In this class we will review Bayesian likelihood methods for solving statistical problems, determining the posterior probabilities of model parameters, and selecting between two models.
At the end of this class you should be able to ...

- ... understand the application of Bayes’ theorem in model-fitting and the role of priors
- ... obtain parameter values and confidence ranges via likelihood methods
- ... search parameter space with MCMC algorithms
- ... apply model selection tests using the Bayes factor or Akaike information criteria
Bayesian Methods

• Recall from Class 1 that Bayesian statistics is a framework that allows us to **assign probabilities to a model**

• It makes use of conditional probabilities, $P(A|B)$, meaning “the probability of $A$ on the condition that $B$ has occurred”

• Remember that $P(A|B) \neq P(B|A)$ in general!
• An important role in Bayesian statistics is played by **Bayes’ theorem**, which can be derived from elementary probability:

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

• Small print: this formula can be derived by just writing down the joint probability of both \( A \) and \( B \) in 2 ways:

\[ P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \]
Bayesian Methods

• The chance of a certain medical test being positive is 90%, if a patient has disease $D$. 1% of the population have the disease, and the test records a false positive 5% of the time. If you receive a positive test, what is your probability of having $D$?

• We are told: $P(+|D) = 0.9$, $P(D) = 0.01$, $P(+|\text{no } D) = 0.05$

• We want to know: $P(D|+)$

• Bayes’ Theorem: $P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\text{no } D)P(\text{no } D)}$

• Substituting in the data: $P(D|+) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.05 \times 0.99} = 0.15$

• Interpretation: although the test is correct 90% of the time, the probability of having $D$ after a positive test is only 15%. This is because only a small fraction of the population have the disease.
A Frequentist might argue “either the person has the disease or not – it is meaningless to apply probability in this way”

A Bayesian might argue “there is a prior probability of 1% that the person has the disease. This probability should be updated in the light of the new data using Bayes’ theorem”
Bayesian Methods

- Bayes’ theorem can be usefully re-written for science as:

\[
P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model}) \cdot P(\text{model})}{P(\text{data})}
\]

**Posterior** probability of the model in light of the data

**Likelihood** function of the data given the model

**Prior** probability of the model

**Evidence** [can typically be absorbed into the normalization of the posterior]
Role of the prior

• Bayesian statistics cannot determine probabilities of a model without assigning a **prior probability**

• The importance of the prior probability is both the strong and weak point of Bayesian statistics

• A **Bayesian** might argue: *"the prior probability is a logical necessity when assessing the probability of a model. It should be stated, and if it’s unknown you can use an uninformative (wide) prior"*

• A **Frequentist** might argue *"setting the prior is subjective – two experiments could use the same data to come to two different conclusions, just by taking different priors"*
Role of the prior

• Let’s take the example of fitting a parameter $a$ to some data. Bayes’ Theorem now reads:

$$P(a|\text{data}) \propto P(\text{data}|a) P(a)$$

• We do not need the denominator, since we will normalize the posterior $P(a|\text{data})$ such that $\int P(a|\text{data}) \, da = 1$

• In the absence of other information, a uniform (or constant) prior is often assumed for $P(a)$. This is effectively equivalent to the fitting range of a parameter

• Assuming Gaussian variables, the likelihood $P(\text{data}|a)$ is:

$$P(\text{data}|a) \propto e^{-\chi^2/2}$$

Hence: $P(a|\text{data}) \propto e^{-\chi^2/2}$
We can use the posterior probability distribution $P(a)$ to determine **summary statistics** and **confidence intervals** for the parameter $a$:

- **Mean:** $\mu_a = \int_{-\infty}^{\infty} a P(a) \, da$

- **Variance:**
  
  $\sigma_a^2 = \int_{-\infty}^{\infty} (a - \mu_a)^2 P(a) \, da$

- [Small print: only if the probability distribution is Gaussian is the mean equal to the best-fitting value, and the standard deviation equal to the 68% confidence region]
Posteriors and confidence limits

- For a general probability distribution, we can determine the confidence intervals by integration:

\[
\int_{-\infty}^{a_{\text{bot}}} P(a) \, da = 0.16 \\
\int_{a_{\text{bot}}}^{a_{\text{top}}} P(a) \, da = 0.68 \\
\int_{a_{\text{top}}}^{\infty} P(a) \, da = 0.16
\]
Marginalization

