Bayesian Inference

- The MLE is a frequentist inference method. There is another approach to inference called Bayesian inference.
- The key differences between frequentist and Bayesian approaches are shown in the next slides.
- See “A primer on Bayesian statistics in Health Economics and Outcomes research” by Anthony O’Hagan and Bryan R. Luce

<table>
<thead>
<tr>
<th><strong>The Nature of Probability</strong></th>
<th><strong>Frequentist</strong></th>
<th><strong>Bayesian</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability is a limiting, long-run frequency</td>
<td>Probability measures a personal degree of belief</td>
<td></td>
</tr>
<tr>
<td>It only applies to events that are (at least in principle) repeatable</td>
<td>It applies to any event or proposition about which we are uncertain</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The Nature of Parameters</strong></th>
<th><strong>Frequentist</strong></th>
<th><strong>Bayesian</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters are not repeatable or random</td>
<td>Parameters are unknown</td>
<td></td>
</tr>
<tr>
<td>They are therefore not random variables, but fixed (unknown) quantities</td>
<td>They are therefore random variables</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian Inference

The Nature of Inference

<table>
<thead>
<tr>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not (although it appears to) make statements about parameters</td>
<td>Makes direct probability statements about parameters</td>
</tr>
<tr>
<td>Interpreted in terms of long-run repetition</td>
<td>Interpreted in terms of evidence from the observed data</td>
</tr>
</tbody>
</table>

Bayesian Inference

Bayesian inference:

1. Choose probability density $f(\theta)$ – called the prior distribution that expresses our beliefs about a parameter $\theta$ before we see any data.
2. We choose a statistical model $f(x|\theta)$ instead of $f(x; \theta)$.
3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the posterior distribution $f(\theta|X_1,\ldots,X_n)$

Bayesian Inference

Suppose we have $n$ independent, identically distributed observations $X_1, \ldots, X_n$. The joint density of the data is:

$$f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta) = L_x(\theta)$$

$$f(\theta | x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n | \theta) f(\theta)}{f(x_1, \ldots, x_n) f(\theta) } = \frac{L_x(\theta)f(\theta)}{\int L_x(\theta)f(\theta)d\theta}$$

$$= \frac{L_x(\theta)f(\theta)}{\int L_x(\theta)f(\theta)d\theta} = \frac{\theta \alpha L_x(\theta)f(\theta)}{\theta \beta L_x(\theta)f(\theta)}$$

$$\therefore f(\theta | x_1, \ldots, x_n) \propto \theta \alpha L_x(\theta) f(\theta)$$

Bayesian Inference

What do you do with the posterior distribution?

- Use the entire distribution (can be clumsy sometimes)
- Get a point estimate by summarizing the center of the posterior – use the mean or mode
- The posterior mean is:

$$\bar{\theta} = E[\theta] = \int \theta f(\theta | x_1, \ldots, x_n)d\theta = \int \frac{\alpha L_x(\theta)f(\theta)}{\int L_x(\theta)f(\theta)d\theta}$$
Conjugate Priors

Let's redo the first candy example except this time, we will put a Beta($\alpha$, $\beta$) prior on $\theta$. Recall that $\theta$ is the probability a candy will be cherry flavored. The posterior has the form:

$$ f(\theta | x_1, ..., x_n) = \frac{f(\theta)L_n(\theta)}{\int f(\theta)L_n(\theta)d\theta} $$

Conjugate Priors

This is the Beta distribution with alpha parameter = $c + \alpha$ and beta parameter = $l + \beta$. Since it is a known pdf, it will integrate to 1.

$$ Beta(c + \alpha, l + \beta) = \int \frac{\Gamma(c + \alpha) \Gamma(l + \beta)}{\Gamma(c + \alpha + l + \beta)} \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta $$

This is the term in the denominator from the previous page. It is almost a Beta distribution except it is missing the normalization constant in front.

$$ \int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta $$

Let's call the normalization constant (the expression with the Gammas) $c$. The expression above becomes:

$$ \int c\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta = \frac{1}{c} \int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta = \frac{1}{c} $$

Conjugate Priors

Continuing from where we left off:

$$ f(\theta | x_1, ..., x_n) = \frac{\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}}{\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta} $$

$$ = \frac{\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}}{\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1} d\theta} $$

$$ = Beta(c + \alpha, l + \beta) $$
Conjugate Priors

• A conjugate prior is a family of prior probability distributions with the property that the posterior also belongs to that family.
• eg. the conjugate prior for a Bernoulli is a Beta distribution
• Other useful conjugate priors:

