Topic 7: Inference

Learning Objectives

1. Define the objective functions for likelihood, and Bayesian inference. Describe the measures of confidence in each paradigm. What are the pros and cons of each paradigm?
2. Describe the data, $\mathcal{D}$, and the model, Θ, for a given biological inference problem
3. Define a likelihood in terms of probability and describe why the likelihood is not a probability.
4. Name the terms in Bayes' theorem and describe their role in this inference paradigm.
5. Describe the Metropolis-Hastings algorithm for simulating posteriors

Advanced Topic Parametric versus Non-parametric Inference

Lecture 7.1 Likelihood Fitting

In this lecture we shift our focus from thinking about the generating processes (how model assumptions and parameters can create data) to how to **infer** the model structure and parameters from data. Inference relies on the specification of an **objective function**. Objective functions formalize a scientific view of the world about what it means for a model to *best explain* the data.

You have probably encountered objective functions before, whether you new it or not. Two examples include Occam's Razor (aka parsimony) and least-squares fitting.

Occam's Razor posits that among competing hypotheses (aka models) the simplest model that can explain the data is best.

Example: 7.1 Parsimony Trees

Consider the following sample of three genome sequences where each site is characterized by 1 of 2 states (states: 0 and 1).

Species	Site 1	Site 2	Site 3
A	0	1	1
B	0	1	0
C	1	0	1

1. What are the possible tree topologies for these 3 species?

2. How many mutations are required to explain the data with each topology?

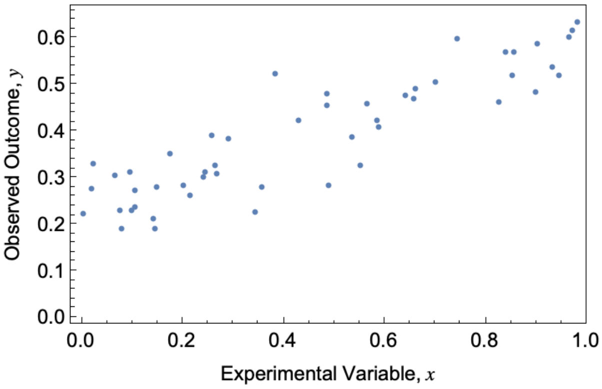
3. What is the maximum-parsimony topology?

4. Explain the objective function here and why might it be wrong?
In this world view whatever evolutionary history requires the fewest changes (our measure of simplicity) is our best guess of the true evolutionary history. This view may be wrong because mutations may not actually be rare. Rather if enough time has passed then mutations a state may mutation from 0 to 1 and back to 0. Such "erased" mutations would be discounted in a parsimony model.

In least-squares fitting we use the objective function that the *lowest order* polynomial who's coefficients *minimize* the squared distance between the fit and the data (the sum squared error) is the best model.

Example: 7.2 Least-squares fitting

Consider the following data set of 50 points.



1. Use least squares fitting to find the best fit model.

- Fitting a constant to the data we have:

$$f(x) = \beta_0$$

Using built-in least squares fitting we estimate $\hat{\beta}_0 = 0.392 \quad CI = (0.355, 0.429)$. Hence intercept is significant.

- Fitting a line to the data we have:

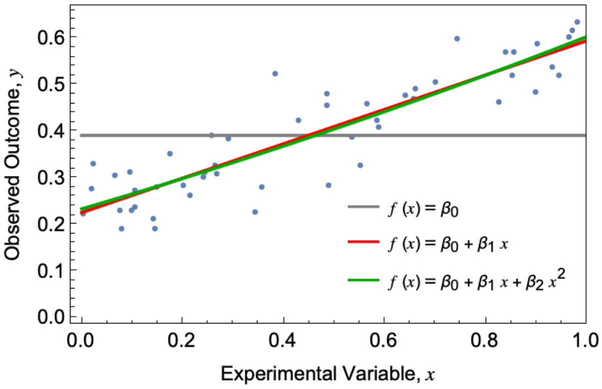
$$f(x) = \beta_0 + \beta_1 x$$

Where $\hat{\beta}_0 = 0.226 \quad CI = (0.197, 0.256)$ and $\hat{\beta}_1 = 0.369 \quad CI = (0.316, 0.423)$. Hence both are significant.

