Statistics for Astronomers: Lecture 17, 2021.01.04

Prof. Sundar Srinivasan

IRyA/UNAM

Bayesian inference

Recall: Bayes' Theorem

Definition (Bayes' Theorem)

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

Under the Bayesian Interpretation of probability, this is read as

$$\text{Updated deg. of belief in } A = \text{Support for } A \text{ from evidence } B \times \text{Original deg. of belief in } A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Cond. prob. of } B \text{ given } A}{\text{Marginal prob. of } B} \times \text{Prior prob. of} A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Likelihood of } A \text{ given } B}{\text{Evidence } B} \times \text{Prior prob. of } A.$$

Recall: Bayes' Theorem

Definition (Bayes' Theorem)

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

Under the Bayesian Interpretation of probability, this is read as

$$\text{Updated deg. of belief in } A = \text{Support for } A \text{ from evidence } B \times \text{Original deg. of belief in } A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Cond. prob. of } B \text{ given } A}{\text{Marginal prob. of } B} \times \text{Prior prob. of} A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Likelihood of } A \text{ given } B}{\text{Evidence } B} \times \text{Prior prob. of } A.$$

Multiple events A_j: normalisation requires computing the sum (integral)

$$P(B) = \sum_{j=1}^{N} P(B|A_j) \times P(A_j).$$

Recall: Bayes' Theorem

Definition (Bayes' Theorem)

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

Under the Bayesian Interpretation of probability, this is read as

$$\text{Updated deg. of belief in } A = \text{Support for } A \text{ from evidence } B \times \text{Original deg. of belief in } A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Cond. prob. of } B \text{ given } A}{\text{Marginal prob. of } B} \times \text{Prior prob. of} A.$$

or

$$\text{Posterior prob. of } A \text{ given evidence } B = \frac{\text{Likelihood of } A \text{ given } B}{\text{Evidence } B} \times \text{Prior prob. of } A.$$

Multiple events A_j: normalisation requires computing the sum (integral)

$$P(B) = \sum_{j=1}^{N} P(B|A_j) \times P(A_j).$$

This is typically the most computationally intensive step – Monte Carlo sampling techniques.

OR can leave it as a proportionality.

Coin toss: prior selection

Observation: 7 heads in 10 tosses. What is $P(\mathrm{Head}) \equiv \theta$?

Need to pick a prior probability distribution for θ. If no information provided, assume coin is fair.

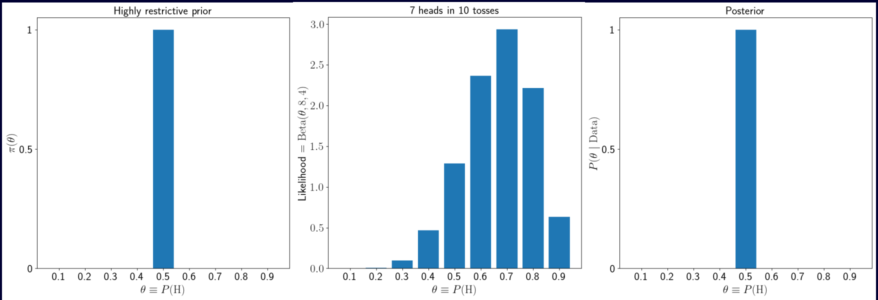
Frequentist: can maximise likelihood. Bayesian: select prior and multiply into likelihood.

Coin toss: prior selection

Observation: 7 heads in 10 tosses. What is $P(\mathrm{Head}) \equiv \theta$?

Need to pick a prior probability distribution for θ. If no information provided, assume coin is fair.

Highly restrictive prior: $P_\ominus(\theta) = 1$ if $\theta = 0.5$, 0 otherwise.

