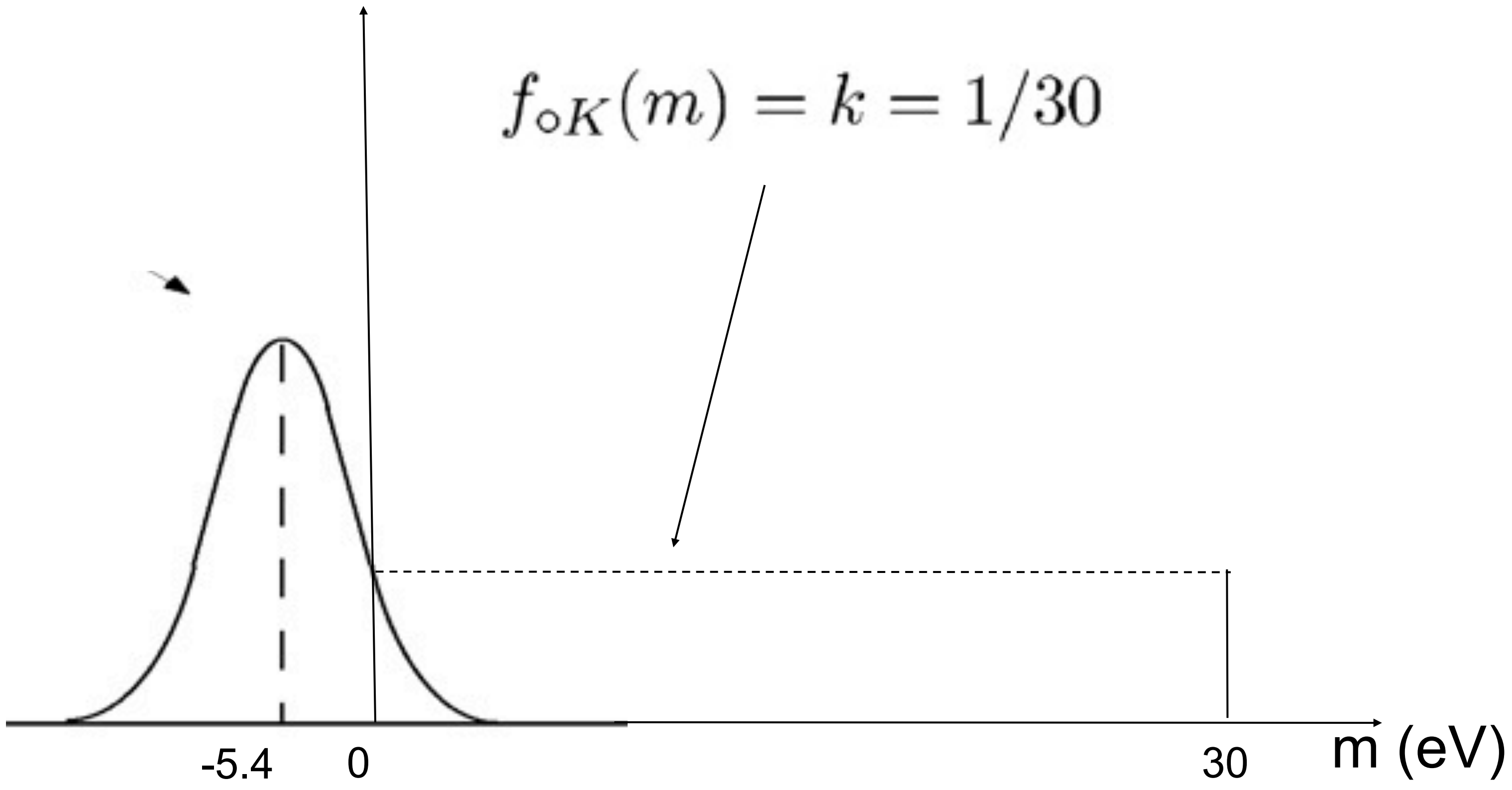
σ_m= 3.3 eV (independent of the mass)

Experimental result: m=-5.4 eV

What can we say about the ν_e mass ?

Prior: positive and not too large mass

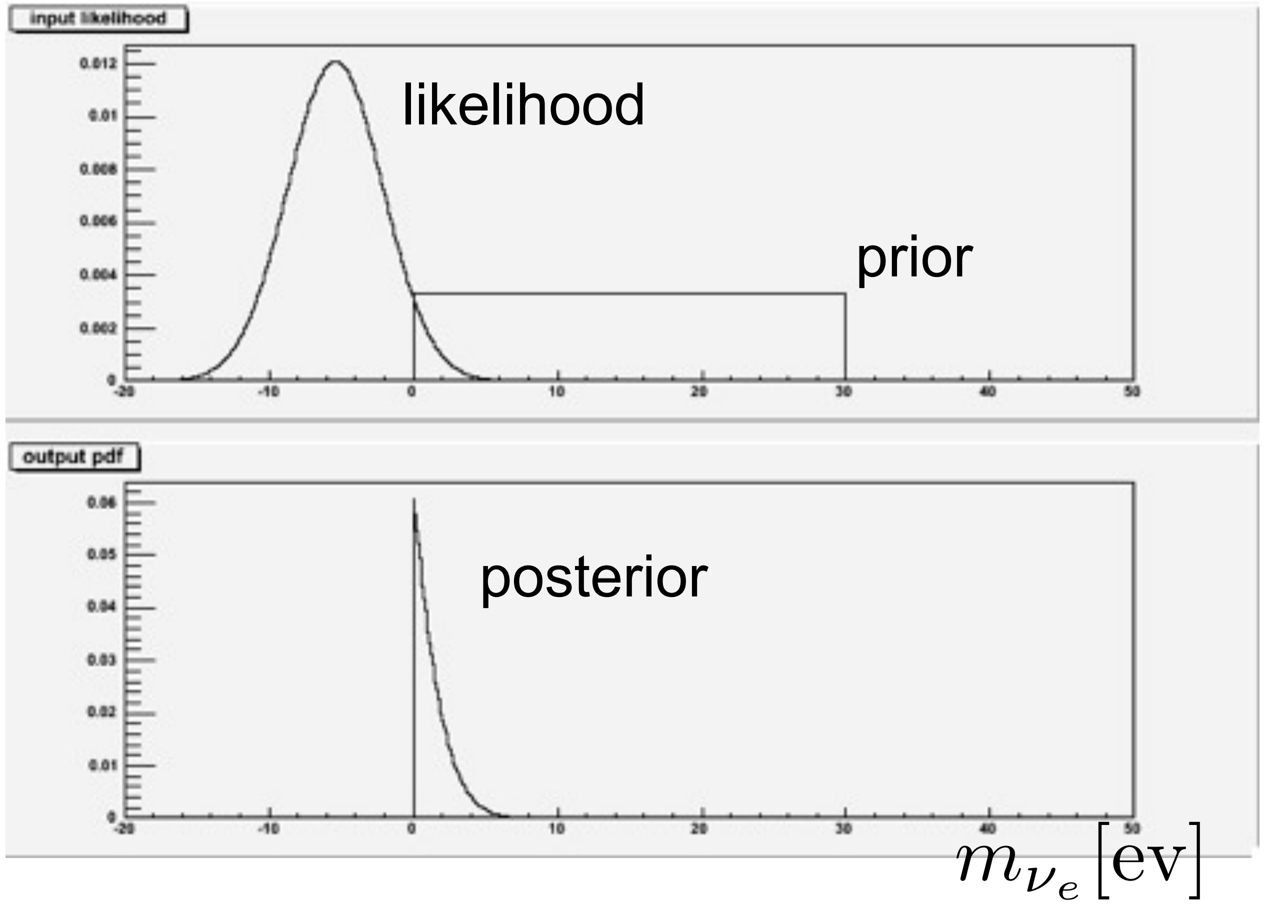
let's assume a constant mass in the range 0 ≤ m≤ 30 eV

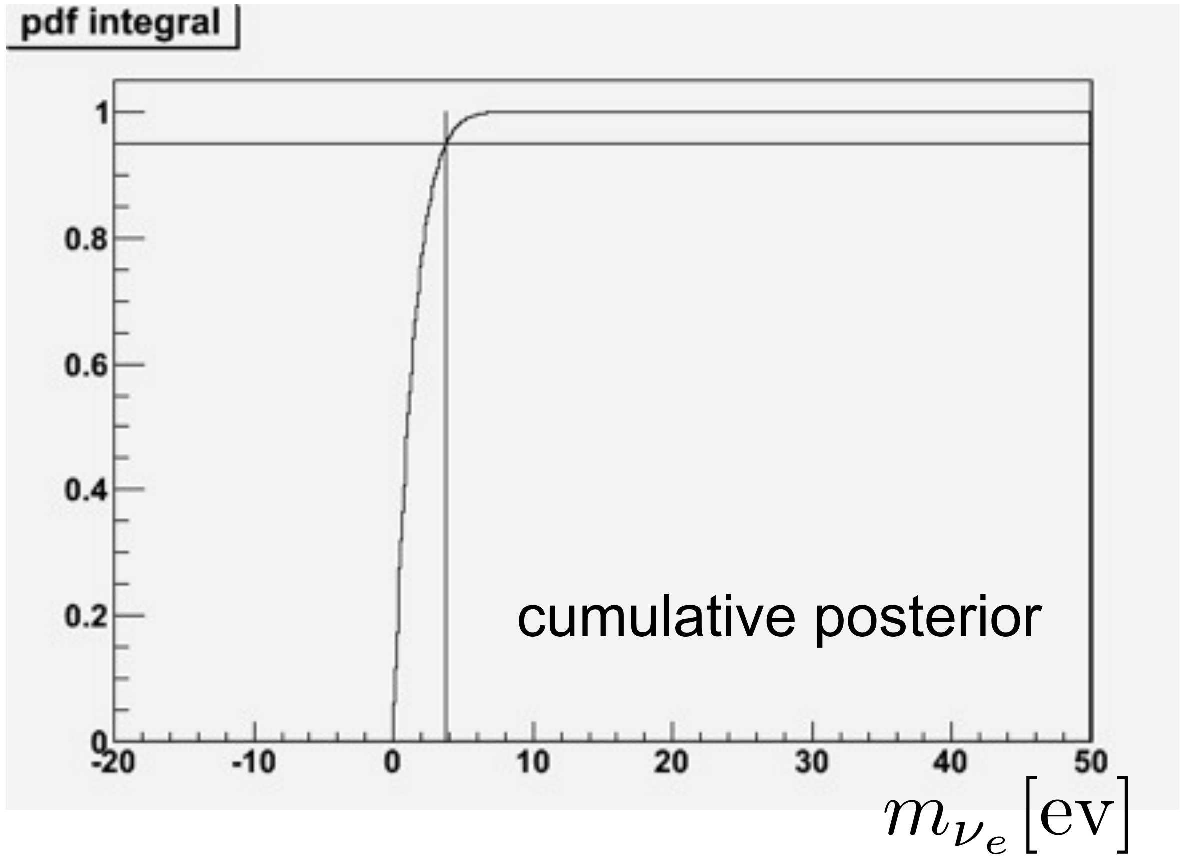


$$x = -5.4\ eV$$

$$p(m_{\nu_e}|x) = \frac{\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right]k}{\int_0^{30}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(m-x)^2}{2\,\sigma^2}\right]k\,\mathrm{d}\mu}$$

$$= \frac{20}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(m-x)^2}{2\sigma^2}\right] \qquad (0 \leq m \leq 30)$$





