Bayesian data analysis is based on the following two principles:

1. Probability is interpreted as a measure of uncertainty, *whatever the source*. Thus, in a Bayesian analysis, it is standard practice to assign probability distributions not only to unseen data, but also to parameters, models, and hypotheses.

2. Uncertainty is quantified both before and after the collection of data and Bayes’ formula is used to update our beliefs in light of the new data.
Suppose that our objective is to use some newly acquired data $D$ to estimate the value of an unknown parameter $\Theta$. A Bayesian treatment of this problem would proceed as follows:

1. We first need to formulate a **statistical model** which determines the conditional distribution of the data under each possible value of $\Theta$. This is specified by the **likelihood function**, $p(D|\Theta = \theta)$.

2. We then choose a **prior distribution**, $p(\Theta = \theta)$, for the unknown parameter which quantifies how strongly we believe that the true value is $\theta$ before we examine the new data.

3. We then collect and examine the new data $D$.

4. In light of this new data, we use Bayes’ formula to revise our beliefs concerning the value of the parameter $\Theta$:

   $$
   p(\Theta = \theta|D) = p(\Theta = \theta) \cdot \left( \frac{p(D|\Theta = \theta)}{p(D)} \right)
   $$

   $p(\Theta = \theta|D)$ is said to be the **posterior distribution** of the parameter $\Theta$ given the data $D$. 
Example: Bayesian Estimation of Sex Ratios

Suppose that our objective is to estimate the birth sex ratio of a newly described species. To this end, we will count the numbers of male and female offspring in each of $b$ broods.

For the sake of the example, we will make the following assumptions:

- The probability of being born female does not vary within or between broods. In particular, there is no environmental or heritable variation in sex ratio.
- The sexes of the different members of a brood are determined independently of one another.
- If we let $\theta$ denote the unknown probability that an individual is born female, then the sex ratio at birth will be $\theta/1 - \theta$.
- We will let $m$ and $f$ denote the total numbers of male and female offspring contained in the $b$ broods.
**Likelihood Function:** Under our assumptions about sex determination, the likelihood function depends only on the sex ratio and the total numbers of males and females in the broods. Conditional on $\Theta = \theta$, the total number of female offspring among the $n$ offspring is binomially distributed with parameters $f + m$ and $\theta$:

$$P(f, m|\Theta = \theta) = \binom{f + m}{f} \theta^f (1 - \theta)^m$$

**MLE**

$$\theta_{MLE} = \frac{f}{f + m}$$
**Prior Distribution:** The prior distribution on the unknown parameter $\theta$ should reflect what we have previously learned about the birth sex ratio in this species. For example, this might be determined by prior observations of this species or by information about the sex ratio of closely-related species. Here I will consider prior distributions for three different scenarios:

- **Uniform: no prior information**
  
  $$p(\Theta = \theta) = 1$$

- **Beta(5,5): even sex ratio**
  
  $$p(\Theta = \theta) = 630 \theta^4 (1 - \theta)^4$$

- **Beta(2,4): male-biased**
  
  $$p(\Theta = \theta) = 20 \theta (1 - \theta)^3$$
We are now ready to use Bayes’ formula to calculate the posterior distribution of \( \theta \) conditional on having observed \( f \) female offspring out of a total of \( n \) offspring. When the prior distribution is of type Beta(\( a, b \)), we have

\[
p(\Theta = \theta | f, m) = p(\Theta = \theta) \cdot \left( \frac{P(f, m|\theta)}{P(f, m)} \right) \quad \text{(Bayes’ formula)}
\]

\[
= \frac{1}{\beta(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \cdot \left( \frac{(f+m)\theta^f (1 - \theta)^m}{P(f, m)} \right)
\]

\[
= \frac{1}{\beta(f + a, m + b)} \theta^{f+a-1} (1 - \theta)^{m+b-1},
\]

which shows that the posterior distribution is of type Beta(\( f + a, m + b \)).

**Remark**

Because the prior and the posterior distribution belong to the same family of distributions, we say that the beta distribution is a **conjugate prior** for the binomial likelihood function.
The figures show the posterior distributions corresponding to each of these three priors for two different data sets: either $f = 7, m = 3$ (left plot) or $f = 70, m = 30$ (right plot).
Summaries of the Posterior Distribution

Although the posterior distribution $p(\Theta = \theta | D)$ comprehensively describes the state of knowledge concerning an unknown parameter, it is common to summarize this distribution by a number or an interval.

- **Point estimators** of $\Theta$ include the mean and the median of the posterior distribution.

- If it exists, the mode of the posterior distribution can be used to estimate $\Theta$, in which case it is called the maximum a posteriori (MAP) estimate.