- Fitting a quadratic to the data we have:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Where $\hat{\beta}_0 = 0.234 \quad CI = (0.191, 0.277)$, $\hat{\beta}_1 = 0.316 \quad CI = (0.096, 0.537)$ and $\hat{\beta}_2 = 0.053 \quad CI = (-0.164, 0.271)$. Only the intercept and slope are significant.



Today we are going to use a different objective function called the **likelihood**.

Definition of the Likelihood

The likelihood of a model θ given some data $\mathcal{D}$, written as $\mathcal{L}(\theta|\mathcal{D})$ is defined as the probability of the data given the model.

$$\mathcal{L}(\theta|\mathcal{D}) = \Pr(\mathcal{D}|\theta)$$

In this world-view we are assuming that whatever process has the highest probability of creating our observed outcome is what happened. Note that in the likelihood the model θ is the focal outcome *given* the data $\mathcal{D}$

The likelihood versus a probability

It is important to note that while the likelihood $\mathcal{L}(\theta|\mathcal{D})$ is equal to a probability $\Pr(\mathcal{D}|\theta)$. It itself is not a probability. Note that conditional probabilities $\Pr(A|B)$ must satisfy the total probability rule that:

$$\sum_a \Pr(A=a|B) = 1$$

In words, for the probability $\Pr(\mathcal{D}|\theta)$, the model θ has to generate something even if the *data* generated by the focal model may not be the data we currently have. Hence we have:

$$\sum_d \Pr(\mathcal{D}=d|\theta) = 1$$

Note that the opposite is not necessarily true:

$$\sum_q \Pr(\theta=q|\mathcal{D}) \neq 1$$

In words, we may have a dataset that is impossible to generate with any conceivable model. Hence, while the likelihood is equal to a probability it *can not* be interpreted as a probability.

$$\mathcal{L}(\theta|\mathcal{D}) \neq \Pr(\theta|\mathcal{D})$$

Obtaining Likelihoods for Stochastic Models

The definition of the model likelihood as the probability of the data arising from a given model makes it very straightforward to derive likelihoods from a stochastic process.

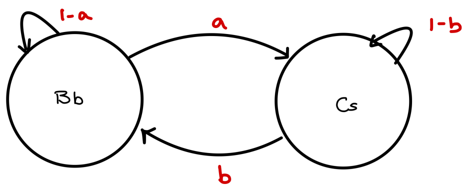
Let's start with an example from a discrete-time discrete-state stochastic process.

Example: 7.3 Species coexistence

A researcher studying competitive exclusion and niches in the classic intertidal barnacle system performs a long-term survey of species presence/absence at a focal field site. The researcher assess the presence or absence of two barnacle species *Balanus balanoides* (Species Bb) and *Chthamalus stellatus* (Species Cs). The presence/absence data for each year in a 10 year survay are shown below:

Year	1	2	3	4	5	6	7	8	9	10
Species	Bb	Bb	Cs	Cs	Bb	Bb	Cs	Bb	Bb	Cs

1. Propose a DTDS stochastic process describing the dynamics of replacement and competitive exclusion in this site.



$$M = \begin{bmatrix} 1-a & b \\ b & 1-b \end{bmatrix}$$

2. What is the data $\mathcal{D}$ and the model θ in this example?

The data $\mathcal{D}$ is the sequence of species occurances. Translating this into states we have:

$$\mathcal{D} = \vec{x} = [1,1,2,2,1,1,2,1,1,2]$$

The model here has two parts. First the model is the assumptions we made about the structure of the stocahstic process. For the purposes of this question we are going to assume that this is fixed. The second part of the model is the parameters $\theta = \{a, b\}$.

