

BAYESIAN APPROACH

Recall The Bayes Theorem for events A and B :

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

Where : $P(A)$ = Prior degree of belief on A before knowing that B has happened

$P(A|B)$ = Posterior degree of belief on A after knowing that B has happened

• Bayesian probability and likelihood = parameter estimation

Suppose we have n measurements of a variable x with a given PDF which depends on some parameters $\vec{\theta} \in \Omega$.

The probability of obtaining the ~~data~~ (exactly) The data \vec{x} given all possible choices of the parameters $\vec{\theta}$ is :

$$P(\vec{x}|\vec{\theta}) = \mathcal{L}(\vec{x}|\vec{\theta})$$

Prior to the measurement, our degree of belief on the parameters is $\pi(\vec{\theta})$

Therefore we can use The Bayes Theorem To find The ^{posterior} probability of the parameters $\vec{\theta}$ ~~the~~ given the measurement \vec{x} :

$$P(\vec{\theta}|\vec{x}) = \frac{\mathcal{L}(\vec{x}|\vec{\theta})\pi(\vec{\theta})}{\int_{\Omega} \mathcal{L}(\vec{x}|\vec{\theta})\pi(\vec{\theta}) d\vec{\theta}}$$

→ depends only on $\vec{\theta}$ (\vec{x} is measured)

→ same dimension of $\vec{\theta}$

↳ This is a constant,
so it's just a normalization term.

Bayes 1

→ From $P(\vec{\theta}|\vec{\pi})$ we can find the global mode, which represents the most probable combination of all parameters.

~~The~~

→ Suppose we are interested only in one of the many parameters $\vec{\theta}$, say θ_1 . Then we define:

$\theta_1 = \text{parameter of interest} = \theta$

~~the~~ $\theta_2, \dots, \theta_n = \text{nuisance parameters} = \vec{v}$

This is just a choice. Mathematically there is no difference between the parameter of interest and the nuisance parameters.

We can ask ourselves: What is the most probable value of θ , given the measurement $\vec{\pi}$?

What is the PDF of θ given $\vec{\pi}$?

What is the ~~68%~~ interval that contains the true / most probable value of θ with 68% probability?

→ Marginalization!

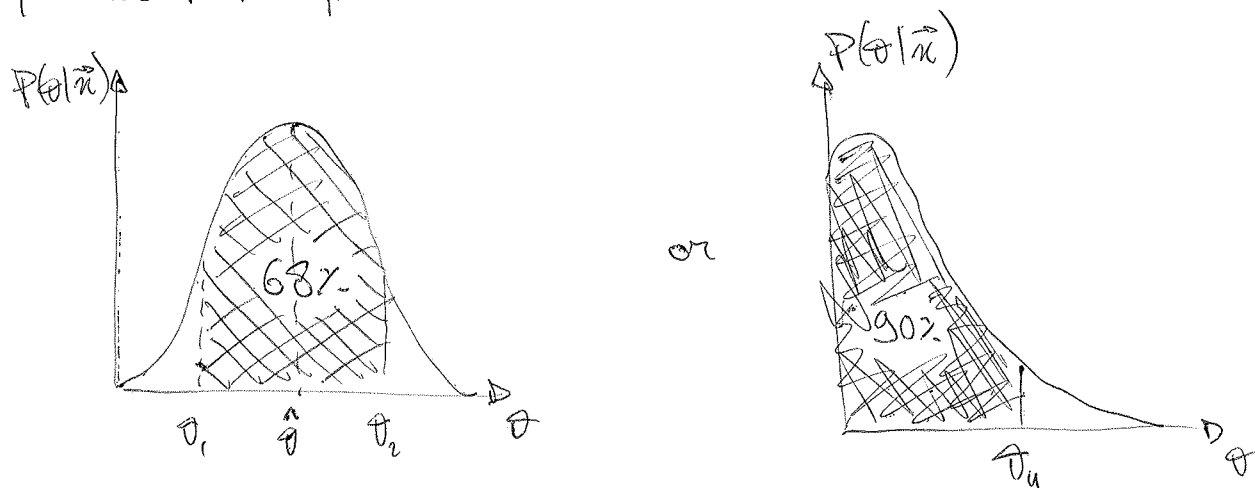
We extract the posterior PDF for θ by marginalizing over all the other ~~per~~ (nuisance) parameters:

~~$P(\theta|\vec{\pi}) = \int P(\theta, \vec{v}|\vec{\pi}) d\vec{v}$~~

$$P(\theta|\vec{\pi}) = \int P(\theta, \vec{v}|\vec{\pi}) d\vec{v} = \frac{\int \mathcal{L}(\vec{\pi}|\theta, \vec{v}) \pi(\theta) \pi(\vec{v}) d\vec{v}}{\int \mathcal{L}(\vec{\pi}|\theta, \vec{v}) \pi(\theta) \pi(\vec{v}) d\theta d\vec{v}}$$

• Credible intervals

The posterior PDF of θ will look like:



From $P(\theta|\vec{n})$ we can quote:

- The mode $\hat{\theta}$ (which could also be at the boundary of the physical region)
- The central / shortest interval $[\theta_1, \theta_2]$ which contains the ~~the~~ most probable value of θ with 68% probability
- If θ is near one of the borders, we can quote an upper or lower limit (Typically at 90% coverage).

These intervals are called "credible intervals".

They tell us that, based on our current knowledge, the true value of θ will be contained in that range with the specified probability.

↳ The concept of coverage here applies to the true parameter and is a property of the interval, not of the method

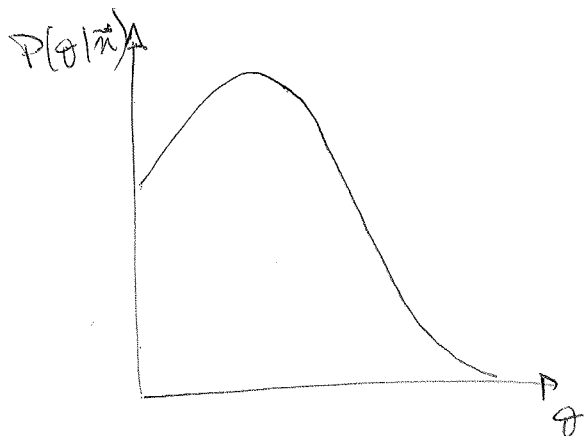
→ A Bayesian 68% credible interval might have 0% frequentist coverage, and this is not a problem.

We will quote as a result: $\theta = \hat{\theta}^{+ (\theta_2 - \hat{\theta})}_{- (\hat{\theta} - \theta_1)}$

Example: Poisson posterior for rate

- How to choose between measurement and limit

Suppose you have a posterior like this:

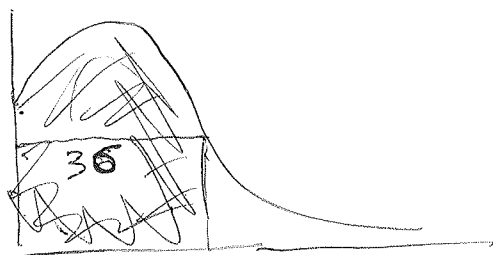


Should I quote the mode and the 68% credible interval, or the 90% credible interval limit?

~~There~~
We have two options:

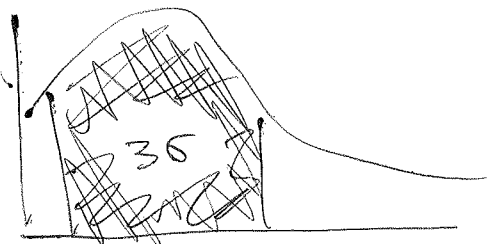
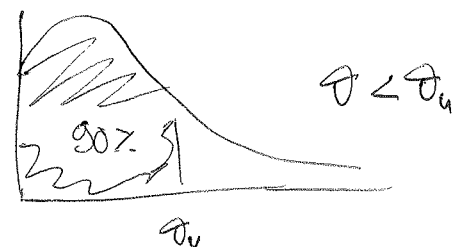
a) Easy but not too solid:

If the 3σ shortest interval hits a physical border, quote a limit, ^{90%} otherwise quote the mode and the upper/lower range:



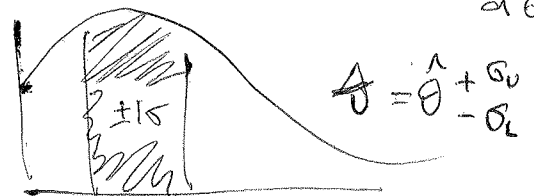
⇒

Quote limit at 90% c.i.:



⇒

Quote central value + uncertainty at 68%:



b) Complicated but more solid

If the parameter θ hits some physical border, maybe we are using the wrong model. This is especially true if we're testing the existence of new physics. (continues)

In such case, we should repeat the fit twice:

- 1) With the background only model H_0
- 2) With the signal + background model H_1 , which includes the parameter σ .

