Bayesian inference

Lecture 14b (STAT 24400 F24)

Review: the Bayesian framework

$$\begin{cases} \theta & \sim g(\cdot) \quad \leftarrow \text{PMF/density of prior distribution} \\ X_1, \ldots, X_n \mid \theta & \sim f(\cdot \mid \theta) \end{cases}$$

The *posterior distribution* is the conditional distribution of θ
(conditioned on the observed data $X_1, \ldots, X_n$).

Posterior PMF/density:

$$h(t \mid X_1, \ldots, X_n) = \frac{g(t) f(X_1, \cdots, X_n | t)}{f(X_1, \cdots, X_n)}$$

$$= \begin{pmatrix} \text{terms that don't} \\ \text{depend on } t \end{pmatrix} \cdot g(t) \cdot \prod_{i=1}^{n} f(X_i \mid t)$$

Example: Exponential with gamma prior

Suppose the data is drawn from an exponential distribution:

$$X_1, \ldots, X_n \mid \lambda \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$$

Our prior on the parameter λ is:

$$\lambda \sim \text{Gamma}(k, r) \leftarrow \text{shape } k > 0, \text{ rate } r > 0$$

Some facts about Gamma(a, b):

- mean $= \frac{a}{b}$, mode $= \frac{a-1}{b}$ (if $a > 1$)
- For integer k: Gamma(k, b) is the distribution of a sum of k independent Exponential(b) r.v.'s
- So by CLT, if a is large, Gamma$(a, b) \approx \text{N}\big(\frac{a}{b}, \frac{a}{b^2}\big)$

Example: Exponential (cont.)

Back to the example: $X_1, \ldots, X_n \mid \lambda \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$

- Prior density $g(t) = \frac{r^k}{\Gamma(k)} t^{k-1} e^{-rt}$, $t \geq 0$
- Data density $f(x \mid t) = te^{-tx}$, $x \geq 0$
- Posterior density: $h(t|X_1, \ldots, X_n) = \frac{g(t) f(X_1, \ldots, X_n | t)}{f(X_1, \ldots, X_n)}$

$$h(t \mid X_1, \ldots, X_n) = \begin{pmatrix} \text{terms that don't} \\ \text{depend on } t \end{pmatrix} \cdot \frac{r^k}{\Gamma(k)} t^{k-1} e^{-rt} \cdot \prod_{i=1}^{n} te^{-tX_i}, \ t \geq 0$$

$$= \begin{pmatrix} \text{terms that don't} \\ \text{depend on } t \end{pmatrix} \cdot t^{k+n-1} e^{-(r+\sum_i X_i)t}, \ t \geq 0$$

$\rightsquigarrow$ the posterior distribution is Gamma$(k + n, r + \sum_i X_i)$

The Bayesian framework (point estimation via posterior)

The posterior gives a distribution of θ (given observed data).

What if we want a "point estimate", i.e. a single value that is a good estimate for θ?

Two standard options:

- Posterior mean:

$$\widehat{\theta} = \mathbb{E}(\theta \mid X_1, \dots, X_n) \quad \longleftarrow \mathbb{E}(\cdot) \text{ with respect to posterior } h(\cdot \mid X_1, \dots, X_n)$$

- Posterior mode (MAP):

$$\widehat{\theta} = \operatorname*{argmax}_{t \in \Theta} h(t \mid X_1, \dots, X_n)$$

Example: Exponential

Our model:

$$\begin{cases} \lambda \sim \text{Gamma}(k, r) \\ X_1, \dots, X_n \mid \lambda \overset{\text{iid}}{\sim} \text{Exponential}(\lambda) \end{cases}$$

Posterior:

$$\lambda \mid X_1, \dots, X_n \sim \text{Gamma}(k + n, r + \textstyle\sum_{i=1}^n X_i)$$

Recall for Gamma(a, b): mean $= \frac{a}{b}$, mode $= \frac{a-1}{b}$ (if $a > 1$)

$$\Rightarrow \quad \text{Posterior mean} = \frac{k+n}{r + \sum_i X_i}, \quad \text{Posterior mode (MAP)} = \frac{k+n-1}{r + \sum_i X_i}$$

Construction of credible intervals

A $(1-\alpha)$ **credible interval** I (calculated as a function of $X_1, \dots, X_n$) contains $(1-\alpha)$ posterior probability:

$$\mathbb{P}(\theta \in I \mid X_1, \dots, X_n) = 1 - \alpha$$

There are various ways to construct a credible interval.