3. Use your answers to 1 and 2 to calculate the likelihood for a given parameter set where $a = 0.1$ and $b = 0.2$

The likelihood $\mathcal{L}(\theta|\mathcal{D}) = \Pr(\mathcal{D}|\theta)$. We start in state 1 (Bb) in year 1. The probability of seeing the observed data then is:

$$\begin{aligned}\mathcal{L}(\theta|\mathcal{D}) = \Pr(\mathcal{D}|\theta) &= \prod_{j=2}^{10} P_{x_{j-1},x_j} \\ &= (1-a)^3 * (1-b) * a^3 * b^2 = 0.000023328\end{aligned}$$

where x_j is the state in year j

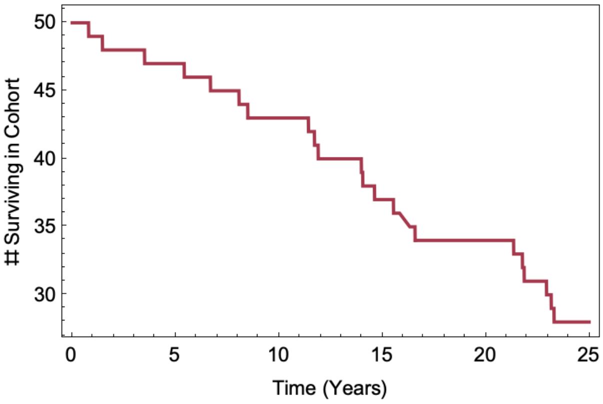
4. Does this make sense?
There are a lot of different sequences of 9 state changes. So the probability of seeing our particular sequences should be quite small.

Or an example with a continuous-time discrete-state stochastic process.

Example: 7.4 Survival Analysis

A researcher is performing a longitudinal study of health outcomes in a community. The research compiles a cohort of 50 participants (all of whom are 60 at the beginning of the study) in year $t = 0$ and follows their development and health for the subsequent decades. Following are data on the date of time to death of the 23 of the 50 participants during the subsequent 25 years. This type of data is of the classic form of a 'survival analysis'.

$$\begin{aligned}\vec{t} = \{&0.79, 1.46, 3.5, 5.42, 6.69, 8.07, 8.49, \\ &11.43, 11.72, 11.9, 13.99, 14.05, 14.62, 15.54, \\ &16.28, 16.58, 21.35, 21.78, 21.86, 22.94, 23.17, 23.31, 24.11\}\end{aligned}$$



1. Propose a stochastic model of this system.

Let's model the data as a Poisson process with rate λ. We are interested in both the timing of the death events and the fact that 27 individuals survive past the 25 year duration of the experiment.

Advanced Topic Hidden Markov Models and fitting an Erlang distributed

2. What is the model and the data here?

The model here is 1) the assumption about the Poisson nature of the model which we will once again assume to be fixed and 2) the parameter $\theta = \lambda$.

Here the data consists of two facts:

1. the times until the death events. $\vec{t}$
2. the fact that 27 individuals live past $t = 25$.

3. What is the probability of our data given $\lambda=0.1

We can model each time to death as the PDF of independent exponential random variables.

We can model the probability of not dying as the CDF of independent random variables

$$\mathcal{L}(\theta|\mathcal{D}) = \Pr(\mathcal{D}|\lambda=0.1) = \underbrace{\prod_{i=1}^{23} 0.1e^{-0.1t_i}}_{\text{prob. } i \text{ death event}} \underbrace{\left(1-\left[1-e^{-0.1*25.}\right]\right)^{27}}_{\text{prob. of not dying}} = 1.14\times 10^{-38}$$

4. Does this make sense?

There are alot of different times that people could have died. So the probability of seeing exactly the set of times that we saw is tiny!

The Utility of the Log-Likelihood

We ultimately want to find the model θ that has the largest probability of creating the data that we observed. In other words, we ultimately want to find the maximum likelihood. But as we saw above the value of the likelihood can be quite small. This can be a computational problem as comparing small numbers can create roundoff error in our computer.

The likelihood is small because the probability of observing a big data set naturally involves calculating the probability of observing obs 1 AND obs 2 AND obs 3... and hence a product of a bunch of things less than 1. It would be much nicer if we could sum over stuff instead of multiplying.

Recall that $\ln(a*b*c) = \ln(a)+\ln(b)+\ln(c)$. Conveniently the function $f(y) = \ln(y)$ is a **monotonic transformation** meaning that although although it changes the value of the dependent variable it does not change the ranking of the dependent variable:

$$\ln(y_1) > \ln(y_2) \quad \text{if and only if} \quad y_1 > y_2$$

Hence the value of x that maximizes the function $y(x)$ also maximizes the transformed function $\ln(y(x))$.