At this point, we need a criterion to ~~understand whether~~ ~~the~~ compare H_0 and H_1 , and see if we actually need the new signal term in H_1 .

If yes \rightarrow quote the mode and 68% c.i. shortest interval

If no \rightarrow still use the fit with H_1 model, but quote 90% c.i. limit on σ .

The criterion on how to compare the models H_0 and H_1 will be discussed next week in the session on hypothesis testing.

• Bayesian upper limits for Poisson counts

Suppose we measure a ~~process where~~ something where we have a signal contribution with s expected signal events and b expected bkg events.

Assume a flat prior on s and b ,

\rightarrow Zero background case

If $b=0$, so if we are sure that we have no background, and if we measure $n=0$ events, the posterior for s (assuming a flat prior on s and b) is:

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

$$P(s|n=0) = e^{-s}$$

The 90% c.i. limit is: $s < 2.3$ counts

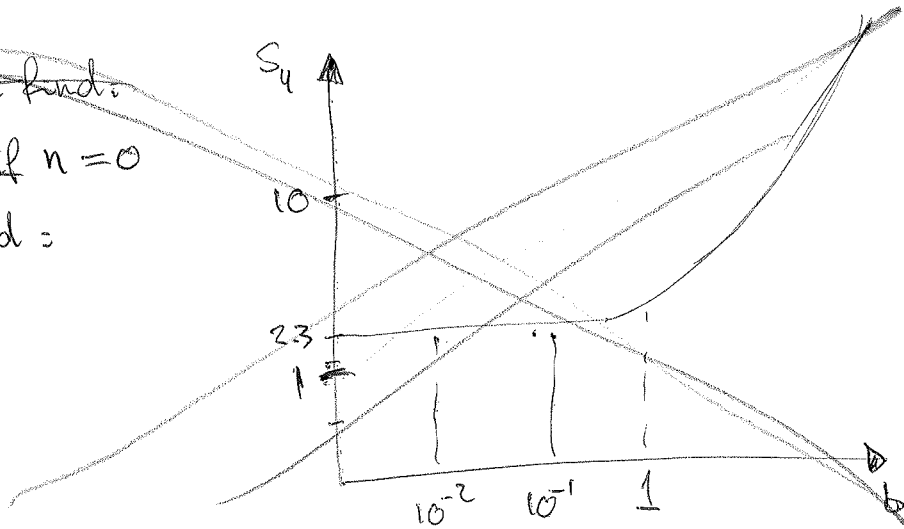
→ Non-zero background

$b \neq 0$ (the expected bkg $b \neq 0$, The measured one can be whatever)

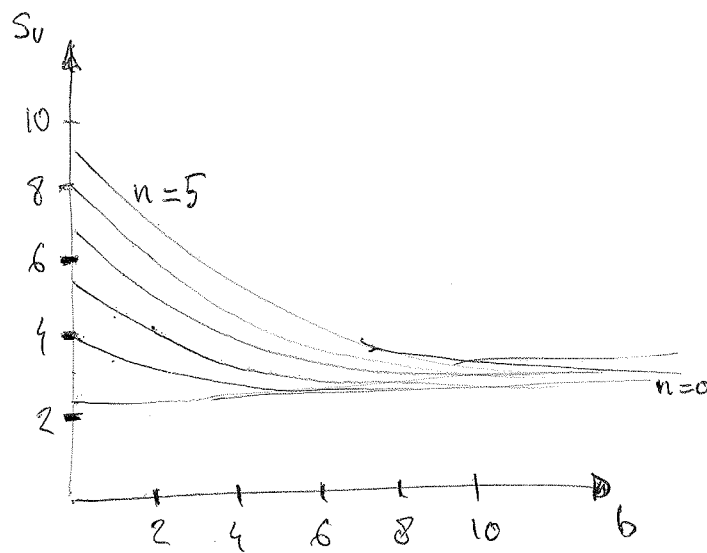
$n=0 \Rightarrow S < 2.3$ @ 90% c.i.

~~In general, we find:~~

In general, if $n=0$
we find:



If $n \neq 0$:



• Parameter ranges

How should I choose my parameter range?