Two common options:

- Equal tailed interval
- High posterior density interval

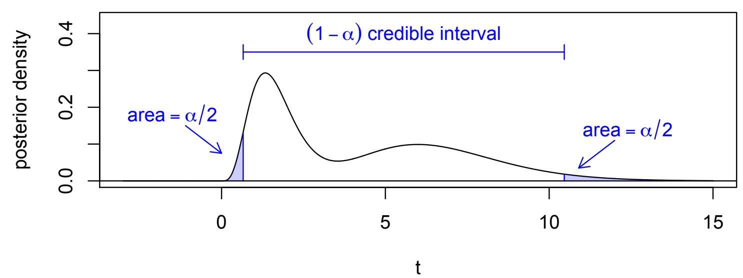
(For a symmetric & unimodal distribution, these options are equivalent)

Remarks: the ideas used in the methods also apply to construction of frequentist confidence intervals with asymmetric and non-unimodal distributions.

Equal tailed credible intervals

- Equal tailed interval: our interval is

$$F^{-1}_{\text{posterior}}(\alpha/2) \le \theta \le F^{-1}_{\text{posterior}}(1 - \alpha/2).$$

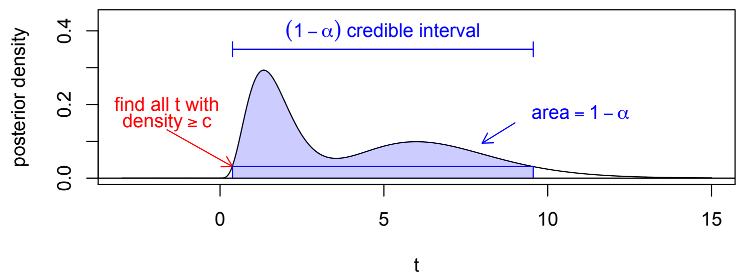


High posterior credible intervals

- High posterior density interval: our interval is given by

$$I = \{t : f_{\theta|X_1,\ldots,X_n}(t \mid x_1,\ldots,x_n) \geq c\}$$

where the density cutoff c is chosen so that prob. $= 1-\alpha$



Note that this region I might not be a single interval!
(In the example above, if α is large, then I splits into two intervals)

Back to Example: Exponential (credible interval)

Our model:

$$\begin{cases} \lambda \sim \text{Gamma}(k,r) \\ X_1,\ldots,X_n \mid \lambda \overset{\text{iid}}{\sim} \text{Exponential}(\lambda) \end{cases}$$

Posterior:

$$\lambda \mid X_1,\ldots,X_n \sim \text{Gamma}(k+n, r+\sum_i X_i)$$

Equal-tailed credible interval:

$$F^{-1}_{\text{Gamma}(k+n,r+\sum_i X_i)}(\alpha/2) \leq \lambda \leq F^{-1}_{\text{Gamma}(k+n,r+\sum_i X_i)}(1-\alpha/2)$$

write $F_{\text{Gamma}(a,b)}$ for the CDF of Gamma(a,b)

Note: the high posterior density interval is more complex to compute (omitted)

Example: exponential (normal approx. credible interval)

- Recall the fact about Gamma(a,b) — For integer a:
  Gamma(a,b) is the distribution of a sum of a indep. *Exponential*(b) r.v.'s.
  So by the CLT, for large integer a, Gamma$(a,b) \approx \text{N}\left(\frac{a}{b}, \frac{a}{b^2}\right)$
- $\Rightarrow$ for any $x \in (0,\infty)$, the CDF $F_{\text{Gamma}(a,b)}(x) \approx \Phi\left(\frac{x-a/b}{\sqrt{a}/b}\right)$
- $\Rightarrow$ for any $t \in (0,1)$,

$$F^{-1}_{\text{Gamma}(a,b)}(t) \approx F^{-1}_{\text{N}(\frac{a}{b},\frac{a}{b^2})}(t) = \frac{a}{b} + \Phi^{-1}(t)\cdot\frac{\sqrt{a}}{b}$$

For example, for $t = 1-\alpha/2$,

$$F^{-1}_{\text{Gamma}(a,b)}(1-\alpha/2) \approx \frac{a}{b} + \Phi^{-1}(1-\alpha/2)\cdot\frac{\sqrt{a}}{b} = \frac{a}{b} + z_{\alpha/2}\cdot\frac{\sqrt{a}}{b}$$

Example: exponential (Bayesian credible interval vs frequentist conf. interval)