For this reason we often use the **log-likelihood** rather than the likelihood.

$$l(\theta|\mathcal{D}) = \ln\left(\mathcal{L}(\theta|\mathcal{D})\right)$$

Example: 7.3 cont Species coexistence

Calculate the log-likelihood of the data in Example 7.3.

We had that:

$$\mathcal{L}(\theta=\{a,b\}|\mathcal{D}=\vec{x}) = (1-a)^3*(1-b)*a^3*b^2$$

Taking the log of both sides we have:

$$l(\theta|\mathcal{D}) = 3*\ln(1-a)+\ln(1-b)+3*\ln(a)+2*\ln(b)$$

Example: 7.4 cont Survival Analysis

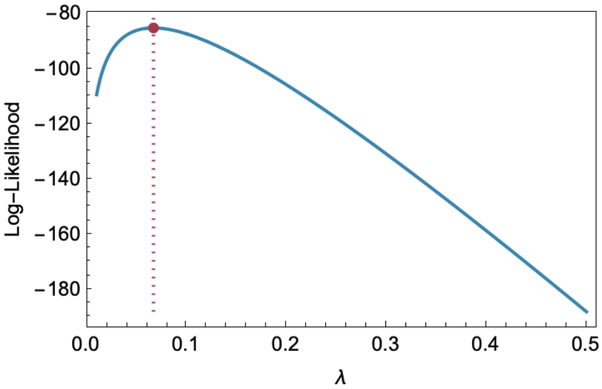
Calculate the log-likelihood of the data in Example 7.4

The Likelihood Surface

For the purposes of this class, I am only going to consider 1D and 2D models (e.g., models with 1 or 2 parameters). In this case we can directly plot the likelihood over parameter space.

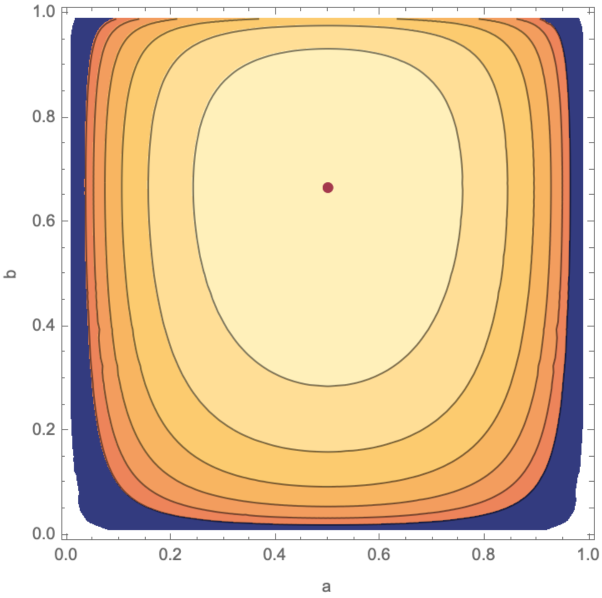
Example: 7.4 cont Survival Analysis

Plot the log-likelihood surface for $\lambda = [0.01, 0.5]$ and identify the value of λ that maximizes this function.



Example: 7.3 cont Species coexistence

1. Plot the log-likelihood surface for $a = [0.01, 0.99]$ and $b = [0.01, 0.99]$ and identify the coordinate pair of (a, b) that maximizes this function.



****

Point Estimates and Uncertainty

Point Estimates: Maximum Likelihood Estimate

When we fit a model we want to derive a single *best* estimate for each parameter, this is known as the **point estimate**. In the case of likelihood fitting we obtain this point estimate by calculating the **maximum likelihood estimate**

Uncertainty

In addition to the single point estimates we also capture our uncretainty in the value of the parameters. In the context of likelihood fitting our goal is to calculate the **confidence interval** about each parameter. In this class we will focus on calculating the 95% confidence interval which gives the range of possible parameter values.

There are several different methods for calculating parameter uncertainty using the likelihood objective function but most of them rely in part on the concept of a **likelihood ratio**.