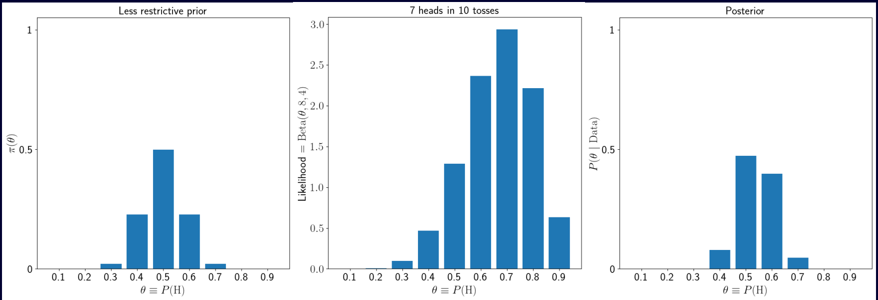


Coin toss: prior selection

Observation: 7 heads in 10 tosses. What is $P(\mathrm{Head}) \equiv \theta$?

Need to pick a prior probability distribution for θ. If no information provided, assume coin is fair.

Less restrictive prior: $P_\ominus(\theta)$ peaks at $\theta = 0.5$, but has finite probability around this value.

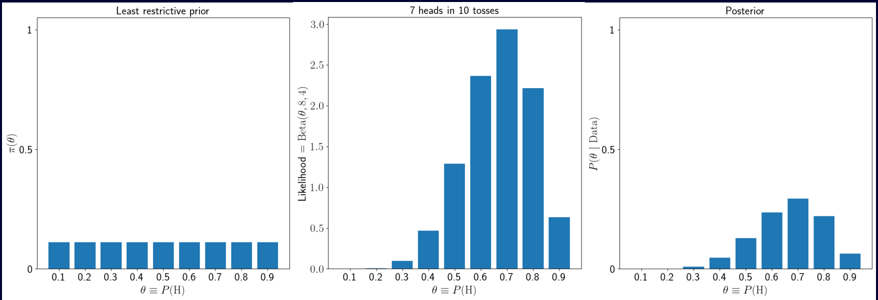


Coin toss: prior selection

Observation: 7 heads in 10 tosses. What is $P(\mathrm{Head}) \equiv \theta$?

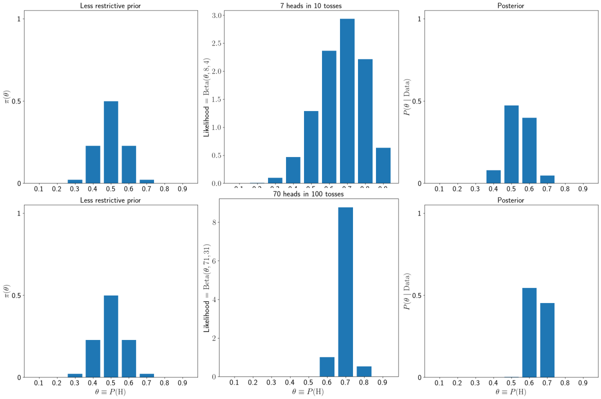
Need to pick a prior probability distribution for θ. If no information provided, assume coin is fair.

Least restrictive prior: $P_\Theta(\theta)$ is constant for $\theta \in [0.1, 0.9]$.



Prior-dominated vs. data/evidence-dominated posterior

Use less restrictive prior, but two datasets: (1) 7 heads in 10 tosses (2) 70 heads in 100 tosses.



Prior choice becomes irrelevant with increasing data size.

Moral: always choose a prior, any prior, as long as it isn't a delta function.
results from large datasets will be independent of the choice of prior.

Bayesian point/location and interval estimates

Once the posterior $p(\theta|\mathrm{data})$ is computed, we can compute the location estimates (mean, median, mode) and interval estimates.

For example, the Bayesian estimator of the parameter mean is $\bar{\theta} = \int d\theta\ \theta\ p(\theta|\mathrm{data})$.

