Chapter 2

Revision of Bayesian inference
Bayes Theorem

Notation
Parameter vector: $\theta = (\theta_1, \ldots, \theta_p)^T$
Data: $x = (x_1, x_2, \ldots, x_n)^T$.

Bayes Theorem for events
For two events $E$ and $F$, the conditional probability of $E$ given $F$ is:

$$Pr(E|F) = \frac{Pr(E \cap F)}{Pr(F)}.$$ 

This gives us the simplest version of Bayes theorem:

$$Pr(E|F) = \frac{Pr(F|E)Pr(E)}{Pr(F)}.$$
Bayes Theorem

If $E_1, E_2, \ldots, E_n$ is a partition i.e. the events $E_i$ are mutually exclusive with $Pr(E_1 \cup \ldots E_n) = 1$, then

$$Pr(F) = Pr((F \cap E_1) \cup (F \cap E_2) \cup \ldots \cup (F \cap E_n)) = \sum_{i=1}^{n} Pr(F \cap E_i)$$

$$= \sum_{i=1}^{n} Pr(F|E_i)Pr(E_i).$$

This gives us the more commonly used version of Bayes theorem:

$$Pr(E_i|F) = \frac{Pr(F|E_i)Pr(E_i)}{\sum_{i=1}^{n} Pr(F|E_i)Pr(E_i)}.$$
Bayes theorem is useful because it tells us how to turn probabilities around.

Often we are able to understand the probability of outcomes $F$ conditional on various hypotheses $E_i$. 
Bayes theorem is useful because it tells us how to turn probabilities around.

Often we are able to understand the probability of outcomes $F$ conditional on various hypotheses $E_i$.

We can then compute probabilities of the form $Pr(F|E_i)$. 
Bayes theorem is useful because it tells us how to turn probabilities around.

Often we are able to understand the probability of outcomes $F$ conditional on various hypotheses $E_i$.

We can then compute probabilities of the form $Pr(F|E_i)$.

However, when we actually observe some outcome, we are interested in which hypothesis is most likely to be true, or in other words, we are interested in the probabilities of the hypotheses conditional on the outcome $Pr(E_i|F)$.

Bayes theorem tells us how to compute these posterior probabilities, but we need to use our prior belief $Pr(E_i)$ for each hypothesis.
- Bayes theorem is useful because it tells us how to turn probabilities around.
- Often we are able to understand the probability of outcomes $F$ conditional on various hypotheses $E_i$.
- We can then compute probabilities of the form $Pr(F|E_i)$.
- However, when we actually observe some outcome, we are interested in which hypothesis is most likely to be true, or in other words, we are interested in the probabilities of the hypotheses conditional on the outcome $Pr(E_i|F)$.
- Bayes theorem tells us how to compute these posterior probabilities, but we need to use our prior belief $Pr(E_i)$ for each hypothesis.
- In this way, Bayes theorem gives us a coherent way to update our prior beliefs into posterior probabilities following an outcome $F$. 
Inference for discrete parameters and data

Suppose our data and parameters have discrete distributions.

The hypotheses $E_i$ are values for our parameter(s) $\Theta$, while the outcome $F$ is a set of data $X$.

Bayes theorem is then

$$Pr(\Theta = \theta | X = x) = \frac{Pr(X = x | \Theta = \theta) Pr(\Theta = \theta)}{\sum_t Pr(X = x | \Theta = t) Pr(\Theta = t)}.$$
Due to inaccuracies in drug testing procedures (e.g., false positives and false negatives), in the medical field the results of a drug test represent only one factor in a physician’s diagnosis.

Yet when Olympic athletes are tested for illegal drug use (i.e. doping), the results of a single test are used to ban the athlete from competition.

In *Chance* (Spring 2004), University of Texas biostatisticians D. A. Berry and L. Chastain demonstrated the application of Bayes’s Rule for making inferences about testosterone abuse among Olympic athletes.

