ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

Bayesian Statistical Inference: I

Dr. Nina Hernitschek
February 28, 2022

Frequentist vs. Bayesian Statistical Inference

There are two major statistical paradigms that address the statistical inference questions:

Key differences	**classical (frequentist) paradigm**	**Bayesian paradigm**
Definition of probabilities:	relative frequency of events over repeated experimental trials	probabilities quantify our subjective belief about experimental outcomes, model parameters, or models
Quantifying uncertainty:	confidence levels describe the distribution of the measured parameter from the data around the true value	credible regions derived from posterior probabilitiy distributions encode our belief in model parameters

Frequentist vs. Bayesian Statistical Inference

we can summarize this as

Frequentist $\theta \to X : P(X|\theta)$

Statistics **Inference**

Bayesian $X \to \theta : P(\theta|X)$

Bayesian Statistical Inference

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

Bayesian Statistical Inference

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

The Bayesian concept of probability is also more **conditional**. In addition to experiment data to predict probabilities (as in the frequentist case), it also uses **prior knowledge**.

Bayesian Statistical Inference

With Bayesian statistics, probability expresses a **degree of belief in an event**. This method is different from the frequentist methodology in a number of ways. One of the big differences is that probability actually expresses the chance of an event happening.

The Bayesian concept of probability is also more **conditional**. In addition to experiment data to predict probabilities (as in the frequentist case), it also uses **prior knowledge**.

example: Measuring the flux of a star.

Bayesian Statistical Inference

example: Measuring the flux of a star.
repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Bayesian Statistical Inference

example: Measuring the flux of a star.
repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Frequentist: probability only has meaning in terms of a limiting case of repeated measurements

Limit of large numbers: the frequency of any given value indicates the probability of measuring that value.
$\Rightarrow$ probabilities fundamentally related to frequencies of events

Bayesian Statistical Inference

example: Measuring the flux of a star.
repeated measurements of flux (from nonvariable star) lead to different values due to the statistical error of the astronomical instrument

Frequentist: probability only has meaning in terms of a limiting case of repeated measurements

Limit of large numbers: the frequency of any given value indicates the probability of measuring that value.
$\Rightarrow$ probabilities fundamentally related to frequencies of events

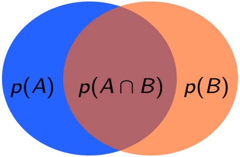
Bayesian:
the concept of probability is extended to cover degrees of certainty about statements.

Bayesian approach claims to measure the flux F with a probability $P(F)$: probability as statement of the knowledge of the measurement outcome.
$\Rightarrow$ probabilities fundamentally related to our own knowledge about an event, the **prior**

Bayes' Rule

recap from lecture 3:

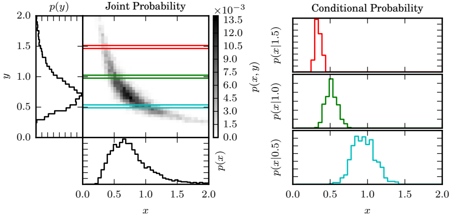
If we have two events, A and B, the possible combinations are:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$p(A \cap B) = p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

We then had seen that the **marginal probability** (projecting onto one axis) is defined as

$$p(x) = \int p(x, y)dy$$
$$= \int p(x|y)p(y)dy$$



Bayes' Rule

Since $p(x|y)p(y) = p(y|x)p(x)$ we can write that

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which gives

Bayes' Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which in words says that

the (conditional) probability of y given x is just the (conditional) probability of x given y times the (marginal) probability of y divided by the (marginal) probability of x, where the latter is just the integral of the numerator.

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.
- Inferences are made by producing probability density functions (pdfs); most notably, model parameters are treated as random variables.

The Bayesian Method

The Essence of the Bayesian Method:

- **Probability statements** are not limited to data, but can be made **for model parameters** and models themselves.
- Inferences are made by producing probability density functions (pdfs); most notably, model parameters are treated as random variables.
- These pdfs represent our belief spread in what the model parameters are. They have nothing to do with outcomes of repeated experiments (although the shape of resulting distributions can often coincide).

Bayesian Statistical Inference

frequentist statistical inference:
We calculated a **likelihood** $p(D \mid M)$.

Bayesian statistical inference:
We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

with data D and model $M = M(\theta)$.

Bayesian Statistical Inference

frequentist statistical inference:
We calculated a **likelihood** $p(D \mid M)$.

Bayesian statistical inference:
We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

We've seen that Bayes' Rule is:

$$p(M \mid D) = \frac{p(D \mid M)\, p(M)}{p(D)},$$

with data D and model $M = M(\theta)$.

