Conditional distributions
&
Introduction to Bayesian inference (part 1)

Lecture 9a (STAT 24400 F24)

Example 1: calibrating the X-ray

We have an X-ray beam whose output follows a Poisson(λ) distribution.

The intensity parameter λ can be set to 100, 110, 120, or 130.

We choose a setting at random, and then observe $X = 108$ as the output.

Question: Can we infer the setting of the machine based on observed data?

E.g., based on observing $X = 108$, how likely the intensity was set at $\lambda = 120$?

In other words, we are interested in $\mathbb{P}(\lambda = 120 \mid X = 108)$.

In general, we may be interested in $\mathbb{P}(\lambda = \ell \mid X = k)$.

Example 1: calibrating the X-ray (cont.)

Recall the definition of conditional probability: $\mathbb{P}(\lambda = \ell \mid X = k) = \frac{\mathbb{P}(X=k,\lambda=\ell)}{\mathbb{P}(X=k)}$.

For each integer $k \geq 0$ and $\ell = 100, 110, 120, 130$:

$$\mathbb{P}(X = k, \lambda = \ell) = \mathbb{P}(\lambda = \ell)\mathbb{P}(X = k \mid \lambda = \ell) = \frac{1}{4} \cdot \frac{\ell^k e^{-\ell}}{k!}$$

$$\mathbb{P}(X = k) = \textstyle\sum_{\ell'} \mathbb{P}(X = k, \lambda = \ell') \quad \leftarrow \text{law of total probability}$$

So:

$$\mathbb{P}(\lambda = \ell \mid X = k) = \frac{\mathbb{P}(X = k, \lambda = \ell)}{\mathbb{P}(X = k)} = \frac{\frac{1}{4} \cdot \frac{\ell^k e^{-\ell}}{k!}}{\sum_{\ell'} \frac{1}{4} \cdot \frac{\ell'^k e^{-\ell'}}{k!}}$$

The PMF of this conditional distribution, for $k = 108$, is:

	100	110	120	130
$\mathbb{P}(\lambda = \ell \mid X = 108)$	0.306	0.411	0.225	0.058

Review properties of density (prep for the continuous version)

- **Fact 1:** Suppose two density functions $f(x)$ and $g(x)$ satisfy

  $$f(x) = h(x) \cdot (\text{constant that doesn't depend on } x)$$

  and

  $$g(x) = h(x) \cdot (\text{constant that doesn't depend on } x)$$

  for the same function h. Then

  $$f \equiv g$$

  Proof: densities must integrate to 1, so in fact both of the unknown constant values must be equal.

  For example, if two density function $f(x) = 5c_1 xe^{-2\lambda x}$, $g(x) = 3c_2 xe^{-2\lambda x}$ on support $x > 0$, then we can conclude that $f \equiv g$.

Review properties of density (prep for the continuous version)

- **Fact 2:** A random variable X has density $f(x)$.
  Then for small $\epsilon > 0$ ($\epsilon \to 0$),

$$\mathbb{P}(x < X < x + \epsilon) \approx \epsilon \cdot f(x)$$

because (as discussed in lecture 6b) from the definition of integral in calculus,

$$Pr(x < X < x + \epsilon) = \int_{x=x}^{x+\epsilon} f(x)dx \approx \epsilon \cdot f(x)$$

The same property applies to conditional density:

If $f(x|y)$ is the conditional density of X given $Y = y$, then

$$\mathbb{P}(x < X < x + \epsilon | Y = y) \approx \epsilon \cdot f(x|y)$$

Example 2: calibrating the X-ray, continuous version

We have an X-ray beam whose output follows a Poisson(λ) distribution.

The intensity parameter λ can be set to any positive value.

We choose a setting at random by drawing λ from the Exponential(0.01) distribution, and then observe $X = 108$ as the output.

Can we infer the setting of the machine?

Example 2 (hierarchical model formulation)

We may formulate the problem as a **hierarchical model**:

$$\begin{cases} \lambda \sim \text{Exponential}(0.01) \\ X \mid \lambda \sim \text{Poisson}(\lambda) \end{cases}$$

We may be interested in asking:

- What is the marginal distribution of X?
- What is conditional distribution of $\lambda \mid X$?