The likelihood ratio of two models Θ_1 and Θ_2 is:

$$LR = \frac{L(\theta_1|\mathcal{D})}{L(\theta_2|\mathcal{D})}$$

If the likelihood ratio is significantly different from 1, it suggests that one hypothesis or model is more supported by the data than the other. In many cases, the likelihood ratio can be used to perform hypothesis tests, construct confidence intervals, and make decisions about the adequacy of models.

It's important to note that the likelihood ratio is a fundamental concept in likelihood-based statistical inference, and its application can vary depending on the specific statistical problem at hand.

The test statistic

$$l_{\text{ratio}} = 2\ln\left(\frac{L(\theta_1|\mathcal{D})}{L(\theta_2|\mathcal{D})}\right) \begin{aligned} &= 2(\ln(L(\theta_1|\mathcal{D})) - \ln(L(\theta_2|\mathcal{D}))) \\ &= 2(l(\theta_1|\mathcal{D}) - l(\theta_1|\mathcal{D})) \end{aligned}$$

Conveniently l_{ratio} is Chi-Squared distributed with 1 degree of freedom meaning that θ_1 is statistically more likely then $\theta 2$ if:

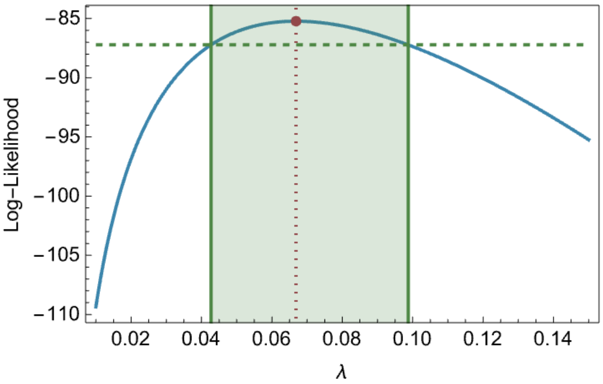
$$\begin{aligned} &2(l(\theta_1|\mathcal{D}) - l(\theta_1|\mathcal{D})) > 3.84 \approx 4 \\ &(l(\theta_1|\mathcal{D}) - l(\theta_1|\mathcal{D})) > 2 \end{aligned}$$

All this is to say, we can approximate the **confidence interval** of our parameters by considering the set of parameters that are within 2 log-likelihood units of the maximum-likelihoood esitmate

---

Example: 7.4 cont Survival Analysis

What is the CI on the death rate, λ?



Example: 7.3 cont Species coexistence

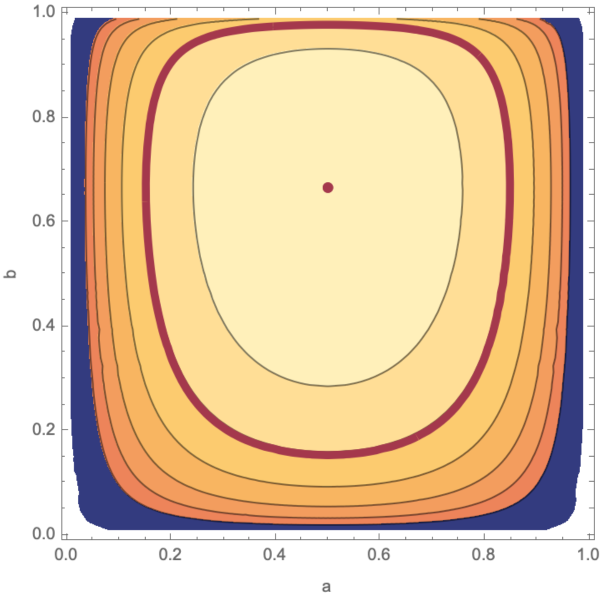
1. What is the point estimate and CI for a and b?

The point estimate is the combination of a and b which maximize the log-likelihood. This is shown in the plot above as is given by the coordinate pair: $(\frac{1}{2}, \frac{2}{3})$.

$l((\frac{1}{2}, \frac{2}{3})|\vec{x}) = -6.07$

To find the CI of a and b we want to plot the space over which the log-likelihood is >-8.07.