They used the following example.
Example 2.1 (Drug testing in athletes)

In a population of 1000 athletes, suppose 100 are illegally using testosterone. Of the users, suppose 50 would test positive for testosterone. Of the nonusers, suppose 9 would test positive.

If an athlete tests positive for testosterone, use Bayes theorem to find the probability that the athlete is really doping.
Example 2.1 (Drug testing in athletes)

In a population of 1000 athletes, suppose 100 are illegally using testosterone. Of the users, suppose 50 would test positive for testosterone. Of the nonusers, suppose 9 would test positive.

If an athlete tests positive for testosterone, use Bayes theorem to find the probability that the athlete is really doping.

Let $\theta = 1$ be the event that the athlete is a user, and $\theta = 0$ be a nonuser; and let $x = 1$ be the event of a positive test result. We need to consider the following questions.

- What is the prior information for $\theta$?
- What is the observation?
- What is the posterior distribution of $\theta$ after we get the data?
Consider a continuous parameter $\Theta$, and observed data $X$. (Notice that both data and parameters are random variables).

Our prior beliefs about the parameters define the prior distribution for $\Theta$, specified by a density $\pi(\theta)$.

We formulate a model that defines the distribution of $X$ given the parameters, i.e. we specify a density $f_{X|\Theta}(x|\theta)$.

This can be regarded as a function of $\theta$ when we have got some fixed observed data $x$, called the likelihood:

$$L(\theta|x) = f_{X|\Theta}(x|\theta).$$
The prior and likelihood determine the full \textit{joint density} over data and parameters:

\[
f_{\Theta, X}(\theta, x) = \pi(\theta)L(\theta|x).
\]

Given the joint density we are then able to compute its marginals and conditionals:

\[
f_X(x) = \int_{\Theta} f_{\Theta, X}(\theta, x) \, d\theta = \int_{\Theta} \pi(\theta)L(\theta|x) \, d\theta
\]

and

\[
f_{\Theta|X}(\theta|x) = \frac{f_{\Theta, X}(\theta, x)}{f_X(x)} = \frac{\pi(\theta)L(\theta|x)}{\int_{\Theta} \pi(\theta)L(\theta|x) \, d\theta}.
\]
The prior and likelihood determine the full *joint density* over data and parameters:

\[
f_{\Theta, X}(\theta, x) = \pi(\theta) L(\theta | x).
\]

Given the joint density we are then able to compute its marginals and conditionals:

\[
f_X(x) = \int_{\Theta} f_{\Theta, X}(\theta, x) \, d\theta = \int_{\Theta} \pi(\theta) L(\theta | x) \, d\theta
\]

and

\[
f_{\Theta | X}(\theta | x) = \frac{f_{\Theta, X}(\theta, x)}{f_X(x)} = \frac{\pi(\theta) L(\theta | x)}{\int_{\Theta} \pi(\theta) L(\theta | x) \, d\theta}.
\]

\(f_{\Theta | X}(\theta | x)\) is known as the *posterior density*, and is usually denoted \(\pi(\theta | x)\).
This leads to the continuous version of Bayes theorem:

\[
\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{\int_{\Theta} \pi(\theta)L(\theta|x) \, d\theta}
\]
This leads to the continuous version of Bayes theorem:

$$
\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{\int_\Theta \pi(\theta)L(\theta|x) \, d\theta}
$$

Now, the denominator is not a function of $\theta$, so we can in fact write this as

$$
\pi(\theta|x) \propto \pi(\theta)L(\theta|x)
$$

where the constant of proportionality is chosen to ensure that the density integrates to one. Hence, *the posterior is proportional to the prior times the likelihood.*
Bayesian computation

- That should be everything!
- However, typically only the kernel can be written in closed form, that is
  \[ \pi(\theta)L(\theta|x). \]
- The denominator has the form
  \[ \int_{\Theta} \pi(\theta)L(\theta|x) \, d\theta \]
  and this integral is often intractable.
Bayesian computation