Well, it depends on the case...

a) If There is a physical constrain in the parameter θ , use it!

For example, if $\theta = m = \text{mass}$, use $m \geq 0$.

b) If There is a prior measurement $\theta = \hat{\theta} \pm \sigma$, a good choice

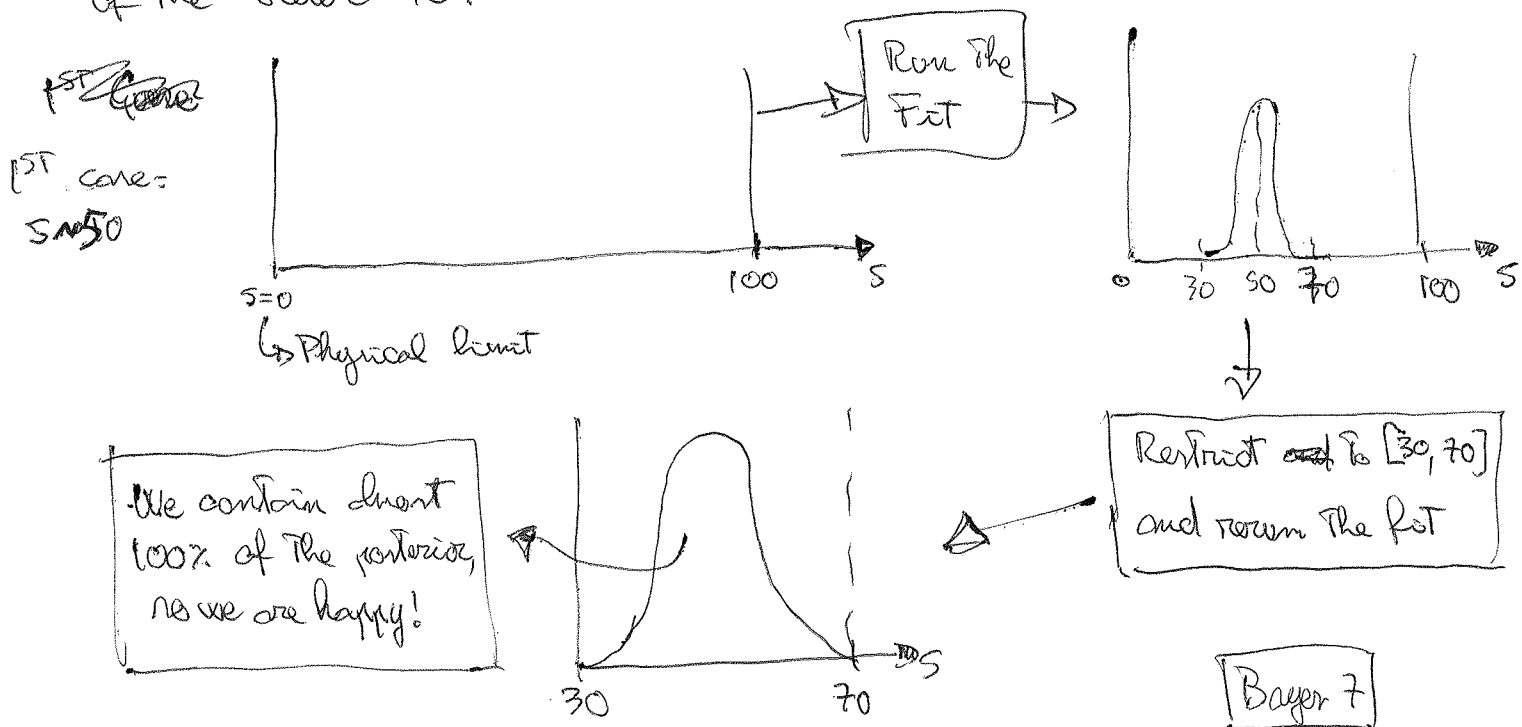
is to use $[\hat{\theta} - k\sigma, \hat{\theta} + k\sigma]$ with $k \geq 5$,

so that we are sure that: 1) The new measurement is likely to be well contained in the parameter range.

