Conditional distributions
&
Introduction to Bayesian inference (part 2)

Lecture 9b (STAT 24400 F24)

Example 3: the biased coin

We have a coin (which might not necessarily be a fair coin).
We flip n times and count # of Heads:

$$X \sim \text{Binomial}(n, \theta),$$

where $\theta \in [0, 1]$ is the probability of Heads for the coin (often denoted as p).

Goal: θ "=" ?

If we don't know θ in advance, we might think of θ itself as a random draw from all possible coin probabilities — i.e., usually around 0.5, but might be biased.

Our hierarchical model:

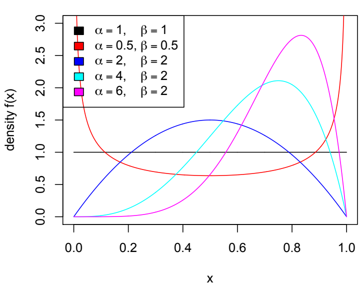
$$\begin{cases} \theta \sim \text{(some distribution over coin probabilities)} \\ X \mid \theta \sim \text{Binomial}(n, \theta) \end{cases}$$

The Beta distribution (definition)

The Beta distribution: supported on $[0, 1]$, with parameters $\alpha, \beta > 0$.
Density:

$$f(t) = \frac{1}{B(\alpha, \beta)} t^{\alpha-1}(1-t)^{\beta-1}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is a normalizing constant (so that density integrates to 1).



- $\text{Beta}(1, 1) = \text{Uniform}[0, 1]$
- $\text{Beta}(c, c)$ for $c < 1$ is symmetric & U-shaped
- $\text{Beta}(c, c)$ for $c > 1$ is symmetric & unimodal
- $\text{Beta}(\alpha, \beta)$ is skewed for $\alpha \neq \beta$

The Beta distribution (basic properties)

Expected value and variance for $\theta \sim \text{Beta}(\alpha, \beta)$:

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}, \qquad \text{Var}(\theta) = \frac{\alpha}{\alpha+\beta} \cdot \frac{\beta}{\alpha+\beta} \cdot \frac{1}{\alpha+\beta+1}$$

- The mean is determined by the *relative* values of α and β
  For example, if $\alpha = \beta$ then $\mathbb{E}(\theta) \equiv \frac{1}{2}$.
- The variance is reduced by increasing the *total* value $\alpha + \beta$

The hierarchical model (for Example 3)

$$\begin{cases} \theta \sim \mathrm{Beta}(\alpha, \beta) & \leftarrow \text{the prior} \\ X \mid \theta \sim \mathrm{Binomial}(n, \theta) & \leftarrow \text{the likelihood} \end{cases}$$

What is $\mathbb{E}(X)$ and $\mathrm{Var}(X)$?

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid \theta)) = \mathbb{E}(n\theta) = n\mathbb{E}(\theta) = \frac{n\,\alpha}{\alpha+\beta}$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}(\mathrm{Var}(X \mid \theta)) + \mathrm{Var}(\mathbb{E}(X \mid \theta)) \\ &= \mathbb{E}(n\,\theta(1-\theta)) + \mathrm{Var}(n\,\theta) \\ &= \frac{n\,\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

The hierarchical model (marginal)

What is the marginal distribution of X?

For each $k = 0, \ldots, n$:

$$\begin{aligned} \mathbb{P}(X = k) &= \mathbb{E}(\mathbb{P}(X = k \mid \theta)) \\ &= \mathbb{E}\left(\binom{n}{k}\theta^k(1-\theta)^{1-k}\right) \\ &= \int_{t=0}^{1} \binom{n}{k} t^k(1-t)^{n-k} f_\theta(t)\,\mathrm{d}t \\ &= \int_{t=0}^{1} \binom{n}{k} t^k(1-t)^{n-k} \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}t \\ &= \frac{\binom{n}{k}}{B(\alpha,\beta)} \int_{t=0}^{1} t^{\alpha+k-1}(1-t)^{\beta+n-k-1}\,\mathrm{d}t \end{aligned}$$

Can we learn about θ? (posterior)

What is the conditional distribution of $\theta \mid X$? (i.e. the posterior of θ)

$$\begin{aligned} f_{\theta|X}(t \mid k) &= \frac{f_\theta(t)\mathbb{P}(X = k \mid \theta = t)}{\mathbb{P}(X = k)} \\ &= \frac{\frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} \cdot \binom{n}{k} t^k(1-t)^{n-k}}{\mathbb{P}(X = k)} \\ &= \begin{pmatrix}\text{value that doesn't} \\ \text{depend on } t\end{pmatrix} \cdot t^{\alpha+k-1}(1-t)^{\beta+n-k-1} \end{aligned}$$

This matches the density of the $\mathrm{Beta}(\alpha+k, \beta+n-k)$ distribution (ignoring the constant)

$$\Rightarrow \qquad \theta \mid X = k \;\sim\; \mathrm{Beta}(\alpha+k, \beta+n-k)$$