- That should be everything!
- However, typically only the kernel can be written in closed form, that is
  \[ \pi(\theta)L(\theta|x). \]
- The denominator has the form
  \[ \int_{\Theta} \pi(\theta)L(\theta|x) \, d\theta \]
  and this integral is often intractable.
- We can try to evaluate this numerically, but
  - the integral may be very high dimensional, i.e. there might be many parameters, and
  - the support of \( \Theta \) may be very complicated.
- Similar problems arise when we want to marginalize or work out an expectation.
Example 2.2 (Normal with unknown mean and variance)

- **Model:**
  
  \[ X_i | \mu, \tau \sim N(\mu, 1/\tau). \]

- **Likelihood for a single observation:**
  
  \[
  L(\mu, \tau | x_i) = f(x_i | \mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp \left\{ -\frac{\tau}{2} (x_i - \mu)^2 \right\}
  \]

- **So for \( n \) independent observations, \( x = (x_1, \ldots, x_n)^T \)**
  
  \[
  L(\mu, \tau | x) = f(x | \mu, \tau) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp \left\{ -\frac{\tau}{2} (x_i - \mu)^2 \right\}
  \]
Example 2.2 (Normal with unknown mean and variance)

\[ L(\mu, \tau|\mathbf{x}) = \left( \frac{\tau}{2\pi} \right)^{\frac{n}{2}} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 \right\} \]

\[ \propto \tau^{\frac{n}{2}} \exp \left\{ -\frac{n\tau}{2} \left[ s^2 + (\bar{x} - \mu)^2 \right] \right\} \]

where

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]
Example 2.2 (Normal with unknown mean and variance)

- Conjugate prior specification possible:

  \[ \tau \sim \text{Gamma}(g, h) \]

  \[ \mu | \tau \sim N \left( b, \frac{1}{c\tau} \right) \].
Example 2.2 (Normal with unknown mean and variance)

- Conjugate prior specification possible:
  \[ \tau \sim \text{Gamma}(g, h) \]
  \[ \mu | \tau \sim N \left( b, \frac{1}{c\tau} \right) . \]

- Undesirable if our prior beliefs for \( \mu \) and \( \tau \) separate into independent specifications.
Example 2.2 (Normal with unknown mean and variance)

- Conjugate prior specification possible:
  \[ \tau \sim \text{Gamma}(g, h) \]
  \[ \mu | \tau \sim N \left( b, \frac{1}{c \tau} \right). \]

- Undesirable if our prior beliefs for \( \mu \) and \( \tau \) separate into independent specifications.

- Take independent priors for the parameters:
  \[ \tau \sim \text{Gamma}(g, h) \]
  \[ \mu \sim N \left( b, \frac{1}{c} \right). \]
Example 2.2 (Normal with unknown mean and variance)

- Conjugate prior specification possible:
  \[ \tau \sim \text{Gamma}(g, h) \]
  \[ \mu | \tau \sim N \left( b, \frac{1}{c\tau} \right). \]

- Undesirable if our prior beliefs for \( \mu \) and \( \tau \) separate into independent specifications.

- Take independent priors for the parameters:
  \[ \tau \sim \text{Gamma}(g, h) \]
  \[ \mu \sim N \left( b, \frac{1}{c} \right). \]

- Specification no longer conjugate, making analytic analysis intractable. Let us see why!
Example 2.2 (Normal with unknown mean and variance)

\[
\pi(\mu) = \sqrt{\frac{c}{2\pi}} \exp \left\{ -\frac{c}{2} (\mu - b)^2 \right\} \propto \exp \left\{ -\frac{c}{2} (\mu - b)^2 \right\}
\]

\[
\pi(\tau) = \frac{h^g}{\Gamma(g)} \tau^{g-1} \exp\{-h\tau\} \propto \tau^{g-1} \exp\{-h\tau\}
\]
Example 2.2 (Normal with unknown mean and variance)

\[
\pi(\mu) = \sqrt{\frac{c}{2\pi}} \exp \left\{ -\frac{c}{2} (\mu - b)^2 \right\} \propto \exp \left\{ -\frac{c}{2} (\mu - b)^2 \right\}
\]

\[
\pi(\tau) = \frac{h^g}{\Gamma(g)} \tau^{g-1} \exp \{-h\tau\} \propto \tau^{g-1} \exp \{-h\tau\}
\]