- A **credible region** is a region that contains a specified proportion of the probability mass of the posterior distribution, e.g., a 95% credible region will contain 95% of this mass. These can be chosen in several ways, including:
  - quantile-based intervals;
  - highest probability density (HPD) regions.
Credible Intervals for \( N(0,1) \)

MAP = mean = median = 0

95% median = \((-1.96, 1.96)\)

95% HPD = \((-1.96, 1.96)\)

Credible Intervals for \( \text{Exp}(1) \)

MAP = 0.0

Median = 0.692

Mean = 1.0

95% HPD = (0, 2.996)

95% quantiles = (0.025, 3.69)

MAP = 0, median = 0.692, mean = 1.0

95% quantiles = (0.025, 3.689)

95% HPD = (0, 2.996)
The table lists various summary statistics for the posterior distributions calculated in the sex ratio example.

<table>
<thead>
<tr>
<th>data</th>
<th>prior</th>
<th>mean</th>
<th>median</th>
<th>sd</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>7, 3</td>
<td>uniform</td>
<td>0.667</td>
<td>0.676</td>
<td>0.131</td>
<td>0.390</td>
<td>0.891</td>
</tr>
<tr>
<td></td>
<td>even</td>
<td>0.600</td>
<td>0.603</td>
<td>0.107</td>
<td>0.384</td>
<td>0.798</td>
</tr>
<tr>
<td></td>
<td>$\sigma^+$-biased</td>
<td>0.563</td>
<td>0.565</td>
<td>0.120</td>
<td>0.323</td>
<td>0.787</td>
</tr>
<tr>
<td>70, 30</td>
<td>uniform</td>
<td>0.697</td>
<td>0.697</td>
<td>0.045</td>
<td>0.604</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td>even</td>
<td>0.682</td>
<td>0.683</td>
<td>0.044</td>
<td>0.592</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>$\sigma^+$-biased</td>
<td>0.679</td>
<td>0.680</td>
<td>0.045</td>
<td>0.588</td>
<td>0.764</td>
</tr>
</tbody>
</table>

**Interpretation**

Whereas the small data set does not provide strong evidence against the hypothesis that the sex ratio is 1 : 1, all three analyses of the large data set suggest that the true value of $\theta$ is greater than 0.58.
One of the strengths of Bayesian statistics is that it can readily handle sequentially acquired data. This is done by using the posterior distribution from the most recent experiment as the prior distribution for the next experiment.

\[
p_0(\theta) \xrightarrow{D_1} p_1(\theta) = p_0(\theta) \frac{p(D_1|\theta)}{p(D_1)}
\]

\[
p_1(\theta) \xrightarrow{D_2} p_2(\theta) = p_1(\theta) \frac{p(D_2|\theta)}{p(D_2)}
\]

\[
p_2(\theta) \xrightarrow{D_3} \ldots
\]
Example: Sex Ratio Estimation, continued

Suppose that our prior distribution on $\Theta$ was uniform and that we initially collected three broods totaling 7 females and 3 males. We then used Bayes’ formula to determine that the posterior distribution of $\Theta$ is $\text{Beta}(8, 4)$. Since this distribution is broad, we decide to collect additional data to refine our estimate of $\Theta$.

If the second data set contains 15 females and 5 males, then using $\text{Beta}(8, 4)$ as the new prior distribution, we find that the new posterior distribution of $\Theta$ is

$$p(\Theta = \theta | f = 15, m = 5) = \frac{\theta^7 (1 - \theta)^3}{\beta(8, 4)} \cdot \left( \frac{\binom{20}{15} \theta^{15} (1 - \theta)^5}{\text{P}(f = 15, m = 5)} \right)$$

$$= \frac{\theta^{22} (1 - \theta)^8}{\beta(23, 9)}$$
Choosing the Prior Distribution: Some Guidelines

Choosing a prior distribution is both useful and sometimes difficult because it requires a careful assessment of our knowledge or beliefs before we perform an experiment.

- As a rule, the prior should be chosen independently of the new data.
- **Cromwell’s rule**: The prior distribution should assign positive probability to any proposition that is not logically false, no matter how unlikely.
- It is sometimes useful to carry out multiple Bayesian analyses using different prior distributions to explore the sensitivity of the posterior distribution to different prior assumptions.
- When there is very little prior information, it may be appropriate to choose an "uninformative prior". Examples include maximum entropy distributions and Jeffreys priors.
Bayesian Phylogenetics

Bayesian methods have proven to be especially useful in the analysis of genetic sequence data. In these problems, the unknown parameters can often be divided into three categories:

- the unknown tree, $T$
- the parameters of the demographic model, $\Theta_{dem}$
- the parameters of the substitution model, $\Theta_{subst}$.