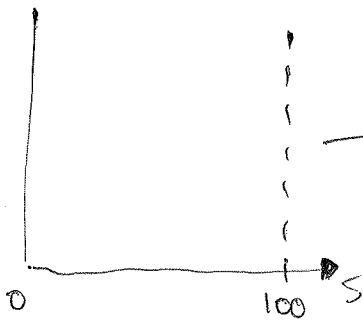
2) The posterior distribution is likely to be fully contained in the range, with the tails not hitting the borders.

c) Otherwise, ~~use a range~~ start with a large range, run the fit, check the posterior, and restrict it to $\pm(5-10)\sigma$, then repeat the fit. If you hit a border that is not justified by physics, enlarge it.

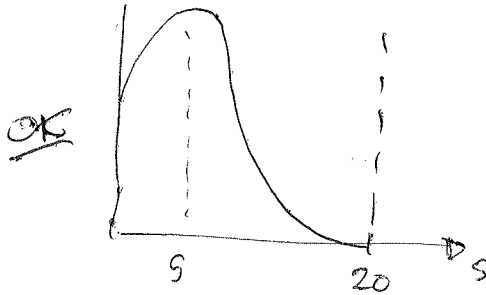
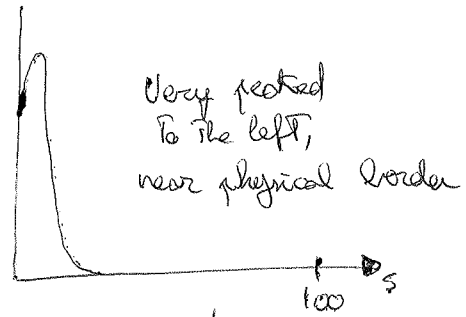
→ Suppose we have a parameter s describing some number of counts, and we have no prior knowledge on it, but we expect s to be of the order 10.



2nd case:
 $S \sim 5$

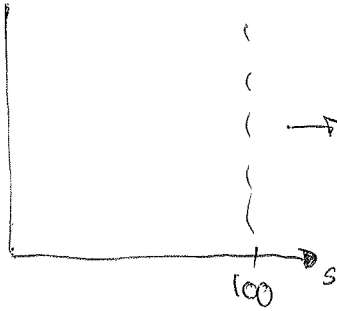


Run the fit

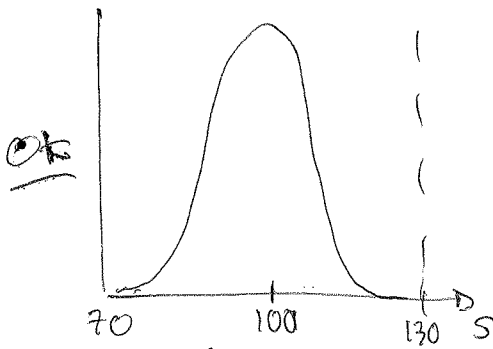
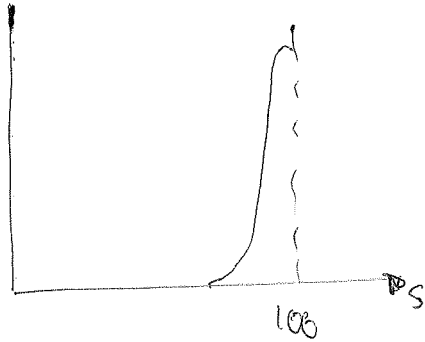


Enlarge Restrict upper border to "zoom" the posterior, keeping the physical boundary. Then rerun the fit

3rd case:
 $S \sim 100$



Run the fit



Enlarge above 100, because it's not a physical boundary.
 Also restrict lower border, because we know we're well above zero.
 Rerun the fit.

NB: This has to be repeated for every parameter, which is a pain in the arse!

- Bayes fits: practical implementation

Let's go back to the Bayes Theorem for parameter estimation:

$$P(\vec{\theta} | \vec{x}) = \frac{\mathcal{L}(\vec{x} | \vec{\theta}) \pi(\vec{\theta})}{\int \mathcal{L}(\vec{x} | \vec{\theta}) \pi(\vec{\theta}) d\vec{\theta}}$$

In principle, we need to:

1) Map the posterior $P(\vec{\theta} | \vec{x})$ and integrate it over the nuisance parameters.

However: a) $P(\vec{\theta} | \vec{x})$ might be very complicated to integrate over $d\vec{\theta}$

b) $\mathcal{L} \cdot \pi$ might be too complicated to integrate over $d\vec{\theta}$

~~Solution Observation:~~

b) ~~We can~~ $\int \mathcal{L} \cdot \pi d\vec{\theta}$ is a constant that does not depend on $\vec{\theta}$

a) We can do

Solution:

1) Map $\mathcal{L} \cdot \pi$ with a Markov Chain (e.g. Metropolis-Hastings) with n tested samples

2) Forget about the denominator $\int \mathcal{L} \pi d\vec{\theta}$, because it's just a constant that does not depend on $\vec{\theta}$.

