Bayesian v Frequentist

Giselle Montamat

Harvard University

Spring 2020

Bayesian v Frequentist

Frequentist:

- Assumes data is sampled from some "true" distribution $F_0 = F(., \theta_0)$, where θ_0 is the "true" parameter and is fixed.
- Data is random so every estimator has a sampling distribution; there's a notion of **"repeated experiments"**, i.e, if we could repeat the experiment of drawing a sample of a given size N from the population distribution under the "true" parameter...
  - ...what are the properties of the estimator? Bias (is it correct on average over infinite repeated samples?), variance. Estimators judged based on whether they work well across many repetitions.
  - ...what can we say about the "true" parameter? Confidence set: if the "true" parameter where θ_0, then this value would be covered by the set 95% of the time over infinite repeated samples.
- Confidence sets: statements about the probability of observing the data that we observe if θ_0 adopted a certain value and one could repeat the experiment of drawing data from model. Conclusions that are valid 95% of the time in repeated experiments where new data is drawn from the true distribution given the fixed θ_0.

Bayesian v Frequentist

Bayesian:

- Has beliefs (a prior) about the unknown parameter θ before we collect/see the data. Then data updates beliefs. Posterior distribution summarizes everything we know about θ after seeing the data.
- Conditions on data, doesn't ask what would've happened under repeated experiements but rather what is the one thing that did happen. Given this one sample draw and my prior beliefs...
  - ...what is the updated belief on the distribution of θ? Posterior density. Summarize with: posterior mean, posterior credible sets, etc.
- Posterior credible sets: statements about the posterior probability of θ adopting a certain value, given the observed data (and the model, and the prior).

Posterior credible sets v Confidence sets

Bayesian $1-\alpha$ posterior credible interval (CR): an interval with posterior probability equal to $1-\alpha$:

$$P_{f(\theta|D)}(\theta \in CR(D)) = 1-\alpha$$

In particular, an **equal-tailed set** considers the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ posterior quantiles to build this interval.

How does it compare to a **frequentist $1-\alpha$ confidence set (CS)**? Remember this has coverage at least $1-\alpha$ for all possible "true" θ_0, i.e.:

$$\inf_{\theta_0} \; P_{f(D|\theta_0)}(\theta_0 \in CS(D)) = 1-\alpha$$

- *Small samples*: credible sets don't in general have correct coverage from frequentist perspective:

$$P_{f(D|\theta_0)}(\theta_0 \in CR(D)) < 1-\alpha$$

- *Large samples*: as $N \to \infty$ (+correct model specification, regularity conditions) theorems (Berstein-von Mises, Cherenozhukov and Hong) imply that credible sets have a frequentist interpretation, aka they are the confidence sets that a frequentist would build based on asymptotic normality.

Posterior credible sets v Confidence sets

Exercise: Pset 7, Exercise 1 asked you to derive a 95% confidence interval and a 95% credible interval in a specific example and show that the coverage of the first is at least 95% while the second has coverage 0.

Some technical hints to remember:

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

If $Z = X$ but restricted to interval $[a, b]$:

$$f(z) = 1\{z \in [a,b]\}\frac{\phi\left(\frac{z-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

$$F(z) = \frac{\Phi\left(\frac{z-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

Posterior credible sets v Confidence sets

You observe a single data point $D \sim N(\theta, 1)$ and want to estimate θ using maximum likelihood. In particular, you are given the additional info that θ is non-negative.

$$\text{Likelihood: } f(d|\theta) = \phi(d-\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(d-\theta)^2}$$

- "Unrestricted" MLE estimator and confidence set:

$$\hat{\theta}_U = \underset{\theta}{arg\,max} - (D-\theta)^2$$

$$\hat{\theta}_U = D$$

$$CS_U = [\hat{\theta}_U - 1.96, \hat{\theta}_U + 1.96] = [D - 1.96, D + 1.96]$$

Coverage:

$$P_{\theta_0}(\theta_0 \in CS_U) = P_{\theta_0}(-1.96 \leq \hat{\theta}_U - \theta_0 \leq 1.96) = 0.95 \quad \forall \theta_0$$

$$\Rightarrow \quad \underset{\theta_0}{inf}\, P_{\theta_0}(\theta_0 \in CS_U) = 0.95$$