Example 2 (marginal distribution of observation X)

Mean, variance, and marginal distribution of X:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid \lambda)) \underset{\substack{\uparrow \\ \text{since } X \mid \lambda \text{ is Poisson}(\lambda)}}{=} \mathbb{E}(\lambda) \underset{\substack{\uparrow \\ \text{since } \lambda \text{ is Exponential}(0.01)}}{=} 100$$

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X \mid \lambda)) + \text{Var}(\mathbb{E}(X \mid \lambda)) = \underset{\substack{\uparrow \\ \text{since } X \mid \lambda \text{ is Poisson}(\lambda)}}{\mathbb{E}(\lambda)} + \underset{\substack{\uparrow \\ \text{since } \lambda \text{ is Exponential}(0.01)}}{\text{Var}(\lambda)} = 100 + 100^2$$

Compute marginal PMF: for each $k \geq 0$,

$$\mathbb{P}(X = k) = \mathbb{E}(\mathbb{P}(X = k \mid \lambda)) = \mathbb{E}\left(\frac{\lambda^k e^{-\lambda}}{k!}\right) = \int_{t=0}^{\infty} \frac{t^k e^{-t}}{k!} \left(0.01 e^{-0.01t}\right) \, dt$$

Example 2 (conditional distribution of λ, heuristic intuition)

Conditional distribution of $\lambda \mid X$:

How can we compute the conditional distribution of $\lambda \mid X$, when λ is continuous but X is discrete?

First: informally, let's treat densities and probabilities as interchangeable:

$$\text{“}\mathbb{P}\text{”}(\lambda = t, X = k) = f_\lambda(t) \cdot \mathbb{P}(X = k \mid \lambda = t) = 0.01e^{-0.01t} \cdot \frac{t^k e^{-t}}{k!}$$

$$f_{\lambda|X}(t \mid k) = \frac{0.01e^{-0.01t} \cdot \frac{t^k e^{-t}}{k!}}{\mathbb{P}(X = k)} = \left(\begin{array}{c}\text{some value that doesn't}\\ \text{depend on } t\end{array}\right) \cdot t^k e^{-1.01t}$$

matches the density of the Gamma$(k+1, 1.01)$ distrib.

Example 2 (conditional distribution of λ, brief derivation)

More formally...

Recall: if two densities $f(x)$ and $g(x)$ satisfy $f(x) = C_1 h(x), g(x) = C_2 h(x)$ for the same $h(x)$, where C_1, C_2 are constants that don't depend on x, then $f \equiv g$.

So, we just need to show that

$$f_{\lambda|X}(t \mid k) = \left(\begin{array}{c}\text{value that doesn't}\\ \text{depend on } t\end{array}\right) \cdot t^k e^{-1.01t}$$

Equivalently, we may just show that as $\epsilon \to 0$,

$$\mathbb{P}(t < \lambda < t + \epsilon \mid X = k) \approx \left(\begin{array}{c}\text{value that doesn't}\\ \text{depend on } t \text{ or } \epsilon\end{array}\right) \cdot \epsilon \cdot t^k e^{-1.01t}$$

Example 2 (conditional distribution of λ, brief derivation)

$$\begin{aligned}
\mathbb{P}(t < \lambda < t + \epsilon, X = k) &= \mathbb{E}(\mathbb{P}(t < \lambda < t + \epsilon, X = k \mid \lambda)) \\
&= \mathbb{E}(\mathbb{1}_{t<\lambda<t+\epsilon} \cdot \mathbb{P}(X = k \mid \lambda)) \\
&= \mathbb{E}\left(\mathbb{1}_{t<\lambda<t+\epsilon} \cdot \frac{\lambda^k e^{-\lambda}}{k!}\right) \\
&= \int_{s=0}^{\infty} \mathbb{1}_{t<s<t+\epsilon} \cdot \frac{s^k e^{-s}}{k!} \cdot 0.01e^{-0.01s}\,\mathrm{d}s \\
&= \int_{s=t}^{t+\epsilon} \frac{s^k e^{-s}}{k!} \cdot 0.01e^{-0.01s}\,\mathrm{d}s \\
&\approx \epsilon \cdot \frac{t^k e^{-t}}{k!} \cdot 0.01e^{-0.01t}
\end{aligned}$$

Example 2 (conditional distribution of λ, brief derivation)