Bayesian Statistical Inference

frequentist statistical inference:
We calculated a **likelihood** $p(D \mid M)$.

Bayesian statistical inference:
We instead evaluate the **posterior probability** taking into account prior information and the likelihood.

We've seen that Bayes' Rule is:

$$\underbrace{p(M \mid D)}_{\text{posterior}} = \frac{\overbrace{p(D \mid M)}^{\text{likelihood}}\,\overbrace{p(M)}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}},$$

with data D and model $M = M(\theta)$.

prior probability
How probable are the possible values of θ in nature?

likelihood
ties the model to the data: how likely is the data given θ?

posterior probability
distribution is updated with information from the data: what is the probability of different θ values given data and model?

Bayesian Statistical Inference

If we **explicitly recognize prior information**, I, and the model parameters, θ, then we can write:

$$p(M, \theta \mid D, I) = \frac{p(D \mid M, \theta, I)\, p(M, \theta \mid I)}{p(D \mid I)},$$

where we will omit the explict dependence on θ by writing M instead of M, θ where appropriate. However, as the prior can be expanded to

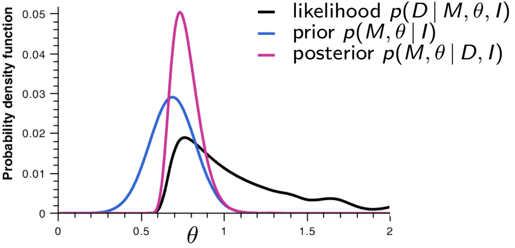
$$p(M, \theta \mid I) = p(\theta \mid M, I)\, p(M \mid I),$$

it will still appear in the term $p(\theta \mid M, I)$.

Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

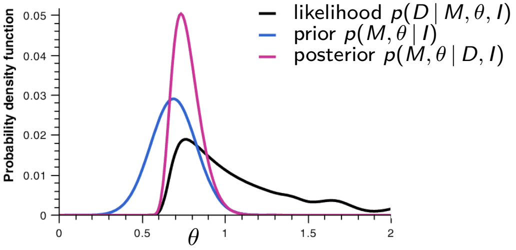
1. formulate the likelihood, $p(D \mid M, \theta, I)$



Bayesian Statistical Inference

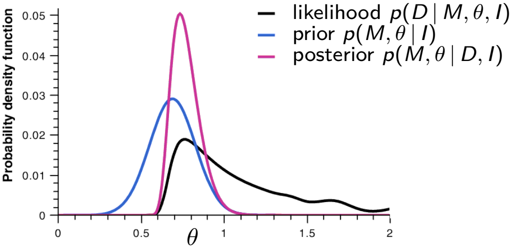
The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D \mid M, \theta, I)$
2. chose a prior, $p(M, \theta \mid I)$, which incorporates other information beyond the data in D



The **Bayesian Statistical Inference process** is then

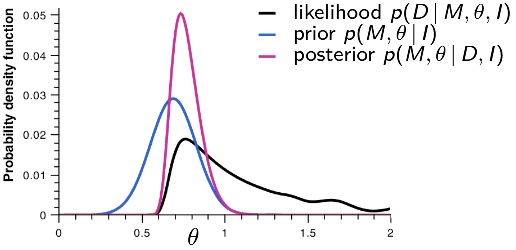
1. formulate the likelihood, $p(D \mid M, \theta, I)$
2. chose a prior, $p(M, \theta \mid I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta \mid D, I)$



Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

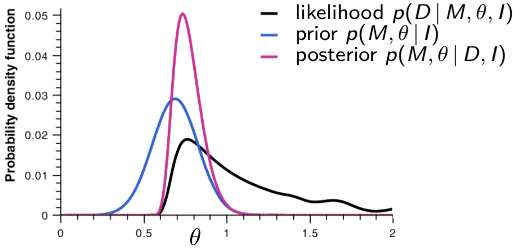
1. formulate the likelihood, $p(D \mid M, \theta, I)$
2. chose a prior, $p(M, \theta \mid I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta \mid D, I)$
4. search for the model parameters that maximize $p(M, \theta \mid D, I)$



Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

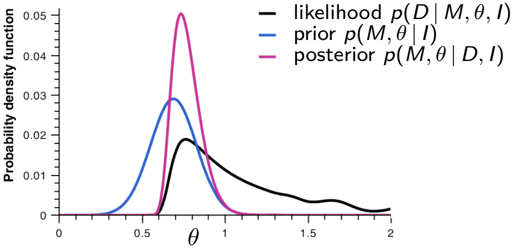
1. formulate the likelihood, $p(D \mid M, \theta, I)$
2. chose a prior, $p(M, \theta \mid I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta \mid D, I)$
4. search for the model parameters that maximize $p(M, \theta \mid D, I)$
5. quantify the uncertainty of the model parameter estimates



Bayesian Statistical Inference

The **Bayesian Statistical Inference process** is then

1. formulate the likelihood, $p(D \mid M, \theta, I)$
2. chose a prior, $p(M, \theta \mid I)$, which incorporates other information beyond the data in D
3. determine the posterior pdf, $p(M, \theta \mid D, I)$
4. search for the model parameters that maximize $p(M, \theta \mid D, I)$
5. quantify the uncertainty of the model parameter estimates
6. perform model selection to find best description of the data



Bayesian Statistical Inference - Priors

Priors can be **informative** or **uninformative**.

based on existing information (including previously obtained data)

"default" priors, i.e. what your prior is when you never saw any data, i.e. a flat pior $p(\theta|M, I) \propto \mathrm{C}$ or a cut-off indicating distances are ≥ 0

Bayesian Statistical Inference - Priors

Priors can be **informative** or **uninformative**.

based on existing information (including previously obtained data)

"default" priors, i.e. what your prior is when you never saw any data, i.e. a flat pior $p(\theta|M, I) \propto \mathrm{C}$ or a cut-off indicating distances are ≥ 0

priors should be (at most) **weakly informative**:

example: Setting the prior distribution for the temperature at noon on a day at a place on Earth to a normal distribution with mean 50 degrees Fahrenheit and standard deviation 40 degrees to constrain the temperature to a reasonable range with a very small chance of being below or above.

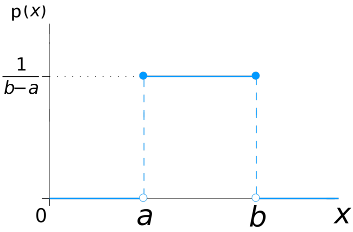
The purpose of a weakly informative prior is for **regularization**, that is, to keep inferences in a reasonable range.

There are three main principles used to choose a prior:

(i) The Principle of Indifference

Essentially this means adopting a uniform prior.

example: assuming 1/2 for heads and tails of a fair coin



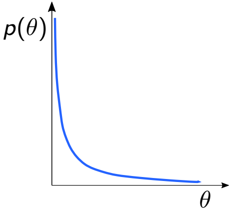
Bayesian Statistical Inference - Priors

There are three main principles used to choose a prior:

(ii) The Principle of Invariance (or Consistency)
This applies to location and scale invariance.

Location invariance suggests a uniform prior, within the accepted bounds: $p(\theta|I) \propto 1/(\theta_{max} - \theta_{min})$ for $\theta_{min} \leq \theta \leq \theta_{max}$.

Scale invariance gives us priors that look like $p(\theta|I) \propto 1/\theta$, which implies a uniform prior for $\ln(\theta)$, i.e. a prior that gives equal weight over many orders of magnitude.



Bayesian Statistical Inference - Priors

There are three main principles used to choose a prior:

(iii) The Principle of Maximum Entropy
Take precisely stated prior data or testable information about a probability distribution function. Consider the **set of all trial probability distributions** that would encode the prior data. According to this principle, the distribution with maximal information entropy is the best choice.

Bayesian Statistical Inference - Priors

There are three main principles used to choose a prior:

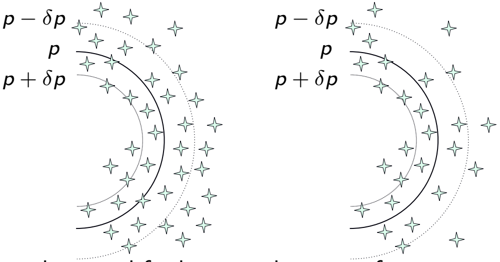
(iii) The Principle of Maximum Entropy
Take precisely stated prior data or testable information about a probability distribution function. Consider the **set of all trial probability distributions** that would encode the prior data. According to this principle, the distribution with maximal information entropy is the best choice.

Since the distribution with the largest entropy is the one that makes the fewest assumptions about the true distribution of data, the principle of maximum entropy can be seen as an application of **Occam's razor**.