We can also compute Bayesian interval estimates, also called posterior intervals or credible intervals (abbreviated in these lectures as CrI).
Same procedure for computing intervals as in frequentist case, but interpretation different.

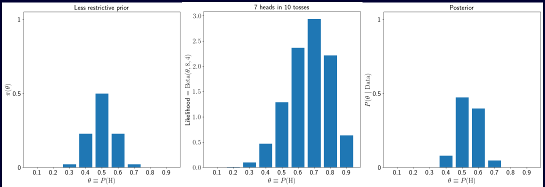
Commonly used CrI: highest posterior density (HPD) interval, defined as the narrowest interval containing $100(1-\alpha)\%$ of the posterior probability.

Interpretation of 95% interval w.r.t. true parameter value:

- Frequentist (confidence interval) - 95% of such intervals will include the true value.
- Bayesian (credible interval) - each interval has 95% probability of including true value!

Coin toss: Bayesian point and interval estimates

The posterior PDF can be used to compute a point estimate for $P(\mathrm{Head})$, as well as an interval estimate (posterior/credible interval).



Approximate *a posteriori* values of the
- max. (MAP): $\theta = 0.5$.
- median: $\theta = 0.5$.
- mean: $\theta = 0.5$.

Approx. 95% HPD interval: $[0.4, 0.7]$.

Examples of Bayesian point estimates: median or mode of posterior PDF.
The mode is the maximum a posteriori (MAP) estimate.

Example of credible interval: highest posterior density (HPD) interval.
Encompasses region with highest probability density.

Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Informative prior

Specific information about parameter(s). Progressively increasing amounts of data $\Longrightarrow$ posterior is evidence-dominated.

Example: "Data from the past ten years suggests that there is a 2% change of rain in Morelia today between 2 and 3 PM."

Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Informative prior

Specific information about parameter(s). Progressively increasing amounts of data $\Longrightarrow$ posterior is evidence-dominated.

Example: "Data from the past ten years suggests that there is a 2% change of rain in Morelia today between 2 and 3 PM."

Non-informative prior

Vague information about parameters, typically based on general principles/objective information (also called objective prior). "Light" modification to observations $\Longrightarrow$ posterior is likelihood-dominated.

Example: "The flux from this star is non-negative" ($0 \leq F < \infty$).

Priors

(from Ivezić et al.)

In terms of information, priors can be informative or "non-informative".

Informative prior

Specific information about parameter(s). Progressively increasing amounts of data $\implies$ posterior is evidence-dominated.

Example: "Data from the past ten years suggests that there is a 2% change of rain in Morelia today between 2 and 3 PM."

Non-informative prior

Vague information about parameters, typically based on general principles/objective information (also called objective prior). "Light" modification to observations $\implies$ posterior is likelihood-dominated.

Example: "The flux from this star is non-negative" ($0 \leq F < \infty$).

Improper prior: prior distribution function doesn't integrate to unity.
However, we are still OK if the resulting posterior is well-defined.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

The prior for a scale parameter should ideally be independent of the choice of units.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

If $y = \theta + c$, $\pi(y) = \pi(\theta(y))\,\dfrac{d\theta}{dy} = \pi(\theta(y)) = A(y - c)^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = 0$.

The prior for a scale parameter should ideally be independent of the choice of units.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

If $y = \theta + c$, $\pi(y) = \pi(\theta(y)) \dfrac{d\theta}{dy} = \pi(\theta(y)) = A(y - c)^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = 0$.

$\Longrightarrow$ the non-informative prior for location parameters is the Uniform distribution.

The prior for a scale parameter should ideally be independent of the choice of units.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

If $y = \theta + c$, $\pi(y) = \pi(\theta(y)) \dfrac{d\theta}{dy} = \pi(\theta(y)) = A(y - c)^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = 0$.

$\Longrightarrow$ the non-informative prior for location parameters is the Uniform distribution.

The prior for a scale parameter should ideally be independent of the choice of units.