Then, given a sequence alignment $D$, the analytical problem is to calculate the posterior distribution of all of the unknowns:

Bayes’ formula for phylogenetic inference, general case

$$p(T, \Theta_{dem}, \Theta_{subst} | D) = p(T, \Theta_{dem}, \Theta_{subst}) \cdot \left( \frac{p(D | T, \Theta_{dem}, \Theta_{subst})}{p(D)} \right)$$
If the substitution process is assumed to be neutral, then the prior distribution and the likelihood functions can be simplified:

Bayes’ formula for phylogenetic inference with neutral data

\[
p(T, \Theta_{\text{subst}}, \Theta_{\text{dem}} | D) = p(\Theta_{\text{subst}}) \cdot p(\Theta_{\text{dem}}) \cdot p(T | \Theta_{\text{dem}}) \cdot \left( \frac{p(D | T, \Theta_{\text{subst}})}{p(D)} \right)
\]

This is a consequence of the following assumptions:

- Under the prior distribution, the parameters of the substitution process \( \Theta_{\text{subst}} \) are independent of the tree \( T \) and the demographic parameters \( \Theta_{\text{dem}} \).
- The demographic parameters \( \Theta_{\text{dem}} \) typically determine the conditional distribution of the genealogy \( T \), e.g., through a coalescent model.
- Conditional on \( T \) and \( \Theta_{\text{subst}} \), the sequence data \( D \) is independent of \( \Theta_{\text{dem}} \).
Example: Bayesian Inference of Effective Population Size

Suppose that our data $D$ consists of $n$ randomly sampled individuals that have been sequenced at a neutral locus and that our objective is to estimate the effective population size $N_e$. For simplicity, we will use the Jukes-Cantor model for the substitution process with a strict molecular clock and we will assume that the demography can be described by the constant population size coalescent.

To carry out a Bayesian analysis, we need to specify $p(\mu)$, $p(N_e)$ and $p(T|N_e)$.

- $p(\mu)$ should be chosen to reflect what we know about the mutation rate, e.g., we could use a lognormal distribution with mean $u$ and variance $\sigma^2$.
- Since $N_e$ is a scale parameter in the coalescent, it is common practice to use the Jeffreys prior $p(N_e) \propto 1/N_e$.
- $p(T|N_e)$ is then determined by Kingman’s coalescent.
Bayesian Analysis

Assuming that we are only interested in $N_e$, then $\mu$ and $T$ are nuisance parameters and so we need to calculate the marginal posterior distribution of $N_e$ by integrating over $\mu$ and $T$:

$$p(N_e|D) = \int \int p(T, N_e, \mu|D) d\mu dT$$

$$= \frac{p(N_e)}{p(D)} \int \int p(D|T, \mu)p(\mu)p(T|N_E)d\mu dT.$$ 

However, unless $n$ is quite small, integration over $T$ is not feasible. For example, if our sample contains 20 sequences, then there are approximately $8 \times 10^{21}$ possible trees to be considered. Even with the fastest computers available, this is an impossible calculation.
Until recently, Bayesian methods were regarded as impractical for many problems because of the computational difficulty of evaluating the posterior distribution. In particular, to use Bayes’ formula,

\[ p(\theta|D) = p(\theta) \cdot \left( \frac{p(D|\theta)}{p(D)} \right), \]

we need to evaluate the marginal probability of the data

\[ p(D) = \int p(D|\theta)p(\theta)d\theta, \]

which requires integration of the likelihood function over the parameter space. Except in special cases, this integration must be performed numerically, but sometimes even this is very difficult.
An alternative is to use Monte Carlo methods to sample from the posterior distribution. Here the idea is to generate a random sample from the distribution and then use the empirical distribution of that sample to approximate $p(\theta|D)$:

$$p(\theta|D) \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{\Theta_i} \quad \text{where} \quad \Theta_1, \cdots, \Theta_N \sim p(\theta|D)$$

For example, the figure shows two histogram estimators for the Beta(8, 4) density generated using either 100 (left) or 1000 (right) independent samples.
In particular, the empirical distribution can be used to estimate probabilities and expectations under $p(\theta|D)$:

$$P(\theta \in A|D) \approx \frac{1}{N} \sum_{i=1}^{N} 1_A(\Theta_i)$$

$$E[f(\theta)|D] \approx \frac{1}{N} \sum_{i=1}^{N} f(\Theta_i).$$

What makes this approach difficult is the need to generate random samples from distributions that are only known up to a constant of proportionality (e.g., $p(D)$). This is where Markov chain Monte Carlo methods come in.
Markov Chains