Posterior credible sets v Confidence sets

- "Restricted" MLE estimator and confidence set (imposes $\theta \geq 0$):

$$\hat{\theta}_R = \underset{\theta}{arg\,max} - (D - \theta)^2$$

$$s.t.\ \ \theta \geq 0$$

$$\hat{\theta}_R = max\{D, 0\}$$

$$CS_R = [\hat{\theta}_R - 1.96, \hat{\theta}_R + 1.96] \cap \mathbb{R}_+ = \begin{cases} CS_U \cap \mathbb{R}_+ \ if\ \ D \geq 0 \\ [-1.96, 1.96] \cap \mathbb{R}_+ \ if\ \ D < 0 \end{cases}$$

Coverage:

$$\underset{\theta_0 \in \mathbb{R}_+}{inf}\ P_{\theta_0}(\theta_0 \in CS_R) \geq 0.95$$

To prove this, need to show $\theta_0 \in CS_U \cap \mathbb{R}_+ \Rightarrow \theta_0 \in CS_R$ because this implies $0.95 = \underset{\theta_0 \in \mathbb{R}_+}{inf}\ P_{\theta_0}(\theta_0 \in CS_U) \leq \underset{\theta_0 \in \mathbb{R}_+}{inf}\ P_{\theta_0}(\theta_0 \in CS_R)$

Posterior credible sets v Confidence sets

- Bayesian estimator considered is the posterior mean. The prior incorporates the knowledge of θ being non-negative but otherwise is flat.

$$\text{Prior: } \pi(\theta) = 1\{\theta \in \mathbb{R}_+\}$$

Find the posterior:

$$f(\theta|d) = \frac{f(d|\theta)\pi(\theta)}{\int f(d|\theta)\pi(\theta)d\theta} = \frac{\phi(d-\theta)\,1\{\theta \in \mathbb{R}_+\}}{\int \phi(d-\theta)\,1\{\theta \in \mathbb{R}_+\}d\theta}$$

$$= 1\{\theta \in \mathbb{R}_+\}\frac{\phi(d-\theta)}{\int_{\mathbb{R}_+}\phi(d-\theta)\,d\theta} = 1\{\theta \in \mathbb{R}_+\}\frac{\phi(\theta-d)}{\int_{\mathbb{R}_+}\phi(\theta-d)\,d\theta}$$

$$= 1\{\theta \in \mathbb{R}_+\}\frac{\phi(\theta-d)}{1-\Phi(-d)}$$

This is a truncated normal density for $\theta \sim N(d,1)$ and the restriction $\theta \in \mathbb{R}_+$!

Posterior credible sets v Confidence sets

Credible set:

$$CR = [\theta_{.025}, \theta_{.975}]$$

$$P_{f(\theta|d)}(\theta \leq \theta_{.025}) = \frac{\Phi(\theta_{.025} - d) - \Phi(-d)}{1 - \Phi(-d)} = 0.025$$

$$P_{f(\theta|d)}(\theta \leq \theta_{.975}) = \frac{\Phi(\theta_{.975} - d) - \Phi(-d)}{1 - \Phi(-d)} = 0.975$$

$\Rightarrow CR =$
$[d + \Phi^{-1}(0.025(1 - \Phi(-d) + \Phi(-d)), d + \Phi^{-1}(0.975(1 - \Phi(-d) + \Phi(-d))]$

Coverage:

$$\inf_{\theta_0 \in \mathbb{R}_+} P_{\theta_0}(\theta_0 \in CR) = 0$$

To see this, take one value in the set $\mathbb{R}_+$, namely 0. Because $0 < \theta_{.025}$ in the posterior distribution, it is not included in CR. So $P_0(0 \in CR) = 0$.