So

\[
\pi(\mu, \tau) \propto \tau^{g-1} \exp \left\{ -\frac{c}{2} (\mu - b)^2 - h\tau \right\}
\]

and we obtain

\[
\pi(\mu, \tau | x) \propto \tau^{g-1} \exp \left\{ -\frac{c}{2} (\mu - b)^2 - h\tau \right\}
\]

\[
\times \tau^{\frac{n}{2}} \exp \left\{ -\frac{n\tau}{2} \left[ s^2 + (\bar{x} - \mu)^2 \right] \right\}
\]

\[
= \tau^{g + \frac{n}{2} - 1} \exp \left\{ -\frac{n\tau}{2} \left[ s^2 + (\bar{x} - \mu)^2 \right] - \frac{c}{2} (\mu - b)^2 - h\tau \right\}.
\]
Example 2.2 (Normal with unknown mean and variance)

- We have the kernel of the posterior for $\mu$ and $\tau$, but it is not in a standard form.
Example 2.2 (Normal with unknown mean and variance)

- We have the kernel of the posterior for $\mu$ and $\tau$, but it is not in a standard form.
- We can gain some idea of the likely values of $(\mu, \tau)$ by plotting the bivariate surface (the integration constant isn’t necessary for that).
We have the kernel of the posterior for $\mu$ and $\tau$, but it is not in a standard form.

We can gain some idea of the likely values of $(\mu, \tau)$ by plotting the bivariate surface (the integration constant isn’t necessary for that).

We cannot work out the posterior mean or variance, or the forms of the marginal posterior distributions for $\mu$ or $\tau$, since we cannot integrate out the other variable.
Example 2.2 (Normal with unknown mean and variance)

- We have the kernel of the posterior for $\mu$ and $\tau$, but it is not in a standard form.
- We can gain some idea of the likely values of $(\mu, \tau)$ by plotting the bivariate surface (the integration constant isn’t necessary for that).
- We cannot work out the posterior mean or variance, or the forms of the marginal posterior distributions for $\mu$ or $\tau$, since we cannot integrate out the other variable.
- There is nothing particularly special about the fact that the density represents a Bayesian posterior.
- Given any complex non-standard probability distribution, we need ways to understand it, to calculate its moments, to compute its conditional and marginal distributions and their moments.

Stochastic simulation is one possible solution.
Example 2.2 (Normal with unknown mean and variance)

- We have the kernel of the posterior for $\mu$ and $\tau$, but it is not in a standard form.
- We can gain some idea of the likely values of $(\mu, \tau)$ by plotting the bivariate surface (the integration constant isn’t necessary for that).
- We cannot work out the posterior mean or variance, or the forms of the marginal posterior distributions for $\mu$ or $\tau$, since we cannot integrate out the other variable.
- There is nothing particularly special about the fact that the density represents a Bayesian posterior.
- Given any complex non-standard probability distribution, we need ways to understand it, to calculate its moments, to compute its conditional and marginal distributions and their moments.
- **Stochastic simulation** is one possible solution.
1. Describe a Monte Carlo algorithm to calculate the normalising constant in a Bayesian analysis.

2. Describe a weighted resampling algorithm for generating draws from a posterior $\pi(\theta|x)$ (known up to proportionality) using a proposal with density $\pi(\theta)$, i.e. the prior. What are the weights? How can the weights be used to estimate the unknown normalising constant?