$a = \{0.151131, 0.848869\}$
$b = \{0.153807, 0.979119\}$



2. Are a and b statistically different? What does this mean?

No they are not statistically different- we can not confirm one species as the superior competitor given the data.

Advanced Topic Model Selection via AIC

Advanced Topic Stochastic Search

Advanced Topic Simulated annealing

Advanced Topic Tree Likelihoods

Advanced Topic Population Genetic Inference from the Coalescent

Lecture 7.2 Bayesian Model Fitting

Bayes' Theorem

Recall that Bayes' Theorem is a method for switching the order of the conditioning for a conditional probability.

$$\Pr(A|B) = \frac{\Pr(B|A)\,\Pr(A)}{\Pr(B)}$$

In Topic 2 we use this theorem to draw probabilistic conclusions. But now let's subsitute in particular quantities for A and B. Specifically if we want to draw conclusions about what model is best given a data set, we want to know $\Pr(\Theta|\mathcal{D})$.

Using Bayes' theorem then we have:

$$\Pr(\Theta|\mathcal{D}) = \frac{\overbrace{\Pr(\mathcal{D}|\Theta)}^{\text{likelihood}}\overbrace{\Pr(\Theta)}^{\text{prior}}}{\underbrace{\Pr(\mathcal{D})}_{\text{marginal prob.}}}$$

We can name each of these terms:

- $\Pr(\mathcal{D}|\Theta)$ is by definition above the likelihood of the model given the data.
- $\Pr(\Theta)$ is known as the prior and is a weighting of the a particular model independent of the likelihood (discussed more below)
- $\Pr(\mathcal{D})$ is known as the marginal probability and is a weird thing to wrap your head around. One way of thinking abou this is using the definition of total probability:

$$\Pr(\mathcal{D}) = \sum_{\Theta} \Pr(\mathcal{D}|\Theta)\Pr(\Theta)$$

In words, it is the probability that any model generated the data weighted by how likely those different models were. This is NOT something we can calculate. So on the surface this means that our Bayes' theorem is a bust. But suppose that we don't look at $\Pr(\Theta|\mathcal{D})$ of any one model, but rather the relative probability of two models:

$$\frac{\Pr(\Theta_1|\mathcal{D})}{\Pr(\Theta_2|\mathcal{D})} = \frac{\Pr(\mathcal{D}|\Theta_1)\Pr(\Theta_1)}{\Pr(\mathcal{D}|\Theta_2)\Pr(\Theta_2)}$$

Now we don't have to calculate the problematic term but we only can compare models. This really isn't a problem as that is how we interpret PDFs anyway (we will just normalize so our total probability is 1 across 'model space').

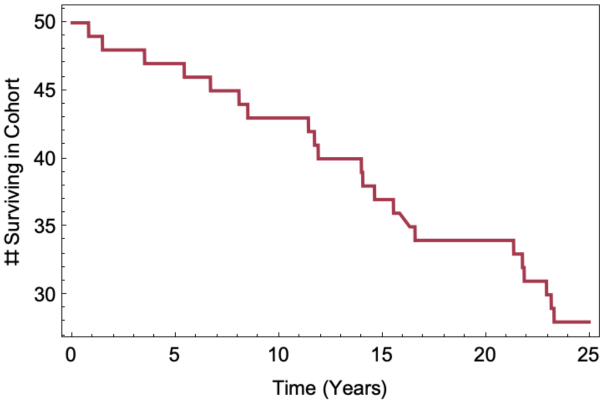
The Power of the Prior

So what is the prior anyway? Priors can be useful ways of explictly or implicitly incorporating information that is NOT in the data set $\mathcal{D}$. As such Bayes theorem is a useful way of bringing together disperate datasests that would be difficult to capture with one unified model. Sometimes this is because that complementary model is not really a data set but rather intuitive reasoning. The intuitive nature of priors can be very powerful, allowing us to combine explicit data with our expert experience, but also risky, if you use your intuition to get an answer you can't then also use your intuition to dummy-check it!