Therefore, the $(1-\alpha)$ (equal-tailed) credible interval

$$F^{-1}_{\text{Gamma}(k+n,r+\sum_i X_i)}(\alpha/2) \leq \lambda \leq F^{-1}_{\text{Gamma}(k+n,r+\sum_i X_i)}(1-\alpha/2)$$

is approximately equal to:

$$\approx \frac{k+n}{r+\sum_i X_i} \pm z_{\alpha/2}\cdot\frac{\sqrt{k+n}}{r+\sum_i X_i}$$

If n is large (while k & r are constant), this credible int. is

$$\approx \frac{n}{\sum_i X_i} \pm z_{\alpha/2}\cdot\frac{\sqrt{n}}{\sum_i X_i} = \overbrace{\frac{1}{\bar{X}} \pm z_{\alpha/2}\cdot\frac{1}{\sqrt{n}\cdot\bar{X}}}^{\text{= frequentist interval (using asymp. normality of the MLE)}}$$

Bayes risk

In the Bayesian framework, what's the best way
to choose an estimator $\widehat{\theta}$ to minimize squared loss?

At a *fixed* parameter value θ, the MSE is

$$\mathbb{E}\big((\widehat{\theta}-\theta)^2\big) \quad \leftarrow \mathbb{E}(\cdot) \text{ with respect to } X_1,\dots,X_n \overset{\text{iid}}{\sim} f(\cdot \mid \theta)$$

In a Bayesian framework, should also account for the random distrib. of θ:

$$\text{Bayes risk} = \mathbb{E}\big((\widehat{\theta}-\theta)^2\big) \quad \leftarrow \mathbb{E}(\cdot) \text{ with respect to } \begin{cases} \theta \sim g(\cdot) \\ X_1,\dots,X_n \mid \theta \overset{\text{iid}}{\sim} f(\cdot \mid \theta) \end{cases}$$

Bayes rule (for squared loss)

The **Bayes rule** is the estimator $\widehat{\theta}$ (i.e., the function $\widehat{\theta}(X_1,\dots,X_n)$)
that minimizes Bayes risk

For squared loss:

$$\mathbb{E}\big((\widehat{\theta}-\theta)^2\big) = \underbrace{\mathbb{E}}_{\mathbb{E}(\cdot) \text{ with respect to marginal distrib. of } X_1,\dots,X_n}\Big(\underbrace{\mathbb{E}}_{\mathbb{E}(\cdot) \text{ with respect to posterior distrib. of } \theta \mid X_1,\dots,X_n}\big((\widehat{\theta}-\theta)^2 \mid X_1,\dots,X_n\big) \Big)$$

This is minimized by $\widehat{\theta}$ = posterior mean

Bayes rule (proof for squared loss)

Why does the posterior mean minimize $\mathbb{E}\big((\widehat{\theta}-\theta)^2\big)$?

- For any random variable T and any constant t,

  $$\mathbb{E}\big((T-t)^2\big) = \text{Var}(T-t) + \big(\mathbb{E}(T-t)\big)^2 = \text{Var}(T) + (\mathbb{E}(T)-t)^2$$

  $\Rightarrow$ $\mathbb{E}\big((T-t)^2\big)$ is minimized by choosing $t = \mathbb{E}(T)$

- For any random variables T and S, and any function $t(S)$,
  $\mathbb{E}\big((T-t(S))^2 \mid S\big)$ is minimized by choosing $t(S) = \mathbb{E}(T \mid S)$

$\Rightarrow$ $\widehat{\theta} = \mathbb{E}(\theta \mid X_1,\dots,X_n)$ = posterior mean is the estimator that minimizes

$$\mathbb{E}\big((\widehat{\theta}-\theta)^2 \mid X_1,\dots,X_n\big)$$

i.e., the expected squared error conditional on the data
$\Rightarrow \widehat{\theta}$ must minimize $\mathbb{E}\big((\widehat{\theta}-\theta)^2\big)$.

Other definitions of Bayes risk

We can generalize this to other loss functions loss$(\widehat{\theta},\theta)$, for example:

- Absolute loss:

  Bayes risk $= \mathbb{E}\big(|\widehat{\theta}-\theta|\big)$

  $\rightsquigarrow$ minimized by $\widehat{\theta}$ = posterior median

- 0/1 loss: $\leftarrow$ for the case of a discrete prior (& so the posterior is discrete)

  Bayes risk $= \mathbb{E}(\mathbb{1}_{\widehat{\theta}\neq\theta}) = \mathbb{P}(\widehat{\theta}\neq\theta)$

  $\rightsquigarrow$ minimized by $\widehat{\theta}$ = posterior mode