most probable value m=0

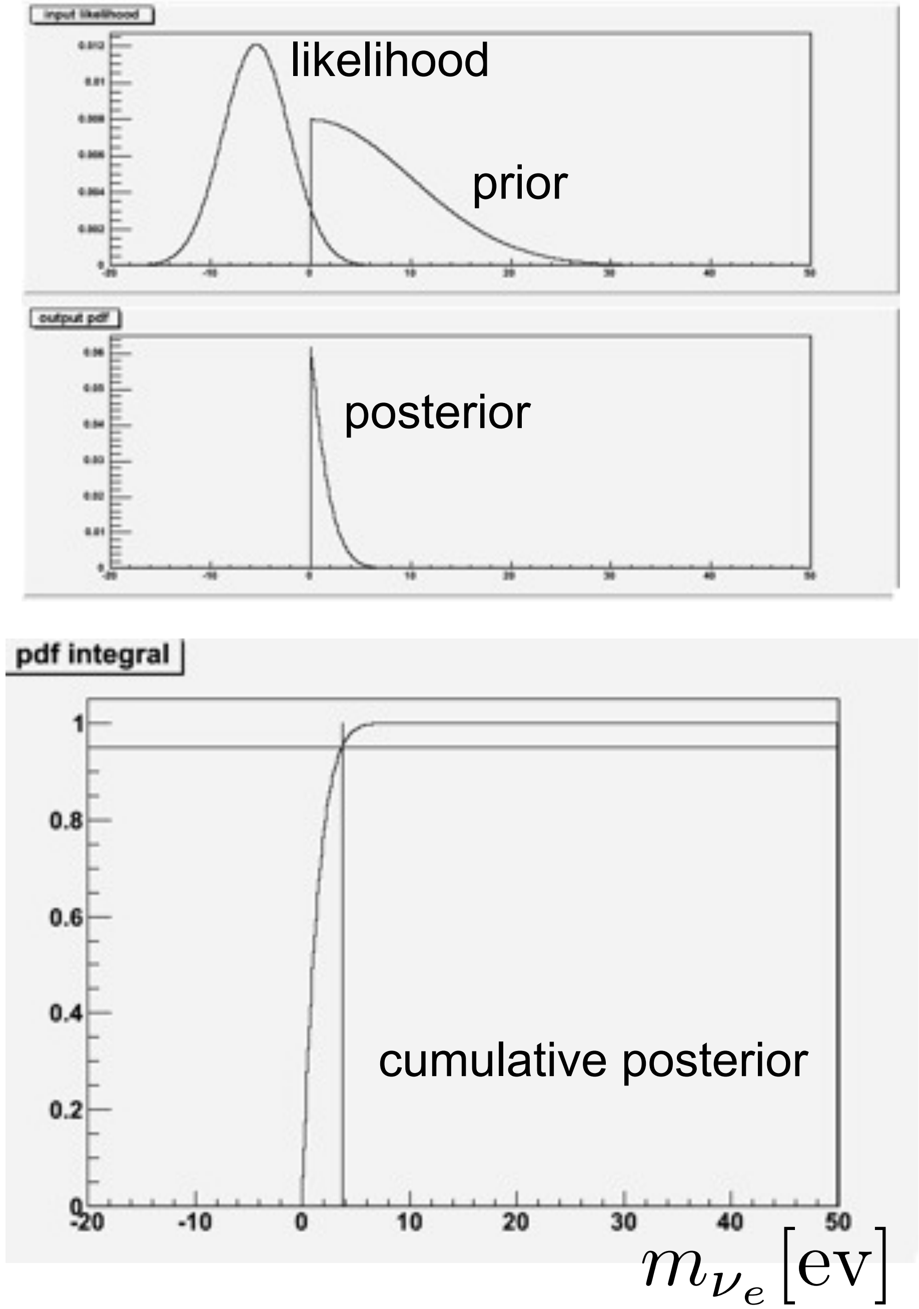
$$\int_0^{m_{95}} p(m_{\nu_e}|x)dm_{\nu_e}$$ → 95% CL m < 3.8 eV

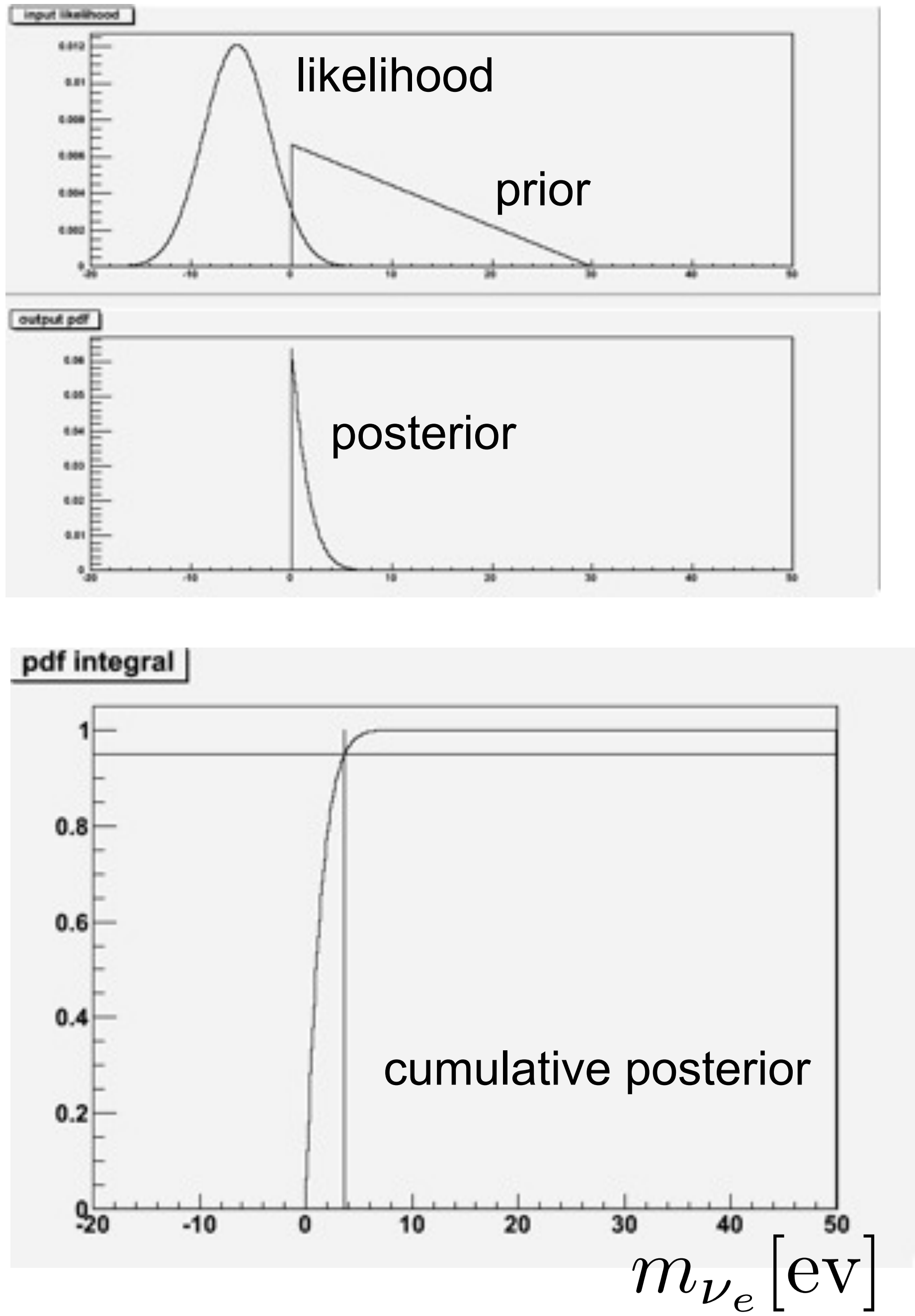
No large differences using different priors

gaussian $\longrightarrow$ $$f_{\circ N}(m) = \frac{2}{\sqrt{2\pi}\,\sigma_\circ}\exp\left[-\frac{m^2}{2\sigma_\circ^2}\right] \qquad (m \geq 0)$$

$$\sigma_0 = 10eV$$

triangular $\longrightarrow$ $$f_{\circ T}(m) = \frac{1}{450}(30 - x) \qquad (0 \leq m \leq 30)$$

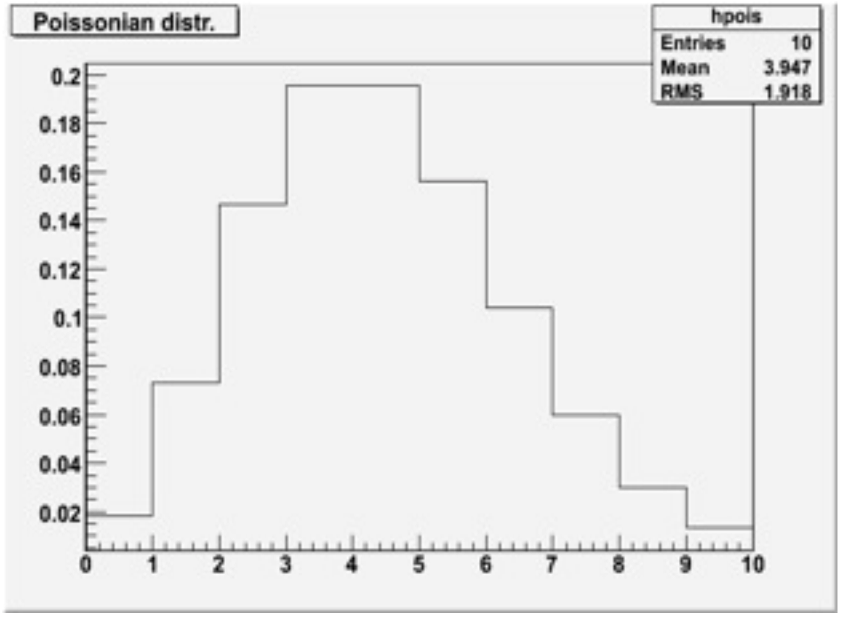




In both cases 95% CL m < 3.7 eV

Example: Poissonian variables

Probability of a production rate of a Beyond Standard Model process



N_{obs}

N_{obs}=5

N_{pred}=4.0

N_{sig}=$\sigma_{BSM}\cdot$lumi$\cdot$eff   lumi=100pb^{-1} eff=0.5

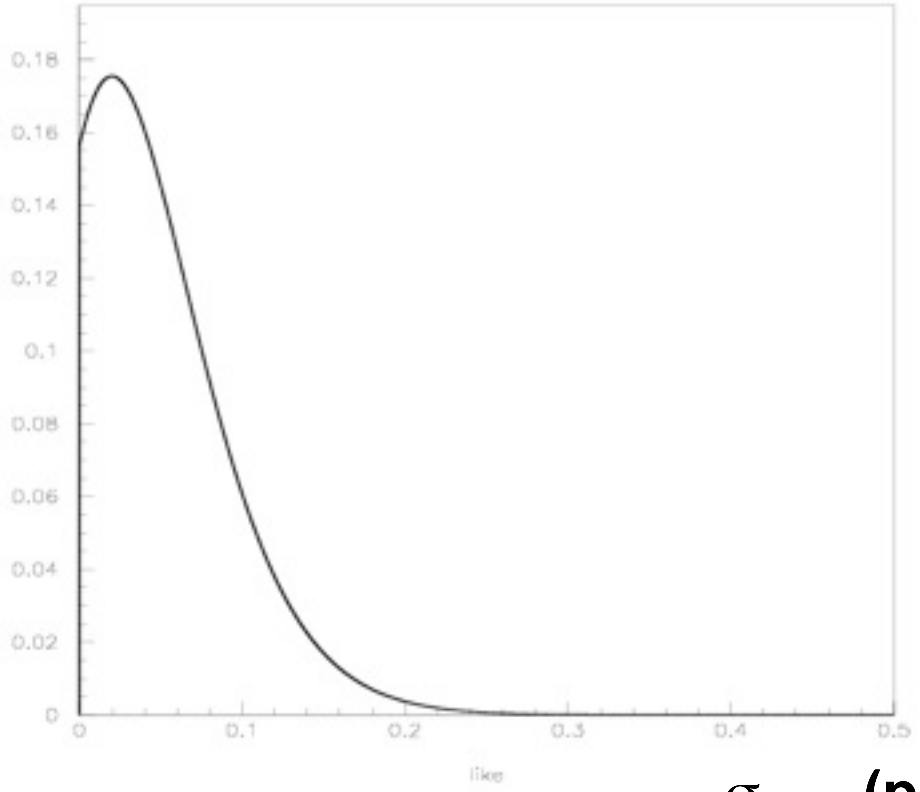
$$P(\sigma_{BSM}\,|\,N_{obs},N_{pred},Lumi,eff)=\frac{P(N_{obs}\,|\,\sigma_{BSM},N_{pred},Lumi,eff)\times P_0(\sigma_{BSM})}{\int_0^{\infty}P(N_{obs}\,|\,\sigma'_{BSM},N_{pred},Lumi,eff)\times P_0(\sigma'_{BSM})d\sigma'_{BSM}}$$