• Now suppose we have determined the 2D posterior probability distribution of a 2-parameter fit, $P_{2D}(a, b)$

• What is the probability distribution for parameter $a$, considering all possible values of parameter $b$? This is known as marginalization of parameter $b$

• Marginalization can be performed by summing (integrating) over one axis of the probability distribution:

$$P_{1D}(a) = \sum_b P_{2D}(a, b)$$

• [Small print: if $P_{2D}(a, b)$ is normalized, then $P_{1D}(a)$ will also be normalized]
Let’s apply these methods to our example from Class 3, fitting a straight line $y = ax + b$ to some data...
• We determine the values of $\chi^2$ over a grid of $(a, b)$ and convert to 2D probability $P(a, b) \propto e^{-\chi^2/2}$
• Then we marginalize to obtain the posterior probability distributions for each parameter, $P(a)$ and $P(b)$ ...
Use of likelihood for parameter fitting

- By integrating under these distributions, we identify the 68% confidence regions ...

![Graphs showing marginalized posterior probability distributions for parameters a and b.](image)
Let’s return to the same supernova distance-redshift dataset we were using in Class 3:

- Convert the $\chi^2$ values into a joint 2D probability distribution in $(\Omega_m, \Omega_\Lambda)$

- Marginalize this probability distribution to obtain the 1D posterior probability distributions for $\Omega_m$ and $\Omega_\Lambda$

- Determine the 68% confidence regions for $\Omega_m$ and $\Omega_\Lambda$
• The grid method becomes inefficient as the number of parameters increases. A powerful alternative is to generate a Monte Carlo Markov Chain (MCMC) in the parameter space.

• There are various algorithms to do this such as python emcee (we won’t go into details here), but the end result is a “chain” (distribution of parameter values) which samples the underlying probability distribution.

Image credit: www.essenceps.com
Image credit: www.newton.ac.uk
Here is a worked example of using python’s emcee algorithm to sample the probability distribution of the straight-line fit:
• Let’s return to the same *supernova distance-redshift dataset* again:

• Run an **MCMC analysis** for parameters \((\Omega_m, \Omega_\Lambda)\)

• Determine the **68% confidence regions** for \(\Omega_m\) and \(\Omega_\Lambda\)
Since Bayesian statistics is related to the probability of models, it allows us to perform **model selection**

A common example: how many model parameters does a dataset justify including in a fit?
In general, given models $M_1$ (parameter $p_1$) and $M_2$ (parameter $p_2$) and a dataset $D$, we can determine the Bayes factor:

$$K = \frac{P(M_1|D)}{P(M_2|D)} = \frac{\int dp_1 P(D|p_1) P(p_1)}{\int dp_2 P(D|p_2) P(p_2)}$$

The size of $K$ quantifies how strongly we can prefer one model to another, e.g. the Jeffreys scale:

<table>
<thead>
<tr>
<th>$K$</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 − 3</td>
<td>“barely worth mentioning”</td>
</tr>
<tr>
<td>3 − 10</td>
<td>“substantial”</td>
</tr>
<tr>
<td>10 − 30</td>
<td>“strong”</td>
</tr>
<tr>
<td>&gt; 30</td>
<td>“very strong”</td>
</tr>
</tbody>
</table>
Model selection

• This quantity is usually difficult to compute, and we can instead use an approximation to this ratio.

• A common approach is to calculate the Akaike information criteria for each model:

$$AIC = \chi^2_{\text{min}} + 2p + \frac{2p(p+1)}{N-p-1}$$

- $p = \text{number of parameters}$
- $N = \text{number of bins}$

• This penalizes models with more parameters (and the final term corrects for sample size).

• The model with the smaller value of AIC is preferred [the likelihood ratio is $e^{(AIC_1-AIC_2)/2}$]
Flat or curved Universe?

• Let’s return to the same **supernova distance-redshift dataset** again:

• Compute the **Akaike information criteria** for a **flat model** (where $\Omega_m + \Omega_\Lambda = 1$) and a **curved model** (where $\Omega_m$, $\Omega_\Lambda$ can take any value). *Which model is preferred, by this metric?*
At the end of this class you should be able to ...

• ... understand the application of Bayes’ theorem in model-fitting and the role of priors

• ... obtain parameter values and confidence ranges via likelihood methods

• ... search parameter space with MCMC algorithms

• ... apply model selection tests using the Bayes factor or Akaike information criteria