Example: 7.4 cont Survival Analysis

Condsider the survival probability analysis from above. In this case the data is:

$$\vec{t} = \{0.79, 1.46, 3.5, 5.42, 6.69, 8.07, 8.49, \\ 11.43, 11.72, 11.9, 13.99, 14.05, 14.62, 15.54, \\ 16.28, 16.58, 21.35, 21.78, 21.86, 22.94, 23.17, 23.31, 24.11\}$$



Where the data is following the time deaths of 50 participants whoa are all 60 at the beginning of the study.

Suppose that in the focal community there is no one over the age of 95, propose a prior for the mortality rate λ of 60 year olds.

Given the observation, 60 year olds in this community are very unlikely to live more than 35 more years. Using the exponential distribution we have:

$$\text{CDF}(35|\lambda^*) \approx 1 = 0.99$$

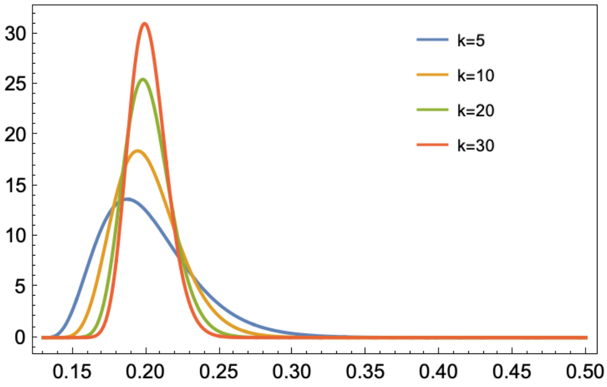
Solving we have: $\lambda^* = 0.13$. So the death rate is likely at least this fast!

We also know that λ can't be too fast and in fact we are pretty sure that it is slowish. So let's create a prior that prioritizes values near 0.13 or slightly larger and has down weights higher estimates. λ is a real number so we need to use a continuous distribution like the Erlang (aka Gamma) distribution.

Let's make the mean of the prior $\lambda \approx 0.2$ which is nearish to the max. In this case we have:

$$\Pr(\lambda) = \text{Erlang}(\lambda - 0.13|k, 14k)$$

Playing with values of k we have:



Let's be least restrictive of these options and use $k = 5$.

Point Estimate: Maximum Posterior Estimate

The Maximum Posterior Estimate (MPE) is a point estimate of a parameter in Bayesian statistics. Much like the Maximum Likelihood Estimate the MPE is the value of the model θ (e.g., model assumptions and parameter values) that maximizes the posterior probability distribution, representing the most probable value of the parameter given the observed data and the prior distribution. The MPE is denoted as $\tilde{\theta}$ and is defined mathematically as:

$$\tilde{\theta} = \arg\max_{\theta} \left(\Pr(\theta | \text{data}) \right)$$

Credible Intervals

A credible interval is a statistical concept used in Bayesian inference to quantify the uncertainty about a parameter estimate. It provides a range of plausible values for the parameter based on the observed data and the posterior distribution. There are several ways to define a credible interval, and the choice may depend on the specific goals or preferences of the analyst. **Note:** the credible interval is NOT a confidence interval as the former is for Bayesian interpretations and the later is for freuqentist (i.e. likelihood) interpretations

1. **Highest Posterior Density (HPD) Interval:**
   - The HPD interval is the narrowest interval that contains a specified probability mass of the posterior distribution.
   - It is often used to provide the most precise interval estimate.
   - The HPD interval is sometimes referred to as the highest density interval.
2. **Equal-Tailed Interval:**
   - The equal-tailed interval is symmetric around the posterior median.
   - It includes the same amount of probability mass in both tails of the posterior distribution.
   - The equal-tailed interval is straightforward to compute but may be wider than the HPD interval.
3. **Central Interval:**
   - The central interval is a symmetric interval around the posterior mean.
   - It may be sensitive to outliers or skewed distributions.

The choice of a credible interval definition may depend on the shape of the posterior distribution, the desire for precision, or the specific interpretability requirements. In practice, it's common to report multiple types of credible intervals to provide a comprehensive understanding of parameter uncertainty.

Advanced Topic Model Selection via BIC and Trans-dimensional MCMC

The Metropolis-Hastings Algorithm

Okay, so we want a posetrior and we now know how to interpret it. But how do we calculate it?