P

$$P(N_{obs}\,|\,\sigma_{BSM},N_{pred},Lumi,eff)=Poiss(N_{obs},\mu)$$

$$\mu=N_{pred}+N_{BSM}\qquad N_{BSM}=\sigma_{BSM}\times Lumi\times eff$$

Likelihood: $$\frac{\mu^{N_{OBS}}e^{-\mu}}{N_{obs}!}\longrightarrow$$



σ_{BSM} (pb)

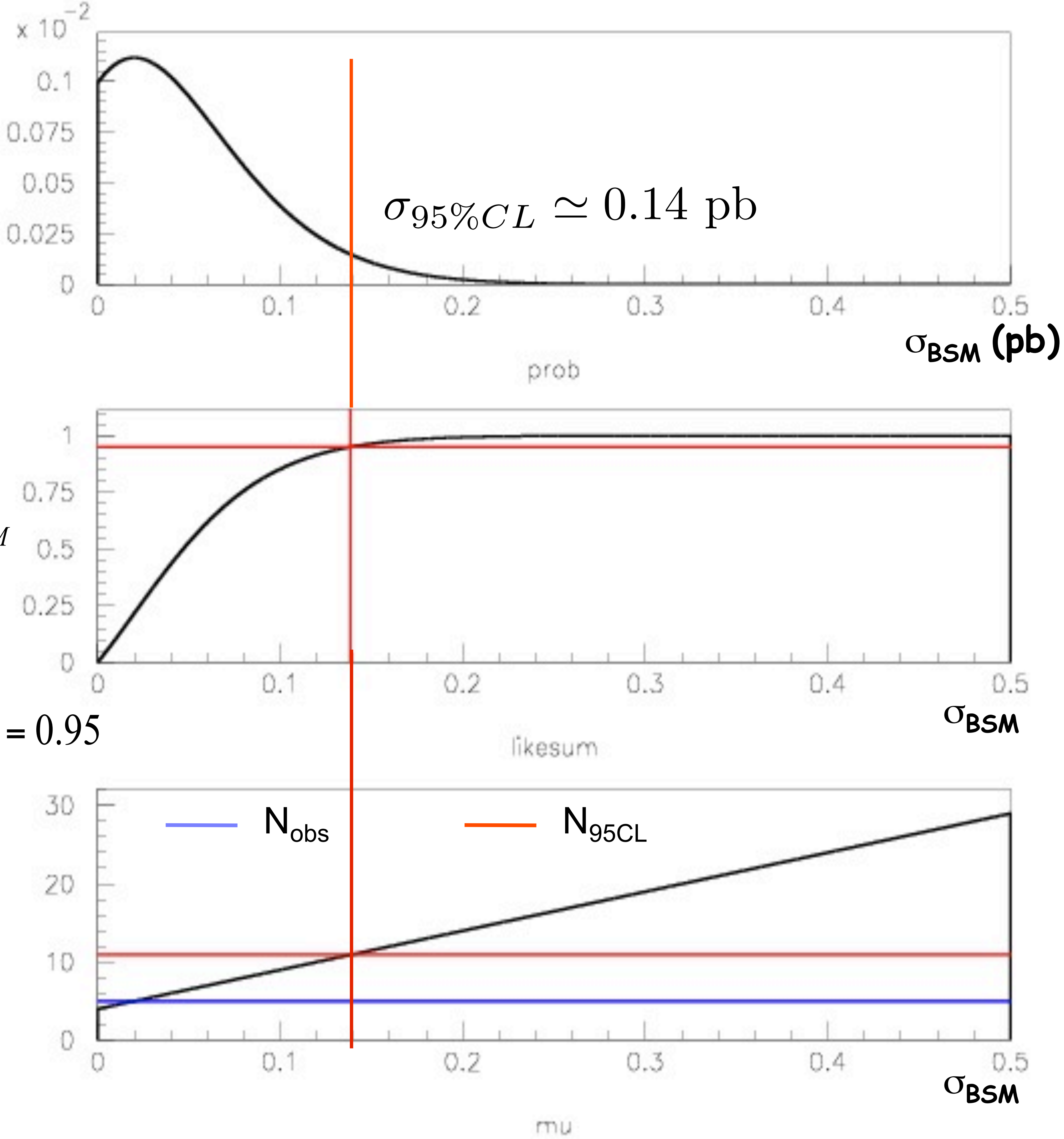
Choice of the prior ⟹ $P_0(\sigma_{BSM}) = \cos t \quad \sigma_{BSM} \geq 0$

$$P(\sigma_{BSM} \mid N_{obs}, N_{pred}, Lumi, eff)$$

$$\int P(\sigma_{BSM} \mid N_{obs}, N_{pred}, Lumi, eff)\, d\sigma_{BSM}$$

95% CL limit:

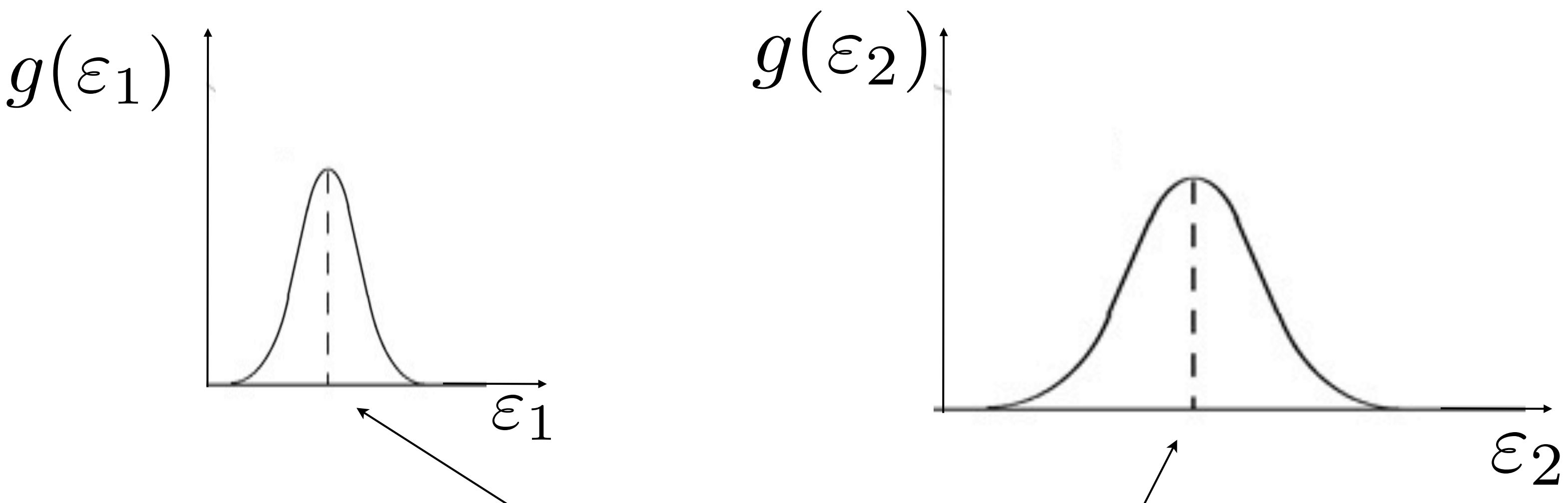
$$\int_0^{\sigma_{95CL}} P(\sigma_{BSM} \mid N_{obs}, N_{pred}, Lumi, eff)\, d\sigma_{BSM} = 0.95$$



$\sigma_{95\%CL} \simeq 0.14$ pb

Systematic uncertainties

The systematic uncertainties are treated as any other probabilistic variables. The dependence of the posterior by the systematic uncertainties can be eliminated integrating the parameters according to their probability distribution function.



$$P(H|E) = \frac{\int\int P(E|H,\varepsilon_1,\varepsilon_2) g(\varepsilon_1) g(\varepsilon_2) d\varepsilon_1 d\varepsilon_2 \times P_0(H)}{P(E)}$$

Example: measurement of a gaussian variable

Let’s come back to the previous example on the measurement of the mean value of a gaussian variable
This time assume that the measurement instrument be affected by a systematic offset uncertainty z, gaussian distributed around 0 (the instrument is correctly calibrated) with a standard deviation σ_z

Likelihood:

$$P(x_1|\mu) = \int \frac{1}{\sqrt{2\,\pi}\,\sigma_1} \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z} \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] dz$$

Posterior for the mean value μ:

$$P(\mu|x_1) = \frac{\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1} \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z} \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}z}{\int\!\!\int \frac{1}{\sqrt{2\,\pi}\,\sigma_1} \exp\left[-\frac{(x_1-\mu-z)^2}{2\,\sigma_1^2}\right] \frac{1}{\sqrt{2\,\pi}\,\sigma_Z} \exp\left[-\frac{z^2}{2\,\sigma_Z^2}\right] \mathrm{d}\mu\,\mathrm{d}z}$$

performing the integration:

$$\longrightarrow \quad \frac{1}{\sqrt{2\,\pi}\,\sqrt{\sigma_1^2+\sigma_Z^2}} \exp\left[-\frac{(\mu-x_1)^2}{2\,(\sigma_1^2+\sigma_Z^2)}\right]$$

The results is that $P(\mu)$ is still gaussian but its global σ is due to stat. and syst. uncertainties added in quadrature.

Final remark: both statistical and systematic uncertainties have probabilistic nature and are treated in the same way.
Different is the way the informations are acquired:
- statistical uncertainties are due to the fact that measurements are based on a finite set of observations,
- systematic uncertainties can be obtained from someone else (the instrument manufacturer), previous experiments, the knowledge of the detector, the simulation, the theoretical models...

Another difference is that usually we are not interested in the pdf of the systematic sources, but only on the effect of the systematic uncertainties on the pdf of the variable we want to measure. This is why we integrate them out (marginalization).