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Binomial</td>
<td>Beta</td>
<td>Beta</td>
</tr>
<tr>
<td>Poisson</td>
<td>Gamma</td>
<td>Gamma</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Dirichlet</td>
<td>Dirichlet</td>
</tr>
</tbody>
</table>

Conjugate Priors

Why are they useful?
• Since we know the form of the posterior, we can easily calculate statistics such as the mean.
• For example, we know:
  \[
  E[Beta(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}
  \]
  • Thus, the mean for the candy example above is:
  \[
  E[Beta(c + \alpha, l + \beta)] = \frac{\alpha + c}{\alpha + \beta + l + c}
  \]

Conjugate Priors

• You can think of \(\alpha\) and \(\beta\) in the posterior distribution as “virtual counts”
• eg. Using a uniform prior Beta(1,1), the mean of the posterior becomes:
  \[
  E[Beta(c + 1, l + 1)] = \frac{\alpha + c}{\alpha + \beta + l + c} = \frac{1 + c}{2 + l + c}
  \]

Conjugate Priors

• If we observe no data, ie. \(c=0, l=0\), the posterior mean is \(\frac{1}{2}\), which is what we would expect since we have to pick between the two flavors of lime and cherry
• If we observe lots of data, then the \(c\) term in the numerator and the \(l+c\) term in the denominator dominate the prior
Conjugate Priors

• The conjugate prior that is of most relevance to parameter estimation is the Multinomial-Dirichlet
• Recall that a Dirichlet distribution is a generalization of a Beta distribution
• And a Multinomial distribution is a generalization of a Binomial distribution
• If a node in a Bayesian network can take 2 values, the analysis is just like the Beta-Binomial example in previous slides
• If it takes more than 2 values, then you have to use a Multinomial-Dirichlet

\[
f(x_1, ..., x_k | n, p_1, ..., p_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1}p_2^{x_2}...p_k^{x_k}
\] for \(\sum x_i = n, p_i \in [0,1], \sum p_i = 1\)

Note: The parameters \(p_1, ..., p_k\) from the multinomial are now the random variables in the Dirichlet prior

\[
f(p_1, ..., p_k | \alpha_1, ..., \alpha_k) = \frac{\Gamma(\alpha_1 + ... + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1-1}...p_k^{\alpha_k-1}
\] for \(p_i \geq 0, \sum p_i = 1\)

Conjugate Priors

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial(x</td>
<td>n, p)</td>
<td>Beta(\alpha, \beta)</td>
</tr>
<tr>
<td>Multinomial(x_1,...,x_k</td>
<td>n, p_1, ..., p_k)</td>
<td>Dirichlet(p_1, ..., p_k</td>
</tr>
</tbody>
</table>

For Beta-Binomial posterior: \(E[p] = \frac{x + \alpha}{n + \alpha + \beta}\)

For Dirichlet-Multinomial posterior: \(E[p_i] = \frac{x_i + \alpha_i}{n + \sum_j \alpha_j}\)

Conjugate Priors

Suppose you were asked to estimate \(P(\text{Price}=\text{Low} \mid \text{Type}=\text{Sedan}, \text{Color}=\text{Silver})\).

Notice that this distribution is a multinomial distribution with \(n = 2\) (because there are 2 records with Color=Silver, Type=Sedan) and \(\mathbb{P}_{\text{Low}}\), \(\mathbb{P}_{\text{Medium}}\), \(\mathbb{P}_{\text{High}}\), corresponding to when Price is low, medium, and high.

Now suppose I tell you to use a Dirichlet prior where all the \(\alpha_i\) are 1.

\[
\text{Estimate } P(\text{Price}=\text{Low} \mid \text{Color}=\text{Silver}, \text{Type}=\text{Sedan}) = \frac{\#(\text{Color}=\text{Silver} \text{ AND Type}=\text{Sedan} \text{ AND Price}=\text{Low}) + 1}{\#(\text{Color}=\text{Silver} \text{ AND Type}=\text{Sedan}) + 3}
\]

\[
= \frac{2 + 1}{2 + 3} = \frac{3}{5}
\]