Bayesian Statistical Inference - Priors

Considering the priors can have significant consequences: two **examples** from astronomy:

a) **Lutz-Kelker bias** (Lutz & Kelker 1973, PASP, 85, 573)

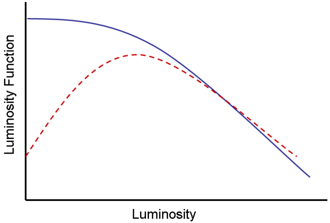


Both stars closer and farther may, because of measurement uncertainty $\pm\delta p$, appear at a given parallax. Assuming uniform stellar distribution in space, the probability density of the true parallax per unit range of parallax will be proportional to $1/p^4$ (where p is the true parallax). Therefore, there will be more stars in the volume shells at farther distance. As a result, more stars will have their true parallax smaller than the observed parallax, and the measured parallax will be systematically biased towards a larger value.

Bayesian Statistical Inference - Priors

Considering the priors can have significant consequences: two **examples** from astronomy:

b) **Malmquist bias** (Malmquist 1922, 1936)
The mean absolute magnitude of observed sample is brighter than the mean absolute magnitude of the population. This bias is caused by astronomical surveys usually being magnitude-limited.



The dashed red line gives a luminosity function when the Malmquist bias is not corrected for. The more numerous low luminosity objects are underrepresented. The solid blue line is the properly corrected luminosity function using the volume-weighted correction method.

Bayesian Statistical Inference - Priors

In special combinations of priors and likelihood functions, the resulting posterior probability distribution is from the same function family as the prior. These **conjugate priors** and give a convenient way for generalizing computations.

Bayesian Statistical Inference - Priors

In special combinations of priors and likelihood functions, the resulting posterior probability distribution is from the same function family as the prior. These **conjugate priors** and give a convenient way for generalizing computations.

example: the conjugate prior for a Gaussian likelihood is a Gaussian, which means: Gaussian likelihood & Gaussian prior $\Rightarrow$ Gaussian posterior

For data drawn from a Gaussian likelihood equal to $\mathcal{N}(\bar{x}, s)$ (where $\bar{x}$ is the sample mean and s is the sample standard deviation), with a prior on the underlying parameters $\mathcal{N}(\mu_p, \sigma_p)$, the posterior is $\mathcal{N}(\mu^0, \sigma^0)$, where

$$\mu^0 = \frac{\mu_p/\sigma_p^2 + \bar{x}/s^2}{1/\sigma_p^2 + 1/s^2}, \quad \sigma^0 = \left(1/\sigma_p^2 + 1/s^2\right)^{-1/2}$$

Bayesian Credible Regions

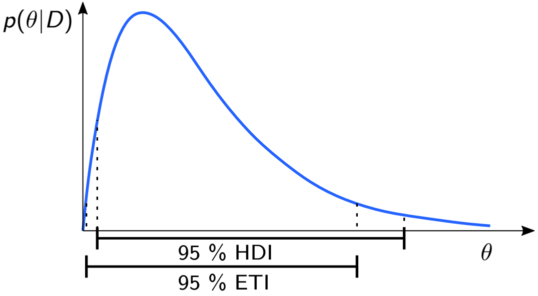
frequentist paradigm: the **confidence interval** $\mu_0 \pm \sigma_\mu$ is the interval containing the true μ (from which the data were drawn) in 68% (or X%) cases of a large number of imaginary repeated experiments (each with a different N values of $\{x_i\}$).

Bayesian paradigm: the **Bayesian credible region** is the interval that contains the true μ with a probability of 68% (or X%), given the given dataset (no imaginary experiments in Bayesian paradigm).

Bayesian Credible Regions

Credible regions can be computed in two different ways:

i) highest (probability) density interval (HDI): integrate downwards from the MAP to enclose $X\%$, or
ii) equal-tailed interval (ETI): integrate inwards from each tail by $(X/2)\%$



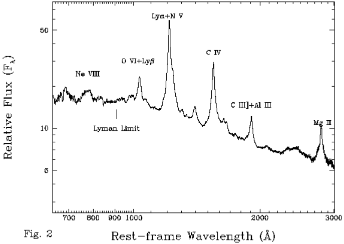
A skewed distribution has different 95% highest density interval (HDI) than 95% equal-tailed interval (ETI).

Gaussian and Uniform Distribution

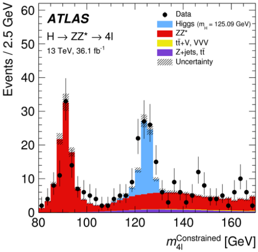
usually distributions aren't exactly Gaussians, but can be modeled as superpositions of Gaussians, uniform distributions...

example: Gaussian distribution embedded in a uniform background distribution

spectral lines superimposed upon a background:



Higgs boson peak embedded in background noise and other particles:



⇒ such distributions are common in physics and astronomy

Gaussian and Uniform Distribution

We assume that

- the location parameter, μ, is known (say from theory) and
- the uncertainties in x_i are negligible compared to σ.