MonteCarlo methods

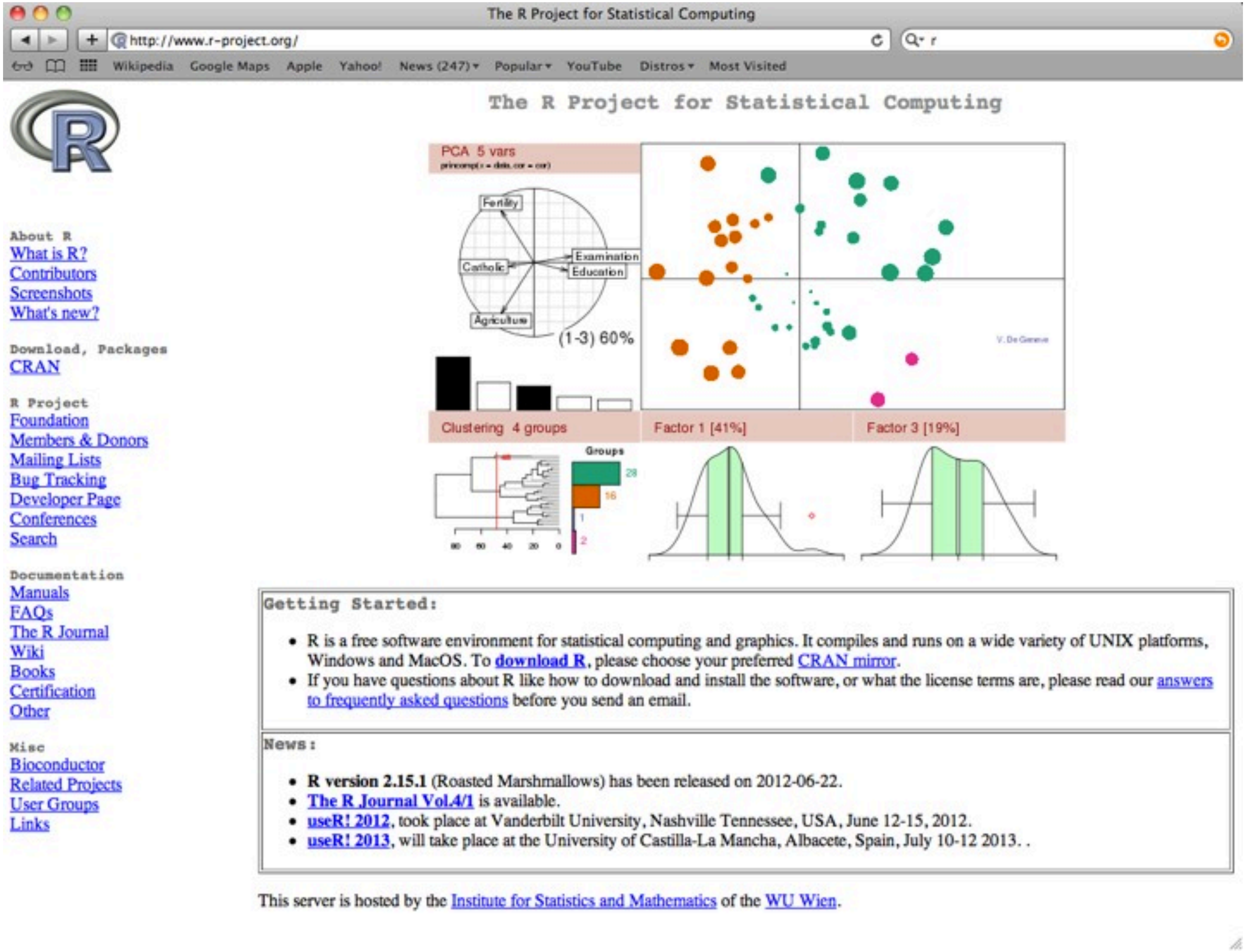
The MonteCarlo methods are a collection of techniques which use pseudo-random generators (computer simulated) to solve numerically mathematical problems too complicated to be solved analytically.

The easy task of the Bayesian approach is the local evaluation of the numerator of the Bayes formula (likelihood x prior), the difficult one is the normalization of such function, the estimation of expectation values, the marginalization respect to the nuisance parameters, the estimation of the credibility intervals. It is clear that if we were able to sample the posterior, the problem is solved at least approximately.

This is why the developments of such techniques have contributed to the spread and the success of the Bayesian approach in many different fields.

The R package: http://www.r-project.org/

free software environment for statistical computing and graphics



The R Project for Statistical Computing

About R
- What is R?
- Contributors
- Screenshots
- What's new?

Download, Packages
- CRAN

R Project
- Foundation
- Members & Donors
- Mailing Lists
- Bug Tracking
- Developer Page
- Conferences
- Search

Documentation
- Manuals
- FAQs
- The R Journal
- Wiki
- Books
- Certification
- Other

Misc
- Bioconductor
- Related Projects
- User Groups
- Links

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News:

- **R version 2.15.1** (Roasted Marshmallows) has been released on 2012-06-22.
- **The R Journal Vol.4/1** is available.
- **useR! 2012**, took place at Vanderbilt University, Nashville Tennessee, USA, June 12-15, 2012.
- **useR! 2013**, will take place at the University of Castilla-La Mancha, Albacete, Spain, July 10-12 2013. .

This server is hosted by the Institute for Statistics and Mathematics of the WU Wien.

Simple example: $\int_a^b g(x)dx \quad g = N(0,1)$ → R macro: norm.R

```
xs = rnorm(100000) # simulate 100000 draws from N(0,1)
xcount = sum((xs>-1) & (xs<0)) # count number of draws between -1 and 0
est=xcount/100000 # Monte Carlo estimate of probability
Rest=pnorm(0)-pnorm(-1) # Compare it to R's answer (cdf at 0) - (cdf at -1)
print("Estimation:")
print(est)
print("Compare it to R estimation (cdf at 0) - (cdf at -1):")
print(Rest)
hist(xs,breaks=100,xlim=c(-5,5),probability=TRUE,col="lightblue",main="Normal
distribution")
abline(v=0.,col="blue",lwd=2)
abline(v=-1.,col="blue",lwd=2)
```

In many cases we want to estimate the expectation value of a function f(x) respect to a pdf p(x):

$$< f >= \int f(x)p(x)dx$$

General method:

1- sampling p(x) generating $x_1,.....x_n$

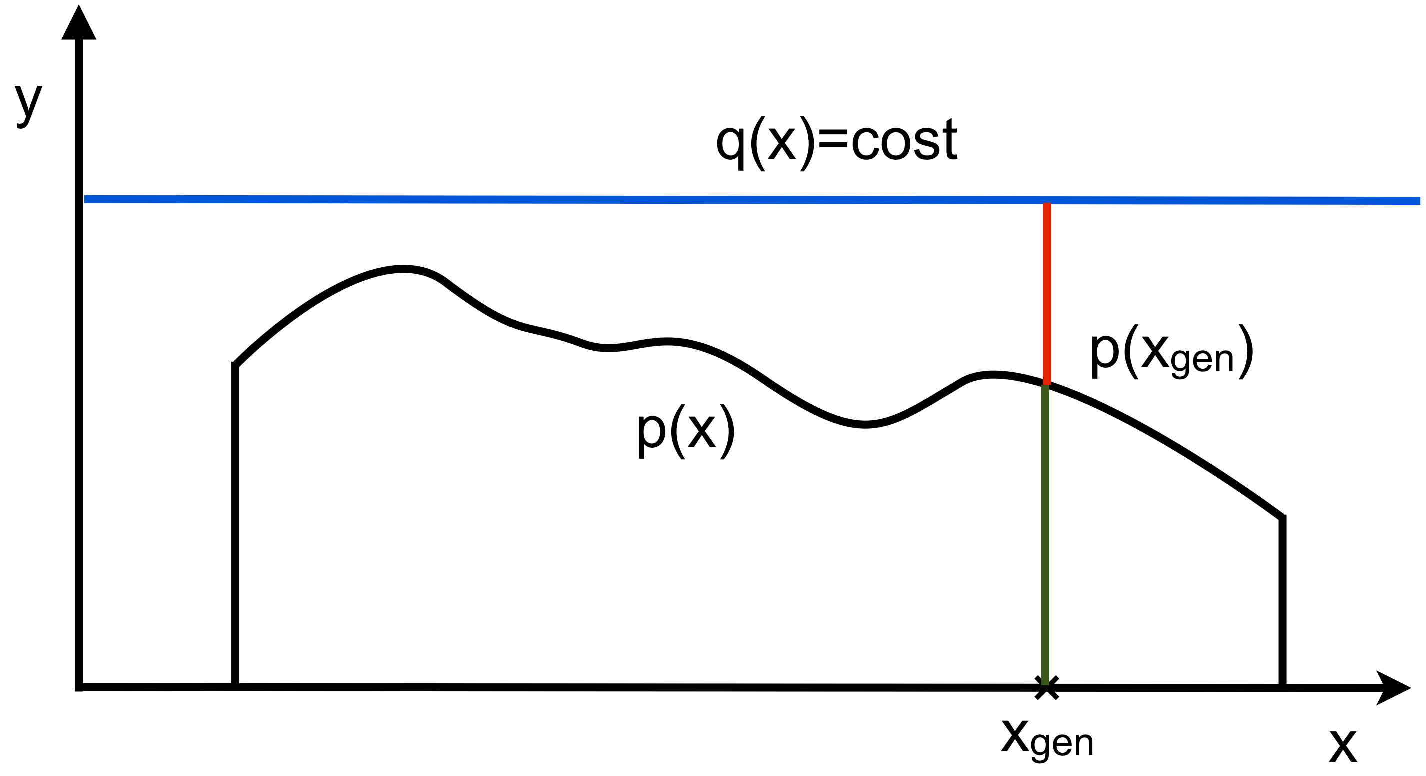
2- estimate <f> : $< f >= \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

So what we need are general sampling methods: →

Sampling methods

Rejection Sampling

If we are able to generate points (x) according to a function q(x) such that p(x)≤cq(x) (c constant), then we accept x with probability p(x)/cq(x).
It is clear that in this way q(x) is reshaped to p(x) and the procedure does not depend on the absolute normalization of p(x). In the simple 1-d case the trivial choice for q(x) is a uniform distribution (hit or miss method).



accepted with prob=p(x_{gen})/cost

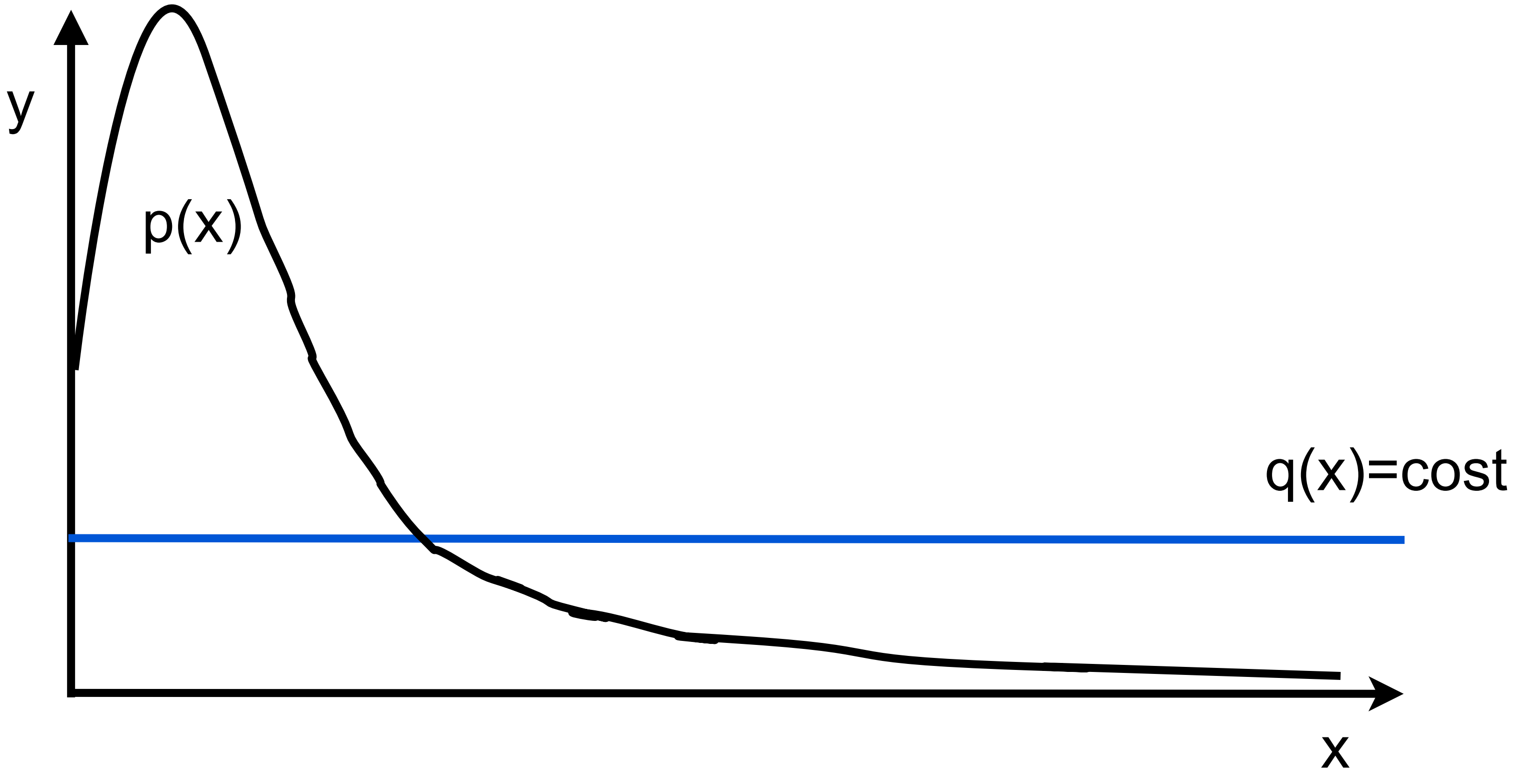
Importance Sampling

Also in this case we start with a sampling distribution q(x), but this time there is no requirement on q(x) (apart that q(x) should be positive):