A discrete-time Markov chain is a stochastic process $X_0, X_1, \cdots$ with the property that the future behavior of the process only depends on its current state. More precisely, this means that for every set $A$ and $t, s \geq 0$,

$$\mathbb{P}(X_{t+s} \in A \mid X_t, X_{t-1}, \cdots, X_0) = \mathbb{P}(X_{t+s} \in A \mid X_t).$$

Markov chains have numerous applications in biology. Some familiar examples include:

- random walks
- branching processes
- the Wright-Fisher model
- chain-binomial models (Reed-Frost)
One consequence of the **Markov property** is that many Markov chains have a tendency to ‘forget’ their initial state as time progresses. More precisely,

**Asymptotic behavior of Markov chains**

Many Markov chains have the property that there is a probability distribution $\pi$, called the **stationary distribution** of the chain, such that for large $t$ the distribution of $X_t$ approaches $\pi$, i.e., for all states $x$ and sets $A$,

$$\lim_{t \to \infty} \mathbb{P}(X_t \in A | X_0 = x) = \pi(A).$$

**Stationary behavior of the Wright-Fisher process:**

($N = 100, \mu = 0.02$)
Some Markov chains satisfy an even stronger property, called **ergodicity**.

**Ergodicity**

A Markov chain with stationary distribution $\pi$ is said to be **ergodic** if for every initial state $X_0 = x$ and every set $A$, we have

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} 1_A(X_t) = \pi(A).$$

In other words, if we run the chain for a long time, then the proportion of time spent visiting the set $A$ is approximately $\pi(A)$.

Ergodic behavior of the Wright-Fisher process:
$$(N = 100, \mu = 0.02)$$
Markov Chain Monte Carlo: General Approach

The central idea in Markov Chain Monte Carlo is to use an ergodic Markov chain to generate a random sample from the target distribution $\pi$.

1. The first step is to select an ergodic Markov chain which has $\pi$ as its stationary distribution. There are different methods for doing this, including the Metropolis-Hastings algorithm and the Gibbs Sampler.

2. We then need to simulate the Markov chain until the distribution is close to the target distribution. This initial period is often called the burn-in period.

3. We continue simulating the chain, but because successive values are highly correlated, it is common practice to only collect a sample every $T$ generations, so as to reduce the correlations between the sampled states (thinning).

4. We can then use these samples to approximate the target distribution:

$$
\frac{1}{N} \sum_{n=1}^{N} \delta_{X_{B+nT}} \approx \pi.
$$
The Metropolis-Hastings Algorithm

Given a probability distribution $\pi$, the Metropolis-Hastings algorithm can be used to explicitly construct a Markov chain that has $\pi$ as its stationary distribution.

Implementation requires the following elements:

- the target distribution, $\pi$, known up to a constant of proportionality;
- a family of proposal distributions, $Q(y|x)$, and a way to efficiently sample from these distributions.

The MH algorithm is based on a more general idea known as rejection sampling. Instead of sampling directly from $\pi$, we propose values using a distribution $Q(y|x)$ that we can easily sample but then we reject values that are unlikely under $\pi$. 
The Metropolis-Hastings algorithm consists of repeated application of the following three steps. Suppose that $X_n = x$ is the current state of the Markov chain. Then the next state is chosen as follows:

**Step 1:** We first propose a new value for the chain by sampling $y$ with probability $Q(y|x)$.

**Step 2:** We then calculate the **acceptance probability** of the new state:

$$
\alpha(x; y) = \min \left\{ \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}, 1 \right\}
$$

**Step 3:** With probability $\alpha(x; y)$, set $X_{n+1} = y$. Otherwise, set $X_{n+1} = x$.

**Remark**

Because $\pi$ enters into $\alpha$ as a ratio $\pi(y)/\pi(x)$, we only need to know $\pi$ up to a constant of proportionality. This is why the MH algorithm is so well suited for Bayesian analysis.
The choice of the proposal distribution $Q$ can have a profound impact on the performance of the MH algorithm. While there is no universal procedure for selecting a ‘good’ proposal distribution, the following considerations are important.

- $Q$ should be chosen so that the chain rapidly converges to its stationary distribution.
- $Q$ should also be chosen so that it is easy to sample from.
- There is usually a tradeoff between these two conditions and sometimes it is necessary to try out different proposal distributions to identify one with good properties.

Many implementations of MH (e.g., BEAST, MIGRATE) offer the user some control over the proposal distribution.
MCMC: Convergence and Mixing

One of the most challenging issues in MCMC is knowing for how long to run the chain. There are two related considerations.