Gaussian and Uniform Distribution

We assume that

- the location parameter, μ, is known (say from theory) and
- the uncertainties in x_i are negligible compared to σ.

The likelihood of obtaining a single measurement, x_i, can be written as a probabilistic mixture of either the Gaussian or the uniform distribution:

$$p(x_i|A,\mu,\sigma,I) = \frac{A}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) + \frac{1-A}{W}.$$

in detail:

- Here the background probability is taken to be $0 < x < W$ and 0 otherwise.
- The feature of interest lies between 0 and W.
- A and $1-A$ are the relative strengths of the two components, which are obviously anti-correlated.
- Note that there will be covariance between A and σ.

Gaussian and Uniform Distribution

The likelihood of obtaining a single measurement, x_i, can be written as a probabilistic mixture of either the Gaussian or the uniform distribution:

$$p(x_i|A, \mu, \sigma, I) = \frac{A}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) + \frac{1-A}{W}.$$

We adopt a uniform prior in both A and σ:

$$p(A, \sigma|I) = C, \text{ for } 0 \leq A < A_{\max} \text{ and } 0 \leq \sigma \leq \sigma_{\max}.$$

The posterior pdf is then given by

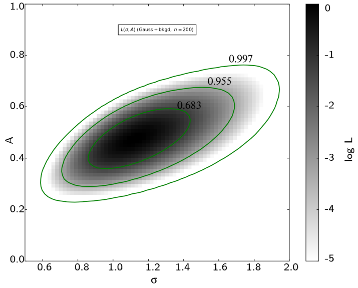
$$\begin{aligned}\log L &= \ln[p(A, \sigma|\{x_i\}, \mu, W)] \\ &= \sum_{i=1}^{N} \ln\left[\frac{A}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) + \frac{1-A}{W}\right].\end{aligned}$$

Gaussian and Uniform Distribution

The posterior pdf is then given by

$$\begin{aligned}\log L &= \ln[p(A, \sigma | \{x_i\}, \mu, W)] \\ &= \sum_{i=1}^{N} \ln \left[\frac{A}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2} \right) + \frac{1-A}{W} \right].\end{aligned}$$

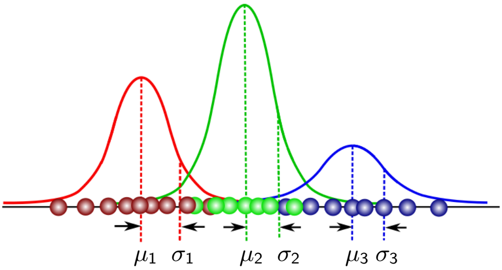
The example below is $\log L$ with $A = 0.5$, $\sigma = 1$, $\mu = 5$, $W = 10$, evaluated on a grid:



Gaussian Mixture Models

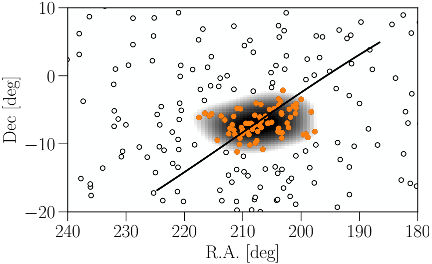
A **Gaussian Mixture** is a function that is comprised of several Gaussians, each identified by $k \in \{1, ..., K\}$ where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters:

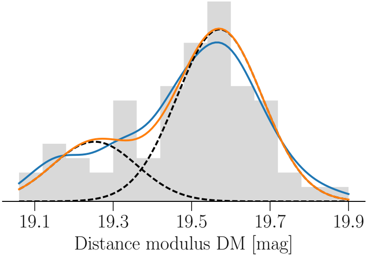
- A mean μ_k that defines its centre.
- A covariance Σ_k that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
- A mixing probability π_k that defines its amplitude.



Gaussian Mixture Models

example:





Spatial distribution of PS1 3π RRab stars in the vicinity of the newly discovered Outer Virgo Overdensity (orange solid circles in the top panel). Density was obtained by running a Gaussian mixture model on the data, optimization with Gaussian kernel density estimation from Python **scikit.learn**.

credit: B. Sesar, N. Hernitschek, M. I. P. Dierickx et al., 2017)

Break & Questions

afterwards we continue with `lecture_6.ipynb` from the `github` repository