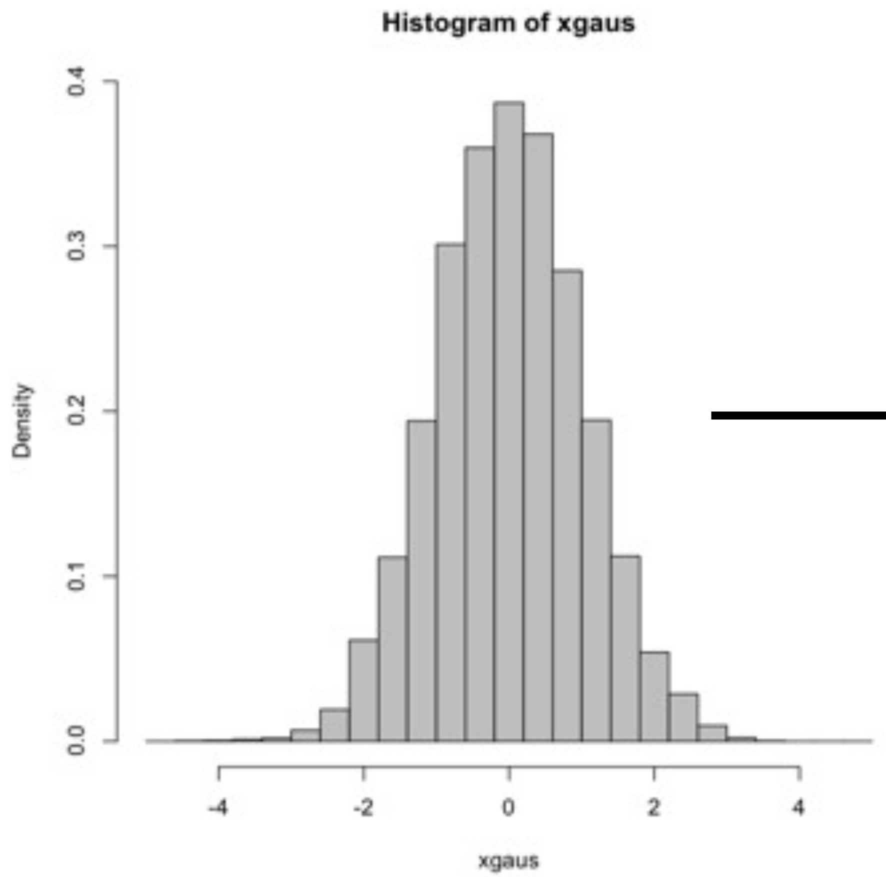
$$< f >= \frac{1}{n}\sum_{i=1}^{n} f(x_i)w(x_i) \quad w(x_i) = \frac{p(x_i)}{q(x_i)}$$

- possible to focus the sampling on part of the p(x)
- different pdf can be sampled using the same sampling distribution

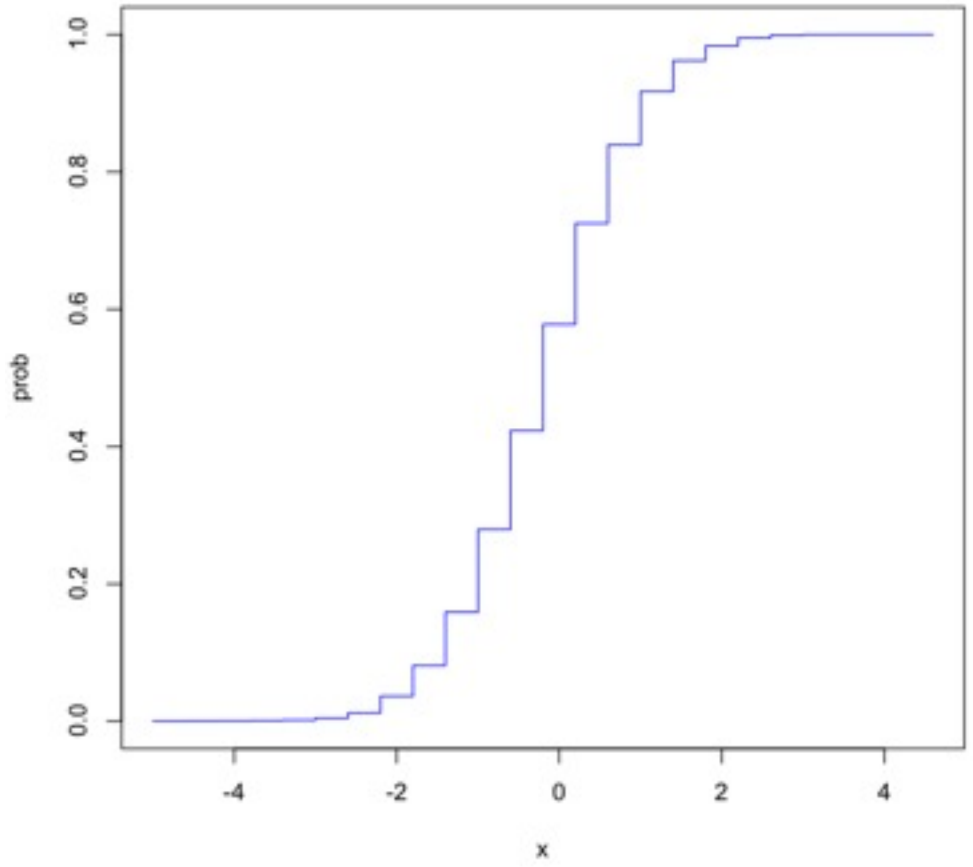
In this case the sampling produces weighted events.



Sampling inverting the cumulative distribution function (CDF)



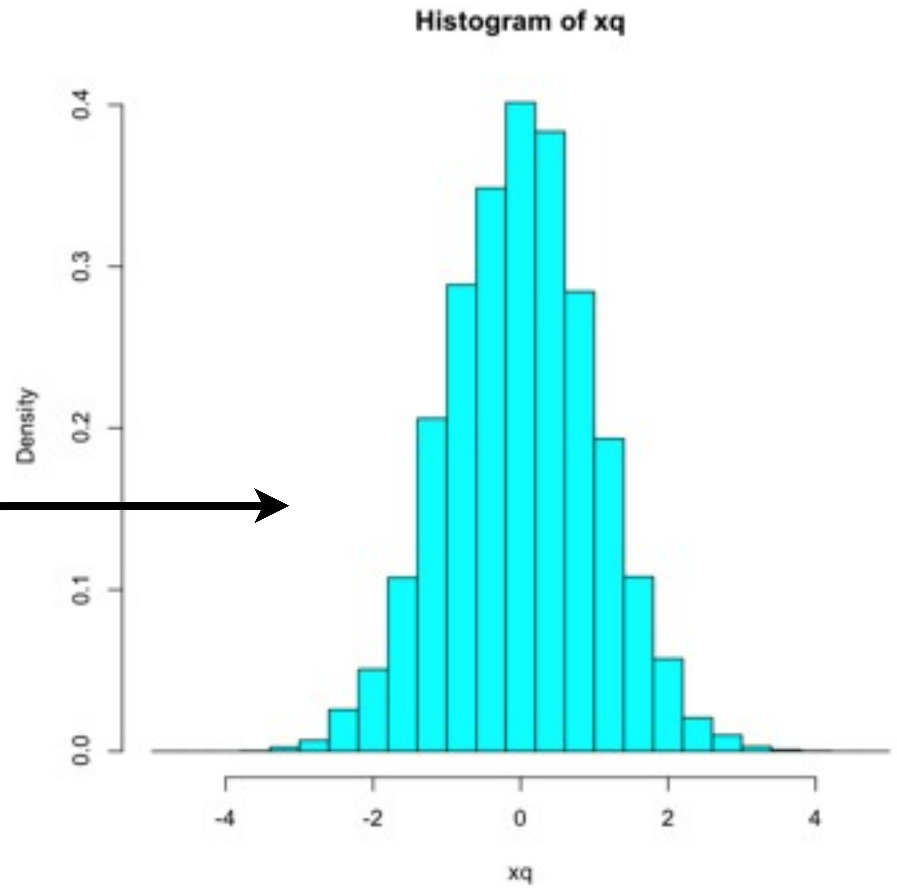
CDF



R macro: cdf.R

```
.........................
# sample a gaussian using the inverse CDF
 i=1
 xq<-c()
 while (i<=np)
 {
  xq<-c(xq,qnorm(runif(1)))
  i=i+1
 }
 dev.new()
 hist(xq,breaks=xb,freq=FALSE,col="cyan") # gaussian sampling using CDF^-1
}
```

uniform sampling of CDF^{-1}



Different approach: Markov chain Montecarlo

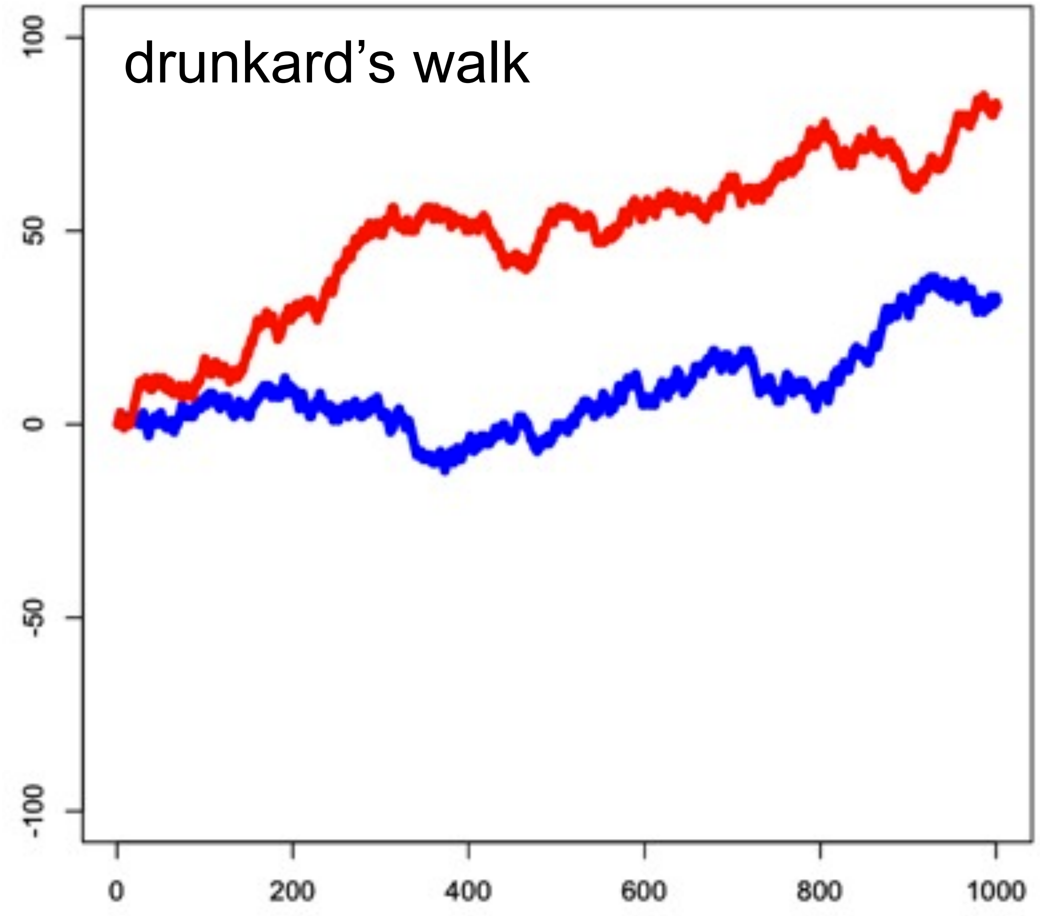
In this case the sequence of generated points form a random walk in the parameter space. The aim is to drive the random walk preferably towards high probability region of the parameter space