Therefore,

$$\begin{aligned}
Pr(t < \lambda < t + \epsilon \mid X = k) &= \frac{\mathbb{P}(t < \lambda < t + \epsilon, X = k)}{\mathbb{P}(X = k)} \\
&\approx \frac{\epsilon \cdot \frac{t^k e^{-t}}{k!} \cdot 0.01e^{-0.01t}}{\text{constant that doesn't depend on } t \text{ or } \epsilon} \\
&\approx \left(\text{value that doesn't depend on } t \text{ or } \epsilon\right) \cdot \epsilon \cdot t^k e^{-1.01t}
\end{aligned}$$

By the property of conditional density we discussed earlier, this implies

$$f_{\lambda|X}(t \mid k) = \left(\begin{array}{c}\text{value that doesn't}\\ \text{depend on } t \text{ or } \epsilon\end{array}\right) \cdot t^k e^{-1.01t}$$

This matches the density of the Gamma$(k+1, 1.01)$ distribution.

Example 2 (conditional distribution of λ)

Conclusion: the conditional density of $\lambda \mid X$ is

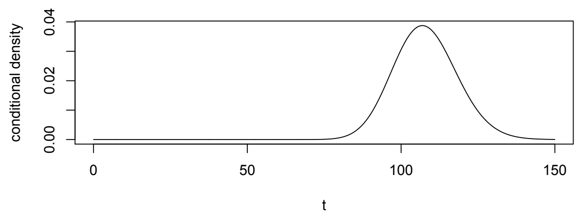
$$f_{\lambda|X}(t \mid k) = \left(\begin{array}{c}\text{value that doesn't} \\ \text{depend on } t \text{ or } \epsilon\end{array}\right) \cdot t^k e^{-1.01t}$$

and therefore, the conditional distribution is

$$\lambda \mid X \sim \text{Gamma}(X+1, 1.01).$$

For example if we observe $X = 108$ then the conditional distribution is

$$\lambda \mid X = 108 \sim \text{Gamma}(109, 1.01)$$



Bayesian statistics (point of view)

In many statistical inference settings, we observe some data (e.g., X)
and want to learn something about the underlying parameter(s) (e.g., λ)

In Bayesian statistics, the underlying parameter is itself treated as a random variable, distributed according to a prior distribution (e.g., Exp(0.01))

Parameters defining the prior distrib. are the hyperparameters (e.g., the 0.01)

The prior distribution may be interpreted in various ways,

– as the true underlying random process the parameter was generated,
– or as reflecting our subjective beliefs,
– or our level of uncertainty about the parameter,
– or may reflect information gathered from past experiments.

Bayesian statistics (prior, likelihood, posterior)

Before observing any data,
our beliefs about λ are expressed via a $\underbrace{\textbf{prior} \text{ distribution}}_{\text{Exponential}(0.01)}$.

After observing data X, we update our beliefs about λ, by using
the $\underbrace{\textbf{likelihood}}_{\text{Poisson}(\lambda)}$ function of X given λ.

The conditional distribution of λ, given X, is the $\underbrace{\textbf{posterior} \text{ distribution}}_{\text{Gamma}(X+1,1.01)}$.

Its conditional expected value is the $\underbrace{\textit{posterior mean}}_{=\frac{X+1}{1.01}}$.

"maximum a posteriori"

The mode of the conditional distribution is the $\underbrace{\textit{posterior mode (MAP)}}_{=\frac{X}{1.01}}$.

(for Example 2 = continuous version)

Bayesian statistics (discrete prior example)

Before observing any data,
our beliefs about λ are expressed via a $\underbrace{\textbf{prior} \text{ distribution}}_{\text{Uniform}\{100,110,120,130\}}$.

After observing data X, we update our beliefs about λ, by using
the $\underbrace{\textbf{likelihood}}_{\text{Poisson}(\lambda)}$ function of X given λ.

The conditional distribution of λ, given X, is the **posterior** distribution.

if observe $X = 108$:

100	110	120	130
0.306	0.411	0.225	0.058

Its conditional expected value is the *posterior mean*.

if observe $X = 108$: $= 0.306 \cdot 100 + 0.411 \cdot 110 + 0.225 \cdot 120 + 0.058 \cdot 130$

The mode of the conditional distribution is the *posterior mode (MAP)*.

if observe $X = 108$: MAP $= 110$

(for Example 1 = discrete version)