The Metropolis-Hastings algorithm, is an algorithm that is designed to generate samples from a *target probability distribution*, in our case the posterior. The Metropolis-Hastings algorithm is especially useful when direct sampling is difficult. It allows us to generate a sequence of samples that converge to the desired distribution.

Initialization:
The alogorithm starts by drawing a random starting point from the sample space of the focal distribution. In the case of the posterior $\Pr(\theta|\mathcal{D})$, this requires generating a random model θ.

Proposal Distribution:

We then propose a new model using the proposal distribution $q(\theta'|\theta)$ is chosen to suggest a new sample θ' given the current sample θ.

Acceptance Probability:

The acceptance probability α is computed as the ratio of the posterior probabilities of the proposed and current samples:

$$\alpha = \min\left(1, \frac{\Pr(\theta'|\mathcal{D}) \cdot q(\theta|\theta')}{\Pr(\theta|\mathcal{D}) \cdot q(\theta'|\theta)}\right)$$

where α is used to decide whether to accept the proposed sample or stay at the current sample.

Given a random variable u is sampled from a uniform distribution in the range $[0, 1]$.

$$\text{If } u < \alpha, \quad \theta_{\text{new}} = \theta', \quad \text{else } \theta_{\text{new}} = \theta$$

Iteration:

Repeat steps 1-3 for a specified number of iterations to obtain a chain of samples.

Convergence

Measures of convergence are essential for assessing whether the algorithm has reached a stationary distribution and whether the samples generated are representative of the target distribution. Two common measures of convergence are the **Convergence Ratio** and the **Effective Sample Size (ESS)**.

1. **Convergence Ratio**
   The convergence ratio is a diagnostic measure that evaluates the convergence of the Markov chain. It is often computed as the ratio of the variances of two independent subsequences of the chain, with the idea that if the chain has converged, these variances should be similar.
   In the context of the Metropolis-Hastings algorithm and other Markov Chain Monte Carlo (MCMC) methods, measures of convergence are essential for assessing whether the algorithm has reached a stationary distribution and whether the samples generated are representative of the target distribution. Two common measures of convergence are the **convergence ratio** and the **Effective Sample Size (ESS)**.
2. Convergence Ratio:

The convergence ratio is a diagnostic measure that evaluates the convergence of the Markov chain. It is often computed as the ratio of the variances of two independent subsequences of the chain, with the idea that if the chain has converged, these variances should be similar.

The convergence ratio (R) is calculated as follows:

$$R = \frac{\text{Var}(\theta_1, \ldots, \theta_n)}{\text{Var}(\theta_{n+1}, \ldots, \theta_{2n})}$$

where $\theta_1, \ldots, \theta_{2n}$ are the samples obtained from the Markov chain.

Interpretation:

- R close to 1 suggests convergence.
- Values significantly greater than 1 indicate lack of convergence.

2. Effective Sample Size (ESS):

The Effective Sample Size (ESS) is a measure of the number of independent samples that carry the same amount of information as the actual number of samples. It accounts for the autocorrelation present in the Markov chain and gives a more accurate estimate of the effective information content.

The ESS is computed using the *autocorrelation* function $\rho(k)$ of the chain:

$$\text{ESS} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho(k)}$$

where n is the total number of samples.

Interpretation:

- A higher ESS implies more independent samples and better representation of the target distribution.
- A low ESS may indicate high autocorrelation, and more samples may be needed for reliable inference.

Additional Considerations:

- **Trace Plots:** Visual inspection of trace plots can also provide insights into convergence. A stable, random-looking trace suggests convergence, while trends or patterns may indicate lack of convergence.
- **Gelman-Rubin Statistic (R-hat):** R-hat compares the within-chain and between-chain variances. $R \approx 1$ suggests convergence.

In practice, it is common to use a combination of these measures to assess convergence. It's important to monitor convergence and consider potential burn-in periods before using samples for inference.

Example: 7.5 cont Survival Analysis

Python: Lecture7_2.ipynb

1. Using the likelihood and prior above, use the Metrapolis-Hastings Algorithm to estimate the MPE and creible interval of λ.

2. Show convergence statistics.