A **Markov process** is a chain of states produced by a transition process such that the probability of the next state depends only on the current state and not of the previous ones:

$$P(X_{n+1} = x|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x|X_n = x_n)$$

⇨ R macro: drunkwalk.R starting point x=0, P(x+1 | x)=P(x-1 | x)=0.5

```
x<-c(0)
y<-c(0)
for (i in 1:1000){
if( runif(1,-1,1) > 0 ) {x<-c(x,x[length(x)]+1)}
else {x<-c(x,x[length(x)]-1)}
}
for (i in 1:1000){
if( runif(1,-1,1) > 0 ) {y<-c(y,y[length(y)]+1)}
else {y<-c(y,y[length(y)]-1)}
}
plot(x,col=4,pch=20,ylim=c(-100.,100.))
par(new=T)
plot(y,col=2,pch=20,ylim=c(-100.,100.))
```

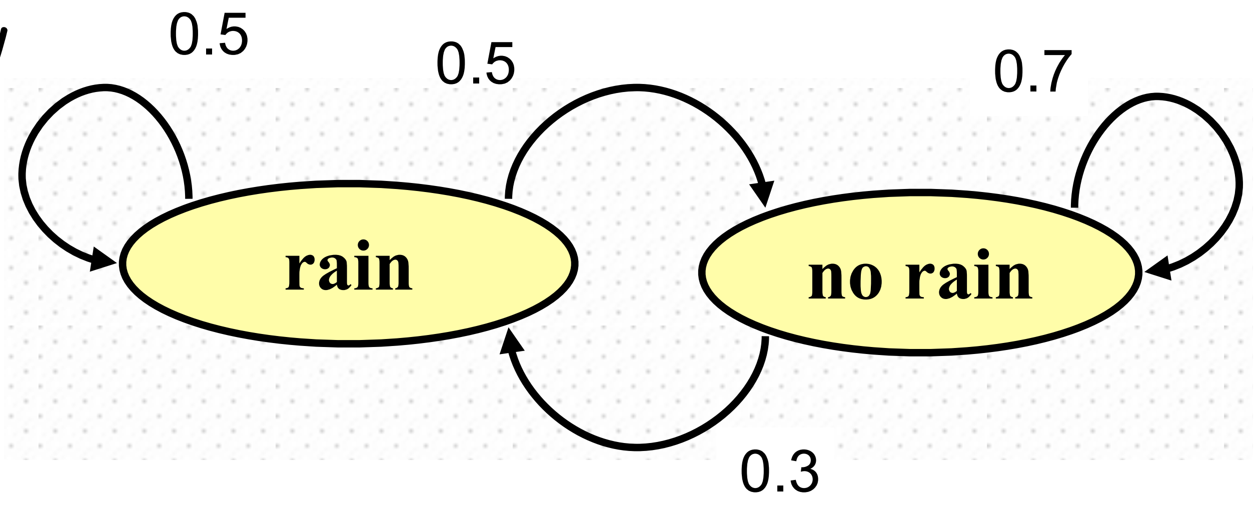


Markov Chains

Example: two state system $\begin{bmatrix} P(rain) \\ P(no\ rain) \end{bmatrix}$

raining today → 50% raining tomorrow
→ 50% not raining tomorrow

not raining today → 30% raining tomorrow
→ 70% not raining tomorrow



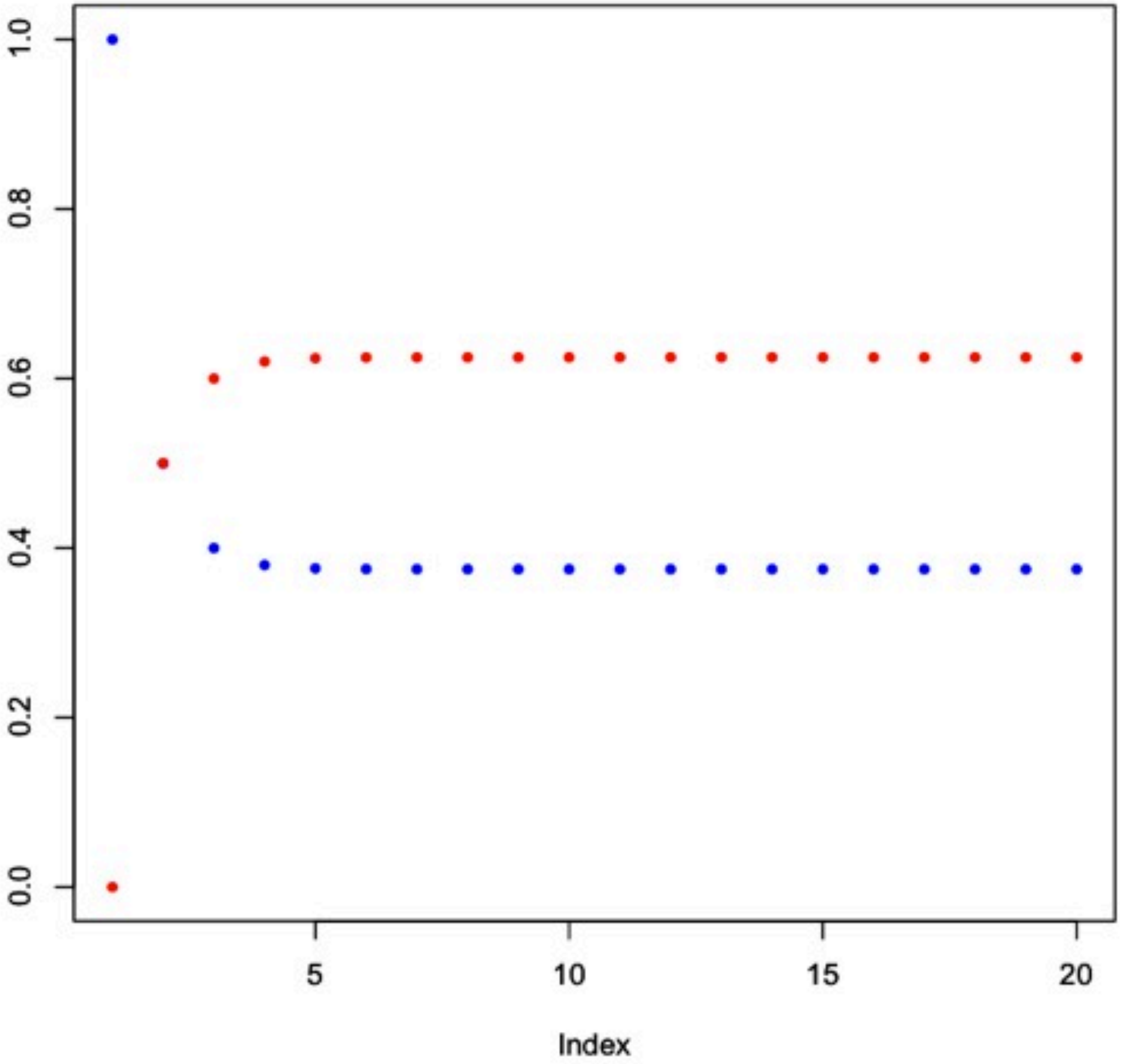
$$\begin{bmatrix} 0.5 & 0.3 \\ 0.5 & 0.7 \end{bmatrix}$$

transition matrix

Stationary assumption: transition matrix independent of time

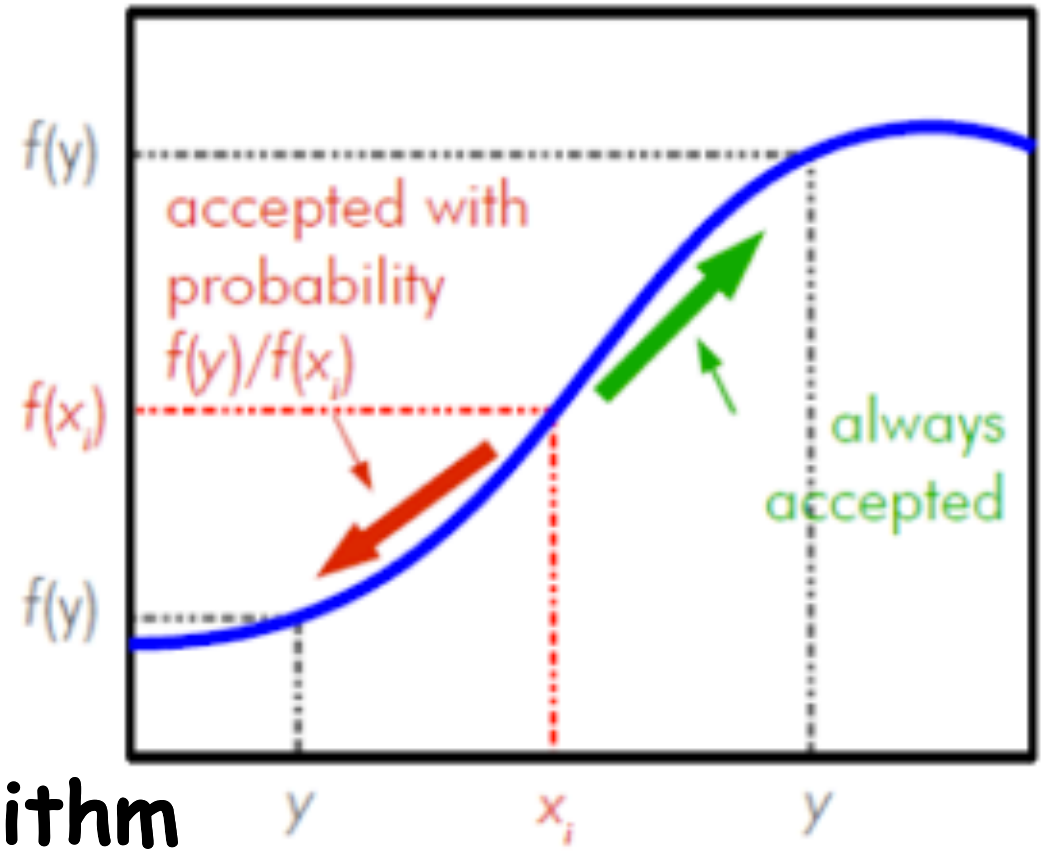
R macro: mcmcrain.R

```
matr<-c(0.5,0.3,0.5,0.7)
trans=matrix(matr,2,2,TRUE)
ini<-c(0.,1.)
story1<-c()
story2<-c()
state=matrix(ini,2,1)
for (i in 1:20){
 story1<-c(story1,state[1])
 story2<-c(story2,state[2])
 print(state)
 state=trans %*% state
# readline ("goto next iteration")
}
plot(story1,col=4,pch=20,ylim=c(0.,1.))
par(new=T)
plot(story2,col=2,pch=20,ylim=c(0.,1.))
```