3) Renormalize the posterior $P(\vec{\theta} | \vec{x})$ by $\frac{1}{n}$

4) To build the marginalized of some parameter θ_i , just put the θ_i values of all ~~test~~ accepted samples of the Markov-Chain into a histogram, and normalize it by $1/n$!

5) Extract $\hat{\theta}_i$ and σ_i from the marginalized histogram.

• Prior

So far, we've neglected the presence of a fundamental ingredient of Bayesian analysis: priors!

Their presence introduces some amount of ~~my~~ subjectivity into the method and, therefore, into the result.

In general, if our ~~measurements~~ data are very informative, any reasonable prior will have very little effect on the result.

~~For example, if we measure thousands of counts and had no pri.~~

→ For example, suppose we want to measure the events associated to some new physics signal for which we have no prior knowledge.

We can use a flat prior in $[0, 100]$: $\pi(s) = \begin{cases} \frac{1}{100} & \text{if } 0 \leq s < 100 \\ 0 & \text{otherwise} \end{cases}$

If we measure 50 events (and 0 background), the posterior for s will be peaked around 50, which is what we expect.

However, if our data are not very informative, we have a strong dependence of the result on the prior.

Example: same as before, but suppose we measure 3 events. The mode of s will be ~ 4 !

→ If we have a prior measurement, use it as a prior!!

→ Possible prior ~~choices~~ choices for parametrizing our ignorance

a) Flat on observable (e.g. number of counts)

↳ Leads to conservative limits

↳ Seems to be ~~an~~ a parametrization of ignorance, but it's not necessarily true

b) Flat on some physics parameter which is not necessarily an observable

Example: If I measure the decay rate of an isotope, I could use

$$T_{1/2} = \frac{\ln 2}{\Gamma} \quad \text{as a fit parameter instead of the}$$

number of counts (observable).

However $\text{flat}(T_{1/2})$ does not correspond to $\text{flat}(N)$!

c) Flat on \log_{10} of some parameter or observable

↳ Sometimes called "scale-invariant" prior

↳ Corresponds to saying that we give the same prior probability to the parameter to be e.g. of order 10 or of order 10^3

Notice however that all these priors ~~are~~ are not invariant under reparametrization

b) none of them is a ^{truly} "uninformative" prior.

In fact, there is an intrinsic arbitrariness in the method that we cannot avoid.

→ Example: religious belief

• Jeffrey's Priors

Jeffrey's priors are a set class of "uninformative" priors that can be used in case we have no knowledge about the parameter, and that are invariant under reparametrization.

Jeffrey's prior are of the type: $\pi(\vec{\theta}) = \sqrt{I_F(\vec{\theta})} \rightarrow$ Fisher information matrix

where:
$$I_F(\vec{\theta}) = \det \left[E \left[\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right] \right]$$

For example, we ~~can~~ have the following Jeffrey's priors for these PDFs:

PDF	Jeffrey's prior
Poisson mean s	$1/\sqrt{s}$
Poisson signal s + background b	$1/\sqrt{s+b}$
Gaussian mean μ	uniform
Gaussian std σ	$1/\sigma$
Binomial efficiency ε	$1/\sqrt{\varepsilon(1-\varepsilon)}$
Exponential decay constant λ	$1/\lambda$

- Improper (= diverging) priors

Suppose I have a parameter s describing some number of counts.

Suppose I want to use a scale invariant prior:

$$\pi(\log_{10} s) = \text{const} \quad \text{with } 0 \leq s \leq \max s$$

How do I choose $\max s$?

~~If I don't have an~~

Notice that ~~the~~ the integral of this prior diverges, so we need a cutoff!

~~In fact~~

Again, there is no general solution.

However, ~~we~~:

- The data might be pushing the ~~data~~ posterior to some restricted region of ~~the~~ parameter space, so we can just ~~cut the~~ set the borders around that region.

This is what we did before.

- There might be some volume effect due to the choice of the parameters used as a fit bases.

If you are interested I show some example, otherwise I'm happy to skip.

Examples \Rightarrow Bayesian efficiency fit

\rightarrow Simultaneous fit