If $y = c\theta$, $\pi(y) = \pi(\theta(y)) \dfrac{d\theta}{dy} = \dfrac{1}{c}\pi(\theta(y)) = Ac^{-(k+1)}y^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = -1$.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

If $y = \theta + c$, $\pi(y) = \pi(\theta(y))\,\dfrac{d\theta}{dy} = \pi(\theta(y)) = A(y-c)^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = 0$.

$\implies$ the non-informative prior for location parameters is the Uniform distribution.

The prior for a scale parameter should ideally be independent of the choice of units.

If $y = c\theta$, $\pi(y) = \pi(\theta(y))\,\dfrac{d\theta}{dy} = \dfrac{1}{c}\pi(\theta(y)) = Ac^{-(k+1)}y^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = -1$.

$\implies$ the non-informative prior for scale parameters is inversely proportionate to the parameter value.

Non-informative priors for location and scale parameters

Let θ be a parameter with prior distribution $\pi(\theta) = A\theta^k$.

The prior for a location parameter should ideally be robust against translation.

If $y = \theta + c$, $\pi(y) = \pi(\theta(y))\,\dfrac{d\theta}{dy} = \pi(\theta(y)) = A(y-c)^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = 0$.

$\Longrightarrow$ the non-informative prior for location parameters is the Uniform distribution.

The prior for a scale parameter should ideally be independent of the choice of units.

If $y = c\theta$, $\pi(y) = \pi(\theta(y))\,\dfrac{d\theta}{dy} = \dfrac{1}{c}\pi(\theta(y)) = Ac^{-(k+1)}y^k$.

We want $\pi(y)$ to have the same form as $\pi(\theta)$. This is only possible if $k = -1$.

$\Longrightarrow$ the non-informative prior for scale parameters is inversely proportionate to the parameter value.

Example: for data drawn from a Gaussian with unknown μ and σ,

$\pi(\mu) = \mathrm{Uniform}(-\infty, \infty)$ (improper, non-informative prior).

$\pi(\sigma) \propto \dfrac{1}{\sigma}$, with $\sigma \in (0, \infty)$ (improper, non-informative prior).

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0, 1)$ so that the prior mean is $1/2$ (expected for a fair coin).

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0,1)$ so that the prior mean is 1/2 (expected for a fair coin).

Likelihood of obtaining s heads: $\mathscr{L}(\theta) \propto \theta^s(1-\theta)^{N-s}$. (Beta distribution!)

Posterior $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)\pi(\theta) \propto \theta^s(1-\theta)^{N-s} =$

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0, 1)$ so that the prior mean is 1/2 (expected for a fair coin).

Likelihood of obtaining s heads: $\mathscr{L}(\theta) \propto \theta^s (1-\theta)^{N-s}$. (Beta distribution!)

Posterior $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)\pi(\theta) \propto \theta^s (1-\theta)^{N-s} = \mathrm{Beta}(\alpha, \beta)$,

What are α and β?

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0,1)$ so that the prior mean is 1/2 (expected for a fair coin).

Likelihood of obtaining s heads: $\mathscr{L}(\theta) \propto \theta^s (1-\theta)^{N-s}$. (Beta distribution!)

Posterior $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)\pi(\theta) \propto \theta^s (1-\theta)^{N-s} = \mathrm{Beta}(\alpha, \beta)$,

What are α and β? $\quad \alpha = s + 1,\ \beta = N - s + 1$.

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0, 1)$ so that the prior mean is 1/2 (expected for a fair coin).

Likelihood of obtaining s heads: $\mathscr{L}(\theta) \propto \theta^s(1-\theta)^{N-s}$. (Beta distribution!)

Posterior $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)\pi(\theta) \propto \theta^s(1-\theta)^{N-s} = \mathrm{Beta}(\alpha, \beta)$,

What are α and β? $\alpha = s + 1, \ \beta = N - s + 1$.