Shrinkage and penalized estimators

Frequentist methods for high-dim models often rely on shrinkage or penalization, which is similar to imposing prior. Examples:

James-Stein estimator

Remember exercise 4) when we were discussing risk functions? You observe $D = (D_1, ..., D_k)$ and want to estimate $\theta = (\theta_1, ..., \theta_k)$:

$$\text{Likelihood: } D_i \sim N(\theta_i, 1)$$

$$\hat{\theta}_{JS} = \left(1 - \frac{k-2}{\sum_{i=1}^{k} D_i^2}\right) D$$

This shrinkage estimator can be derived as a (sort of) Empirical Bayes estimator. In particular, considers the posterior mean derived from this likelihood + a normal prior, and estimates hyperparameters of prior using data.

$$\text{Prior: } \theta_i \sim N(0, \Omega)$$

Shrinkage and penalized estimators

Posterior mean:

$$E[\mu|D] = \Omega(1+\Omega)^{-1}D = \frac{\Omega}{1+\Omega}D = \left(1 - \frac{1}{\Omega+1}\right)D$$

(Note: the MLE estimator is $\hat{\theta} = D$. So $\frac{\Omega}{1+\Omega}$ is shrinking the MLE estimator toward 0.)

Final step is to estimate Ω using the data D. In particular, note that:

$$\text{Marginal: } D \sim N(0, (\Omega+1)I_k)$$

$$||D||^2 = \sum_{i=1}^{k} D_i^2 \sim (\Omega+1)\chi_k^2$$

So an unbiased estimator of $\frac{\Omega}{1+\Omega}$ is $\frac{k-2}{\sum_{i=1}^{k} D_i^2}$ $\Rightarrow$ we arrive at the James-Stein estimator!

(PS: the Empirical Bayes estimator would estimate Ω by maximizing the marginal likelihood $f(D|\Omega)$)

Shrinkage and penalized estimators

Ridge regression Pset 10, Exercise 2 asked you to work through two types of penalized estimators (ridge and lasso). In particular, ridge regression solves:

$$\hat{\beta} = \underset{\beta}{arg\,min}(Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

$$\hat{\beta} = \underset{\beta}{arg\,min} \sum_i (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

Where $\lambda \sum_{j=1}^{k} \beta_j^2$ imposes a "penalty" for choosing a β that is too large.

Shrinkage and penalized estimators

The exercise asked you to find the estimator (solve the min problem) and find the bias and variance of the estimator. For simplicity, it told you to work with the assumption that $X'X$ is diagonal. The more general solution yields:

$$\hat{\beta} = \underset{\beta}{arg\,min} Y'Y - 2Y'X\beta + \beta'X'X\beta + \lambda\beta'\beta$$

FOC:

$$-X'Y + X'X\beta + \lambda\hat{\beta} = 0 \Rightarrow \hat{\beta} = [X'X + \lambda \cdot I]^{-1}X'Y$$

Note that $\lambda \cdot I$ is shrinking the OLS $\hat{\beta}$ closer to 0. The higher the penalty parameter λ, the more it shrinks.
This estimator can be derived even when there is multicollinearity, for example due to the fact that $k > N$ (while OLS can't).

Shrinkage and penalized estimators

The ridge regression estimator can be derived as a Bayesian estimator in the following context:

$$Y = X\beta + \epsilon \;\; ; \;\; \epsilon|X \sim N(0, \sigma^2 \cdot I)$$

$$\text{Likelihood: } \; Y|X\beta \sim N(X\beta, \sigma^2 \cdot I)$$

$$\text{Prior: } \; \beta|X \sim N(0, \gamma^2 \cdot I)$$

Posterior mean:

$$E[\beta|Y, X] = \left(X'X + \frac{\sigma^2}{\gamma^2} \cdot I\right)^{-1} X'Y$$

So the term $\frac{\sigma^2}{\gamma^2} \cdot I$ is shrinking the OLS estimator closer to 0 (but less so the higher the variance of the prior). If we take $\lambda = \frac{\sigma^2}{\gamma^2}$, then this is the ridge regression estimator and the penalty parameter can be motivated as the ratio of the variance of the likelihood over the prior (the lower the variance of the prior, the more you penalize in order to remain close to 0).

Bayesian v Frequentist

So we've seen examples where the Bayesian estimator (e.g., the posterior mean) shrinks the MLE toward 0 (could be some other a priori value). An Empirical/Hierarchical Bayes estimator, moreover, would go one step further to learn from the data how much shrinkage is needed.

Frequentist penalized/shrinkage estimators do the same (examples: JS, ridge regression). In fact, we've seen how these estimators can be derived as Bayesian estimators.

In conclusion, one can find links between the Bayesian and Frequentist approaches: consistency of posterior mean and the frequentist MLE, posterior credible sets and frequentist confidence sets, Bayesian estimators and frequentist shrinkage/penalized estimators, etc.