1. We need to run the chain until its distribution is sufficiently close to the target distribution (convergence).

2. We then need to collect a large enough number of samples that we can estimate any quantities of interest (e.g., the mean TMRCA) sufficiently accurately (mixing).

Unfortunately, there is no universally-valid, fool-proof way to guarantee that either one of these conditions is satisfied. However, there are a number of convergence diagnostics that can indicate when there are problems.
Convergence Diagnostics: Trace Plots

**Trace plots** show how the value of a parameter changes over the course of a simulation. In general, what we want to see is that the mean and the variance of the parameter are fairly constant over the duration of the trace plot, as in the two examples shown below.
Problems with convergence or mixing may be revealed by trends or sudden changes in the behavior of the trace plot.

The increasing trend indicates that the chain has not yet converged.  The sudden changes in mean indicate that the chain is poorly mixing.
Trace Plots: Some Guidelines

1. You should examine the trace plot of every parameter of interest, including the likelihood and the posterior probability. **If any of the trace plots look problematic, then all of the results are suspect.**

2. The fact that a trace plot appears to have converged is not conclusive proof that it has. Especially in high-dimensional problems, a chain that appears to be stationary for the first 500 million generations may well show a sudden change in behavior in the next.

3. The program **Tracer** ([http://tree.bio.ed.ac.uk/software/tracer/](http://tree.bio.ed.ac.uk/software/tracer/)) can be used to display and analyze trace plots generated by BEAST, MrBayes and LAMARC.
Convergence Diagnostics: Effective Sample Size

Because successive states visited by a Markov chain are correlated, an estimate derived using \( N \) values generated by such a chain will usually be less precise than an estimate derived using \( N \) independent samples. This motivates the following definition.

**Effective Sample Size (ESS)**

The effective sample size of a sample of \( N \) correlated random variables is equal to the number of independent samples that would estimate the mean with the same variance.

For a stationary Markov chain with autocorrelation coefficients \( \rho_k \), the ESS of \( N \) successive samples is equal to

\[
ESS = \frac{N}{1 + 2 \sum_{k=1}^{\infty} \rho_k}.
\]
Effective Sample Size: Guidelines

1. Each parameter has its own ESS and these can differ between parameters by more than order of magnitude.

2. Parameters with small ESS’s indicate that a chain either has not converged or is slowly mixing. As a rule of thumb, the ESS of every parameter should exceed 1000 and larger values are even better.

3. Thinning by itself will not increase the ESS. However, we can increase the ESS by simultaneously thinning and increasing the duration of the chain, e.g., collecting a 1000 samples from a chain lasting 100000 generations is better than collecting a 1000 samples from a chain lasting 10000 generations.

4. The ESS of a parameter can usually only be estimated from its trace. For this reason, large ESS’s do not guarantee that the chain has converged.
Formal Tests of Convergence

There are several formal tests of stationarity that can be applied to MCMC.

- The **Geweke diagnostic** compares the mean of a parameter estimated from two non-overlapping parts of the chain and tests whether these are significantly different.

- The **Raftery-Lewis diagnostic** uses a pilot chain to estimate the burn-in and chain length required to estimate the $q$'th quantile of a parameter to within some tolerance.

- Both of these methods suffer from the defect that “you’ve only seen where you’ve been” (Robert & Casella, 2004). In other words, these methods cannot detect that the chain has failed to visit part of the support of the target distribution.

- These and other diagnostic tests are implemented in the R package coda.
Convergence Diagnostics: Multiple Chains

Another approach to testing the convergence of a MCMC analysis is to run multiple independent chains and compare the results.

- Large differences between the posterior distributions estimated by the different chains indicate problems with convergence or mixing.
- It is often useful to start the different chains from different, randomly-chosen initial conditions.
- If the different chains give similar results, then their traces can be combined using a program such as LogCombiner (http://beast2.org).

Best Practice

It is always a good idea to run at least two independent chains in any MCMC analysis.
The first chain (the **cold chain**) is constructed so that the target distribution \( \pi \) is its stationary distribution.

The \( i \)'th chain is constructed so that its stationary distribution is proportional to

\[
\pi_i(x) \propto \pi(x)^{1/T_i}
\]

where \( T_i \) is said to be the **temperature** of the chain. As \( T_i \) increases, the distribution \( \pi_i(x) \) becomes flatter, which makes it easier for this chain to converge.

The MH algorithm is used to swap the states occupied by different chains in such a way that the stationary distribution of each chain is maintained.

The output from the cold chain is then used to approximate the target distribution.