How does MCMC work and why it is so important in Bayesian analysis

- Output of Bayesian analysis are posterior probability densities, often functions of a large numbers of parameters (n-dim space)
- Sampling n-dim functions is a difficult task
- use a random walk preferably driven to high probability region to efficiently explore the n-dim space



→ **Metropolis-Hastings algorithm**

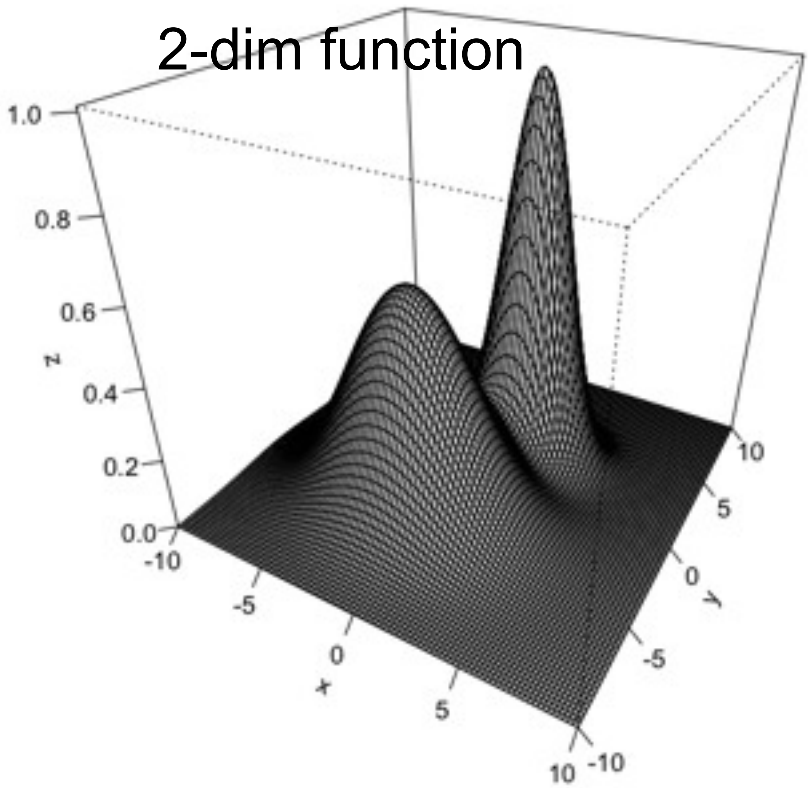
N.Metropolis et al., J. Chem. Phys. 21 (1953) 1087.

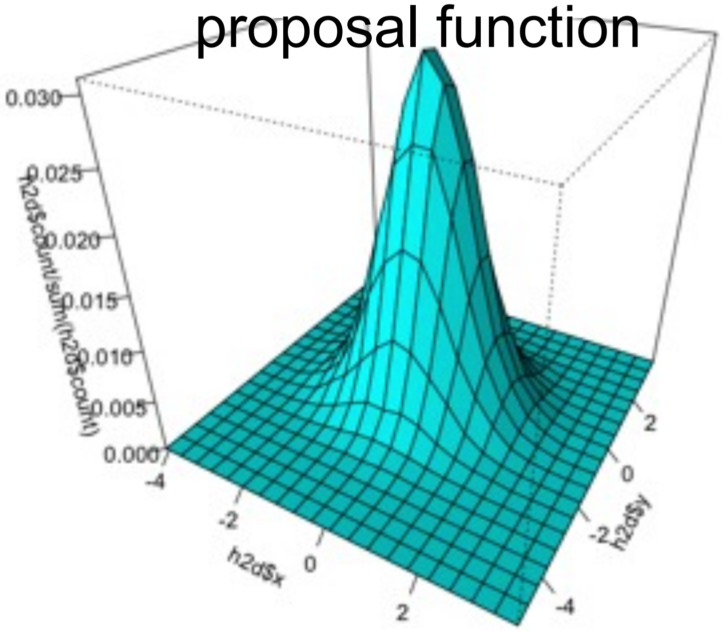
- Start at some randomly chosen x_i
- Randomly generate y around x_i → proposal function
- If $f(y) \geq f(x_i)$ set $x_{i+1} = y$
- If $f(y) < f(x_i)$ set $x_{i+1} = y$ with prob. $f(y)/f(x_i)$
- if y not accepted $x_{i+1} = x_i$ (stay where you are)
- start over

MCMC: Sampling a 2-dim function

R macro: func.R

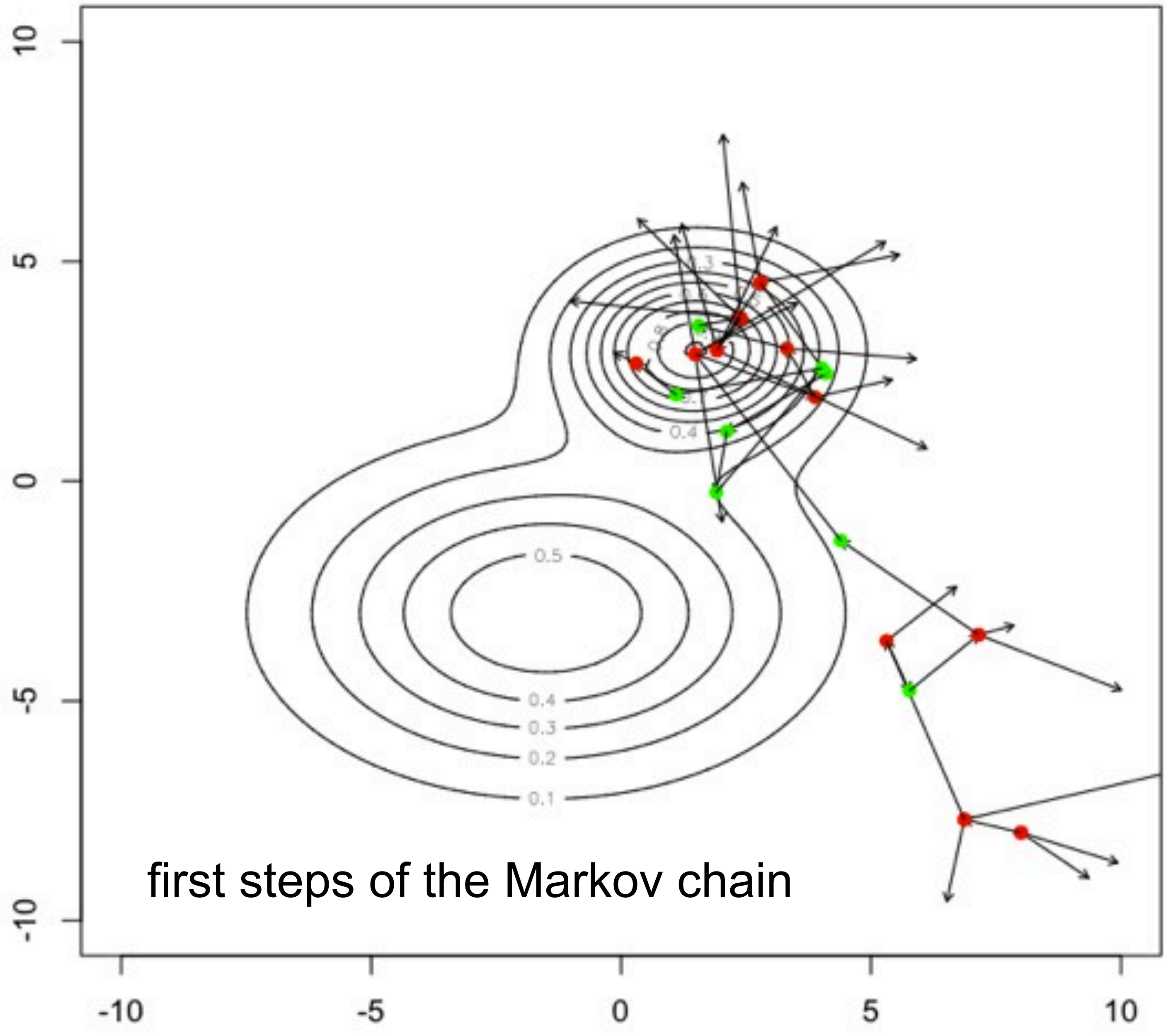
plot the 2-dim function


2-dim function


proposal function

R macro: mcmc_step.R

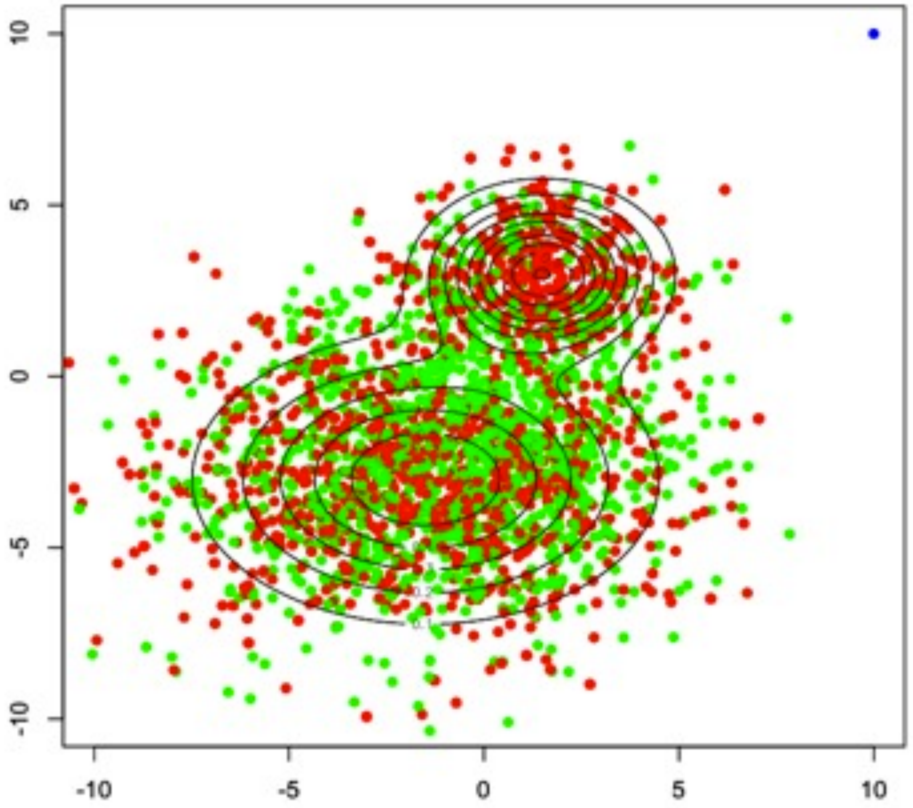
Show the developments of the Markov chain step by step. Two independent gaussians are used as proposal function


first steps of the Markov chain

MCMC: Sampling a 2-dim function

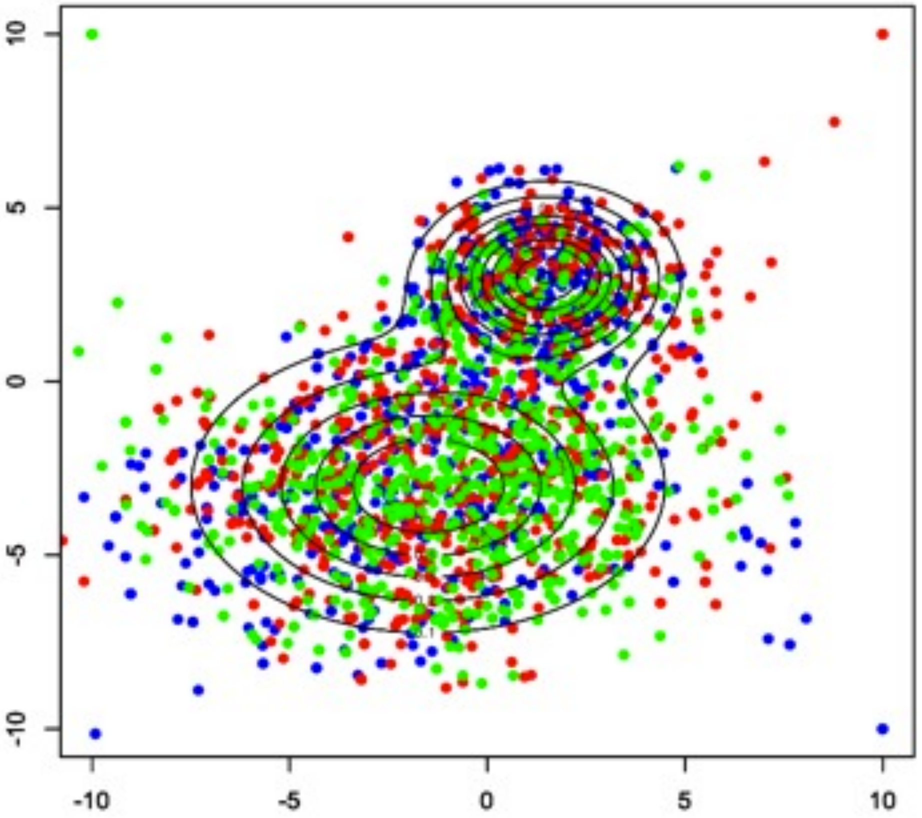
R macro: mcmc_acc.R

Show how the efficiency of accepting the next step and the proper sampling of the distribution is related to the width of the proposal function: green/red → step accepted/rejected