Posterior mean $\bar{\theta} = \dfrac{\alpha}{\alpha + \beta} =$

Example from Wasserman's "All of Statistics"

A coin has an unknown probability θ of coming down heads. Flipping the coin N times, we observe s heads. Find the posterior distribution of θ.

Non-informative prior $\pi(\theta) = U(0,1)$ so that the prior mean is 1/2 (expected for a fair coin).

Likelihood of obtaining s heads: $\mathscr{L}(\theta) \propto \theta^s (1-\theta)^{N-s}$. (Beta distribution!)

Posterior $p(\theta|\mathrm{data}) = \mathscr{L}(\theta)\pi(\theta) \propto \theta^s(1-\theta)^{N-s} = \mathrm{Beta}(\alpha, \beta)$,

What are α and β? $\alpha = s+1,\ \beta = N-s+1$.

Posterior mean $\bar{\theta} = \dfrac{\alpha}{\alpha+\beta} = \dfrac{s+1}{N+2}$.

Rearrange the above:

$$\bar{\theta} = \frac{s+1}{N+2} = \frac{s}{N+2} + \frac{1}{N+2} = \underbrace{\frac{s}{N}}_{\text{data mean}} \times \frac{N}{N+2} + \underbrace{\frac{1}{2}}_{\text{prior mean}} \times \frac{2}{N+2}$$

The posterior mean is thus the weighted average of the data mean and the prior mean. The effective sample size is then $N+2$.

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Example: inferring the true (photon) count rate from a faint source.
Observe a faint source once, get photon count rate of $S_{\mathrm{obs}} = 4\ \mathrm{s}^{-1}$.
Based on this observation, get constraint on the true photon count rate S from source.

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Example: inferring the true (photon) count rate from a faint source.
Observe a faint source once, get photon count rate of $S_{\mathrm{obs}} = 4\ \mathrm{s}^{-1}$.
Based on this observation, get constraint on the true photon count rate S from source.

If photon-count distribution from sources in the Universe were uniform (uniform prior),

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Example: inferring the true (photon) count rate from a faint source.
Observe a faint source once, get photon count rate of $S_{\rm obs} = 4\ {\rm s}^{-1}$.
Based on this observation, get constraint on the true photon count rate S from source.

If photon-count distribution from sources in the Universe were uniform (uniform prior), photon-counting uncertainty will scatter values symmetrically around population mean $\Longrightarrow$ 95% CI from data nicely constrains true count rate.

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Example: inferring the true (photon) count rate from a faint source.
Observe a faint source once, get photon count rate of $S_{\rm obs} = 4\ \mathrm{s}^{-1}$.
Based on this observation, get constraint on the true photon count rate S from source.

If photon-count distribution from sources in the Universe were uniform (uniform prior), photon-counting uncertainty will scatter values symmetrically around population mean $\Longrightarrow$ 95% CI from data nicely constrains true count rate.

However, there are way more faint sources in the Universe.

Euclidean space: $\dfrac{dN}{dS} \equiv p(S) \propto S^{-5/2}$ (steep prior, small intersection with likelihood).

Prior-dominated posterior

(from Andreon & Weaver, "Bayesian Methods for the Physical Sciences")

The prior can drive the posterior away from the data (likelihood) if it is steep and/or has very little overlap with the region where the likelihood dominates.

Example: inferring the true (photon) count rate from a faint source.
Observe a faint source once, get photon count rate of $S_{\rm obs} = 4\ {\rm s}^{-1}$.
Based on this observation, get constraint on the true photon count rate S from source.

If photon-count distribution from sources in the Universe were uniform (uniform prior), photon-counting uncertainty will scatter values symmetrically around population mean $\Longrightarrow$ 95% CI from data nicely constrains true count rate.

However, there are way more faint sources in the Universe.