Can we learn about θ? (posterior mean & variance)

We therefore have

$$E(\theta \mid X = k) = \underbrace{\frac{\alpha+k}{\alpha+\beta+n}}_{\approx \frac{k}{n} \text{ if } k\&n \text{ large}}$$

$$\mathrm{Var}(\theta \mid X = k) = \underbrace{\frac{\alpha+k}{\alpha+\beta+n} \cdot \frac{\beta+n-k}{\alpha+\beta+n} \cdot \frac{1}{\alpha+\beta+n+1}}_{\text{scales as } O(\frac{1}{n}) \text{ if } k\&n \text{ large}}$$

Can we learn about θ? (summary)

In the Beta-Binomial example:

- If n is large, posterior distrib. has mean $\approx \frac{X}{n}$, and low variance
  (i.e., our posterior belief is that θ is quite close to the observed fraction $\frac{X}{n}$)
- If instead n is small, then our posterior may be quite similar to our prior
  (i.e., our posterior belief isn't very different from our prior belief, since we haven't learned much from a small sample)
- One possible interpretation for α and β is that we previously observed
  α many Heads and β many Tails.

Bayesian statistics

Before observing any data,
our beliefs about θ are expressed via a $\underbrace{\textit{prior distribution}}_{\text{Beta}(\alpha,\beta)}$.

After observing data X, we update our beliefs about θ, by using
the $\underbrace{\textit{likelihood}}_{\text{Binomial}(n,\theta)}$ of X given θ.

The conditional distribution of θ, given X, is the $\underbrace{\textit{posterior distribution}}_{\text{Beta}(\alpha+X,\beta+n-X)}$.

Its conditional expected value is the $\underbrace{\textit{posterior mean}}_{=\frac{\alpha+X}{\alpha+\beta+n}}$.

The mode of the conditional distribution is the $\underbrace{\textit{posterior mode (MAP)}}_{=\frac{\alpha+X-1}{\alpha+\beta+n-2}}$.

Example 4: random start time

Suppose we are running a random process
whose lifespan is distributed as Exponential(1).

We start the process at a random time T drawn uniformly from $[0, 10]$,
and observe that the random process terminates at time X.

What can we infer about the start time T?

Example 4: random start time (likelihood)

What is the distribution of $X \mid T$?

Conditional on $T = t$, $X = t +$ (a draw from the Exponential(1) distribution)

Conditional CDF of $X|T = t$:

$$\mathbb{P}(X \leq x \mid T = t) = 1 - e^{-(x-t)}, \quad x \geq t$$

Conditional density:

$$f_{X|T}(x \mid t) = e^{-(x-t)} \cdot \mathbb{1}_{x \geq t}$$

$$\begin{cases} T \sim \text{Uniform}[0, 10] & \leftarrow \text{the prior} \\ X \mid T \sim \text{density } e^{-(x-T)} \cdot \mathbb{1}_{x \geq T} & \leftarrow \text{the likelihood} \end{cases}$$

Example 4: random start time (posterior)

What is the posterior distribution of T?

Joint density:

$$f_{T,X}(t,x) = f_T(t) f_{X|T}(x \mid t) = 0.1 \cdot \mathbb{1}_{0 \le t \le 10} \cdot e^{-(x-t)} \cdot \mathbb{1}_{x \ge t}$$

Conditional density:

$$f_{T|X}(t \mid x) = \frac{f_{T,X}(t,x)}{f_X(x)} = \frac{0.1 \cdot \mathbb{1}_{0 \le t \le 10} \cdot e^{-(x-t)} \cdot \mathbb{1}_{x \ge t}}{f_X(x)}$$

$$= \begin{pmatrix}\text{value that doesn't} \\ \text{depend on } t\end{pmatrix} \cdot e^t \cdot \mathbb{1}_{0 \le t \le \min\{10,x\}}$$

Since a density must integrate to 1, we can solve for the constant:

$$f_{T|X}(t \mid x) = \frac{e^t \cdot \mathbb{1}_{0 \le t \le \min\{10,x\}}}{e^{\min\{10,x\}} - 1}$$

Bayesian statistics

Before observing any data,
our beliefs about T are expressed via a $\underbrace{\textit{prior distribution}}_{\text{Uniform[0,10]}}$.

After observing data X, we update our beliefs about T, by using
the $\underbrace{\textit{likelihood}}_{\text{density } e^{-(x-t)} \cdot \mathbb{1}_{x \ge t}}$ of X given T.

The conditional distribution of T, given X, is the $\underbrace{\textit{posterior distribution}}_{\text{density } f_{T|X}(t \mid x) \text{ from last slide}}$.

Its conditional expected value is the $\underbrace{\textit{posterior mean}}_{\text{calculate via density } f_{T|X}(t \mid x) \text{ from last slide}}$.

The mode of the conditional distribution is the $\underbrace{\textit{posterior mode (MAP)}}_{=\min\{10,x\}}$.