R macro: mcmc.R

Show the convergence of independent Markov chains starting from different points



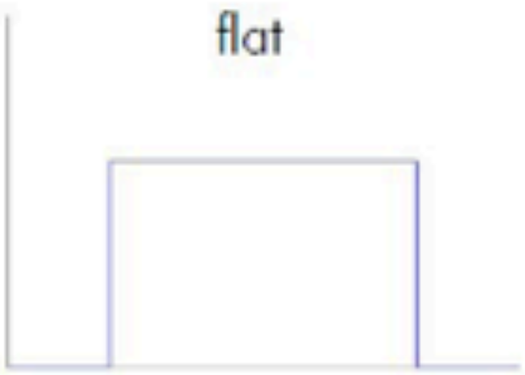
Summary of some relevant features of the Markov chain

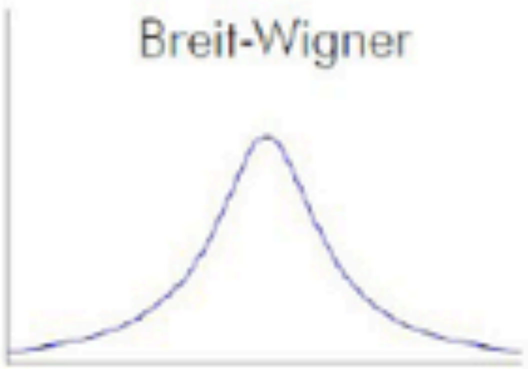
Generation of the next point

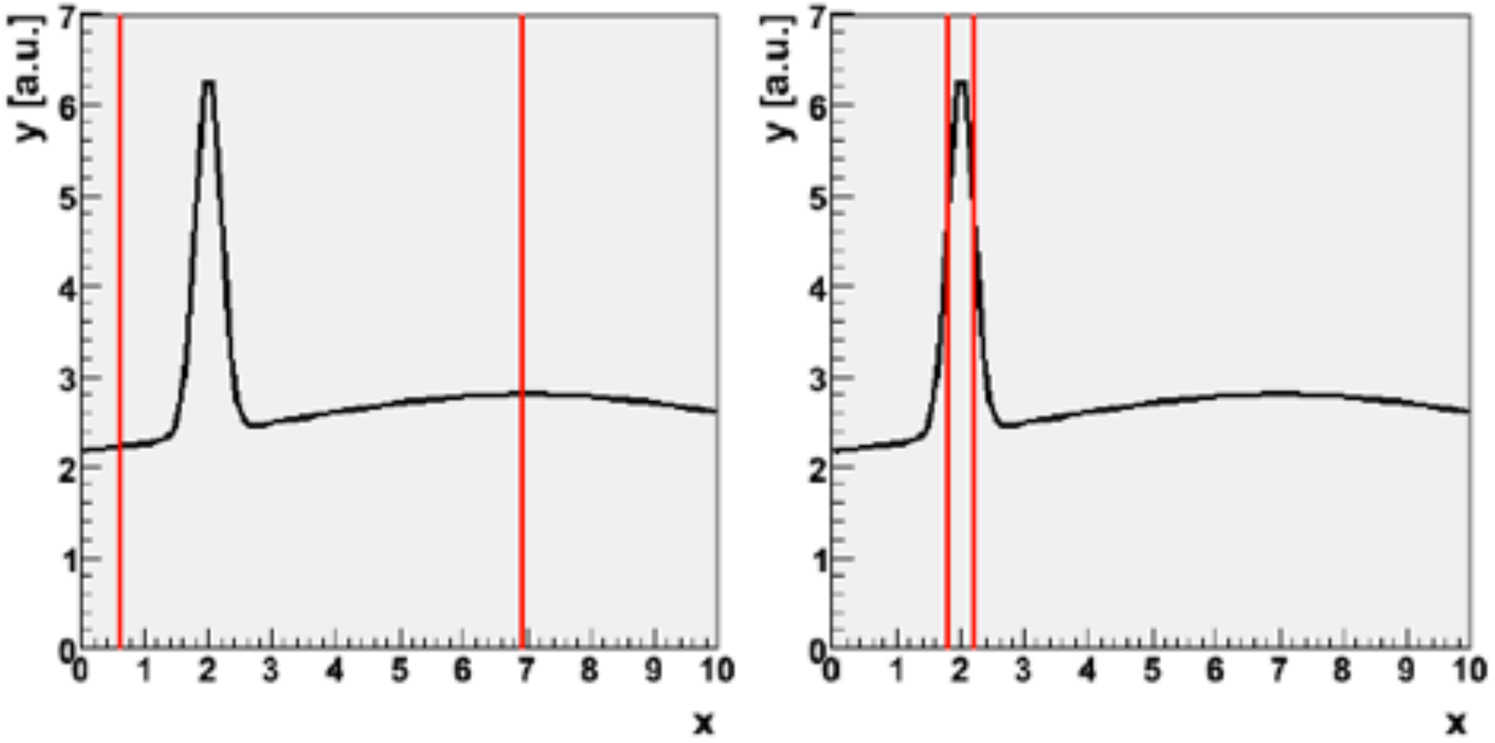
proposal function: probability density used to generate the next step

should be independent of the distribution we want to sample

- shape is very important (Breit-Wigner for example is a good choice)
- width related to the efficiency (fraction of the accepted points)







small width = large efficiency

large width = small efficiency

Use of MCMC in bayesian inference

- Use MCMC to sample the posterior probability distribution

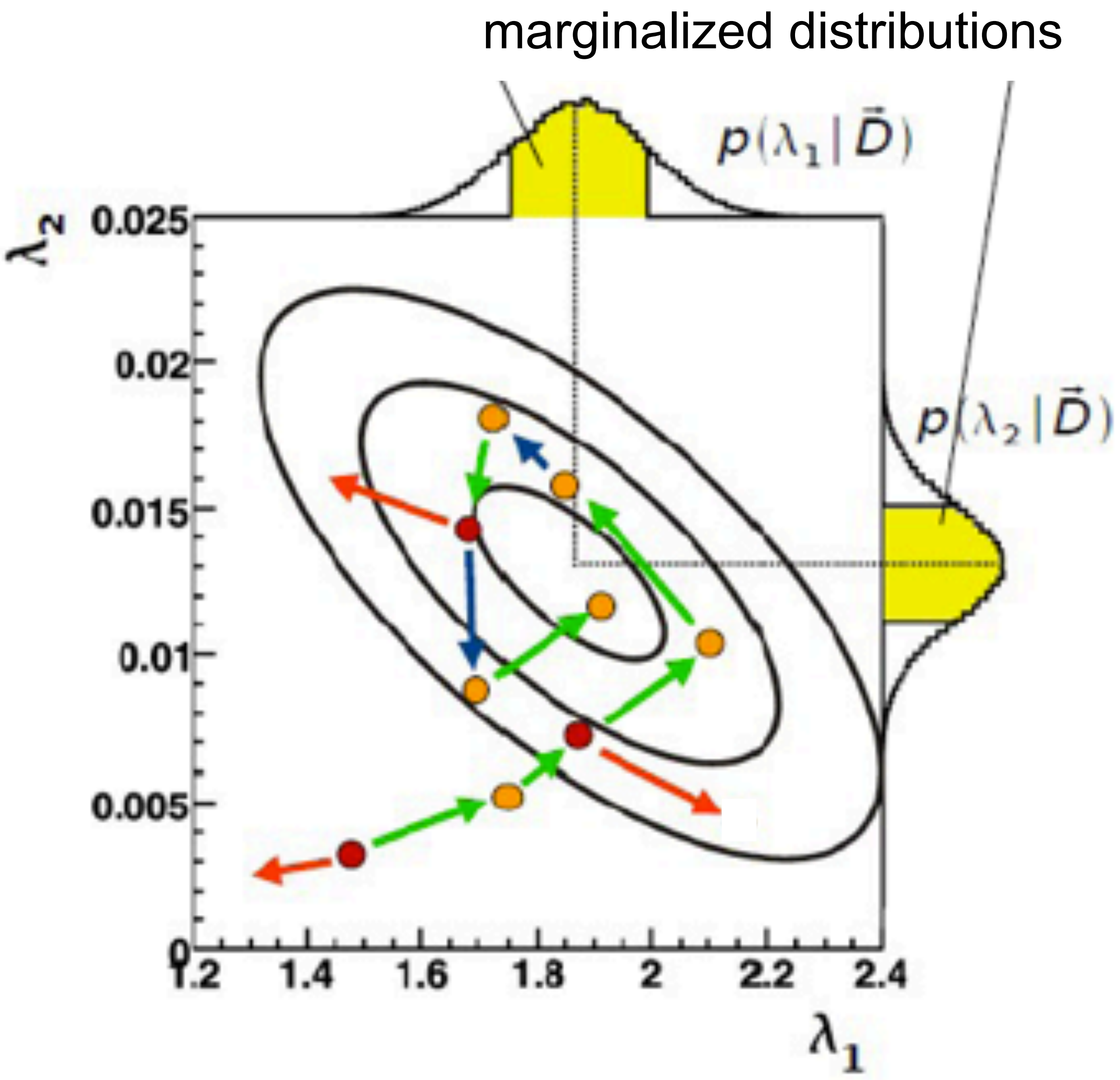
$$p(\vec{\lambda}|\vec{D}) \propto p(\vec{D}|\vec{\lambda})p_0(\vec{\lambda})$$

- Marginalization of the posterior:

$$p(\lambda_i|\vec{D}) \propto \int p(\vec{D}|\vec{\lambda})p_0(\vec{\lambda})d\lambda_{j\neq i}$$

just the projection on the corresponding parameter axis

- Error propagation: any function of the parameters can be evaluated during sampling
- Point estimate: find mode during sampling



Conclusions

My personal view after some years of involvements mainly as user in statistical issues:

- Statistics is often quite complicated and there are on the market a lot of recipes and programs often used as black-box.
- On the other hand one of the advantage in using the Bayesian approach is that you always control what you are doing.
- Its basic concepts are simple and natural, the connection of the probability, meant as “degree of belief”, with the status of information is exactly what we all do (mostly unconsciously) when we have to make decision.
- As we have seen some care is needed in understanding the problem and setting it up correctly, after that the road towards the result is well designed.

I strongly encourage you to use it, I’m pretty sure you’ll enjoy statistical analysis using Bayesian inference !