Euclidean space: $\dfrac{dN}{dS} \equiv p(S) \propto S^{-5/2}$ (steep prior, small intersection with likelihood).

$\Longrightarrow$ more likely that a lower photon count gets observed as a higher value due to Poisson uncertainty.

This is a form of Eddington Bias.

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\rm obs}} \exp\left[-S\right] = S^4 \exp\left[-S\right]$.

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\rm obs}} \exp\Big[-S\Big] = S^4 \exp\Big[-S\Big]$.

Posterior $p(S|S_{\rm obs}) \propto S^{3/2} \exp\Big[-S\Big]$

$$= \mathrm{Gamma}\left(\frac{5}{2}, 1\right).$$

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\rm obs}} \exp\Big[-S\Big] = S^4 \exp\Big[-S\Big]$.

$$\text{Posterior } p(S|S_{\rm obs}) \propto S^{3/2} \exp\Big[-S\Big]$$
$$= \mathrm{Gamma}\left(\frac{5}{2}, 1\right).$$

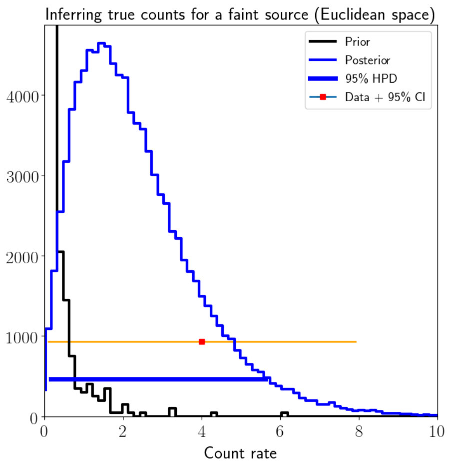
$\Longrightarrow$ Mean: 5/2; Mode: 3/2.
Can also compute HPD.

Prior-dominated posterior (contd.)

Prior: $p(S) \propto S^{-5/2}$.
Likelihood of obtaining data $S_{\rm obs} = 4$ from Poissonian uncertainties acting on S:
$\mathscr{L}(S) \propto S^{S_{\rm obs}} \exp\Big[-S\Big] = S^4 \exp\Big[-S\Big]$.

Posterior $p(S|S_{\rm obs}) \propto S^{3/2} \exp\Big[-S\Big]$

$$= \mathrm{Gamma}\left(\frac{5}{2}, 1\right).$$



$\Longrightarrow$ Mean: 5/2; Mode: 3/2.
Can also compute HPD.

Example of prior-dominated posterior

from Andreon, "Bayesian Methods for the Physical Sciences".

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Example of prior-dominated posterior

from Andreon, “Bayesian Methods for the Physical Sciences”.

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Example of prior-dominated posterior

from Andreon, “Bayesian Methods for the Physical Sciences”.

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Observation: $\log M/M_{\odot} = 14.6 \pm 0.3$.

Example of prior-dominated posterior

from Andreon, "Bayesian Methods for the Physical Sciences".

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Observation: $\log M/M_{\odot} = 14.6 \pm 0.3$.
Prior: Schechter mass function.

Example of prior-dominated posterior

from Andreon, "Bayesian Methods for the Physical Sciences".

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Observation: $\log M/M_\odot = 14.6 \pm 0.3$.
Prior: Schechter mass function.

Prior changes drastically near observed value, similar to previous example.

Example of prior-dominated posterior

from Andreon, "Bayesian Methods for the Physical Sciences".

Andreon et al. 2009 mass measurement for most distant ($z \geq 2$) galaxy cluster, JKCS041.

Mass estimate important to constrain parameters of ΛCDM model.

Observation: $\log M/M_\odot = 14.6 \pm 0.3$.
Prior: Schechter mass function.

Prior changes drastically near observed value, similar to previous example.

Posterior mean is therefore lower than observed value: $\log M/M_\odot = 14.3 \pm 0.3$.
(lower by 2x!)