Mathematics/Statistics Bootcamp

Part VI: Bayesian Statistics

Steven Winter   Christine Shen

Department of Statistical Science
Duke University

MSS Orientation, August 2022

Overview

Introduction to Bayesian Statistics

Axioms of Probability

1. For any event A, $\mathbb{P}(A) \in [0, 1]$;
2. Let Ω denote the sample space, $\mathbb{P}(\Omega) = 1$;
3. If $A_1, A_2, \ldots$ are disjoint events, then

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_{i=1} \mathbb{P}(A_i).$$

Interpretations of Probability

Three classical interpretations of probability are:

1. **Symmetry**: if exactly one of $k \in \mathbb{N}$ events A_i will occur and each equally likely, then $\mathbb{P}[A_i] = 1/k$.
2. **Frequency**: if an event A may be repeated independently over and over,

$$\mathbb{P}[A] = \lim_{n \to \infty} \frac{1}{n}(\text{number of times event } A \text{ occurs}).$$

3. **Degree of Belief**: if you are indifferent between two games:
   - win \$1 if event A occurs and 0 otherwise;
   - win \$1 if a blue ball is drawn from a well-mixed urn containing $100p\%$ blue balls and 0 otherwise,

   then your subjective probability (belief) of event A is p.

Interpretations of Probability

These three interpretations all satisfy the axioms of probability, but with increasing applicability. For example,

1. Symmetry
   - what about the probability of "rain vs sunshine"?
2. Frequency
   - what about the probability of "Duke beats UNC in basketball this year?"

Two Paradigms: Frequentist vs Bayesian

Frequentists view probability as a measure of long-term frequency.

Bayesians use probability to quantify individual degree of belief. The goal is to update one's uncertainty and belief based on data. **Bayesian inference** refers to process of inductive learning via Bayes' rule.

Frequentist vs Bayesian - Example

Suppose we are interested in the probability of landing on heads of a coin.

- Goal: learn about θ, probability of landing on heads
- Parameter space: $\Theta = [0, 1]$
- Data: x, total number of heads in a sample of $n = 10$ tosses
- Sample space: $\Omega = \{0, \dots, 10\}$
- Let X be a random variable for the (random) data to be collected. Posit the following sampling model:

$$X \mid \theta \sim Bin(n, \theta).$$

Note the difference in notations typically used by Frequentists vs Bayesians: $P_\theta(x)$ *vs* $P(x \mid \theta)$.

Classical Frequentist Inference - Exercise

Find the MLE $\hat{\theta}$, calculate its bias, variance and MSE.

Also,

1. follow the *Inference* slides, compute the Fisher Information $I(\theta)$ and find the asymptotic distribution of the MLE.
2. plug in the MLE to get the observed Fisher Information $I(\hat{\theta})$, and construct an asymptotic level (1-α) confidence interval for θ.

Frequentist vs Bayesian View

Frequentist view

1. If we toss the coin infinite number of times, the proportion of tosses landing on heads is θ.
2. θ is **fixed** and unknown. What's **random** is the sample.
3. Uncertainties come from sampling errors in the experiments.

Bayesian view

1. While θ is unknown, we might have certain beliefs/ knowledge about θ before seeing the data.
2. The data, once observed, is **fixed**.
3. We want to use the data to update our beliefs/ uncertainties about θ.

Prior Distribution

We typically use a **prior distribution** $p(\theta)$ to quantify our beliefs on parameter θ prior to seeing the data.

- E.g., unless we have any specific reasons, we might a priori believe that it's likely a fair coin, though perhaps with high uncertainty.
- We can use $p(\theta) \sim Beta(a, b)$ with $a = 2,\ b = 2$ to capture this prior belief.

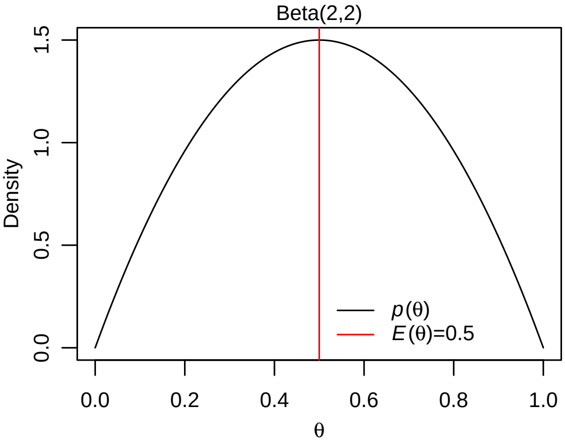
Note:

- *Beta* distributions are defined on [0,1]. $Beta(a, b)$ has a mean of a/b, and variance

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

- Mean and SD for $Beta(2, 2)$ are 0.50 and 0.22.

Density of Prior Distribution



Bayes Rule and Posterior Distribution

We want to update our belief about θ based on the observed data $X = x$ and the sampling model, i.e., we are interested in $p(\theta \mid x)$.

Recall the Bayes' theorem:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}.$$

Therefore,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

This is called the **posterior distribution**. It quantifies our beliefs about parameter θ after observing data $X = x$.

Recap

Elements of classical Frequentist inference:

- fixed parameter θ;
- sampling model $p(x \mid \theta)$ (or alternatively denote as $p_\theta(x)$);
- (imaginary) random data X and the observed data x.

Elements of Bayesian inference:

- prior distribution of the parameter $p(\theta)$;
- sampling model $p(x \mid \theta)$;
- posterior distribution $p(\theta \mid x)$ via the Bayes rule after observing data $X = x$.

Note: a key difference is, Bayesian admits *prior information* for inference on θ.

Discussion

- Does the prior distribution contain additional information compared to the data? Will it affect inference results?
- Suppose two researchers observe the same data, but have different priors and hence reach different conclusions. Is this reasonable? Is it legitimate to incorporate subjective beliefs in inference?
- Why and when is prior information helpful?
- What if we don't have any prior information? Can we, and should we still use Bayesian inference?

Derivation of the Posterior Distribution

How to derive the posterior?

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

That is,

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{normalizing constant}}$$

In practice, normalizing constant is often intractable. But recall the kernel trick!

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \left[\binom{n}{x}\theta^x(1-\theta)^{n-x}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right] \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}. \end{aligned}$$

Derivation of the Posterior Distribution

Notice

$$p(\theta|x) \propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}$$

is the kernel for a $Beta(x+a, n-x+b)$ distribution. But is this sufficient to conclude this is the posterior distribution?

Yes! The posterior is a proper probability distribution and thus its PDF integrates to 1 over the parameter space. Therefore recognizing the kernel is sufficient to identify the posterior.

Derivation of the Posterior Distribution

$$
\begin{aligned}
1 &= \int_0^1 c(x)\theta^{x+a-1}(1-\theta)^{n+b-x-1}d\theta \\
&\implies 1 = c(x)\int_0^1 \theta^{x+a-1}(1-\theta)^{n+b-x-1}d\theta \\
&\implies 1 = c(x)\frac{\Gamma(x+a)\Gamma(n+b-x)}{\Gamma(n+a+b)} \\
&\implies c(x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n+b-x)} \\
&\implies p(\theta|x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n+b-x)}\theta^{x+a-1}(1-\theta)^{n+b-x-1} \\
&\implies p(\theta|x) \sim Beta(x+a, n+b-x).
\end{aligned}
$$

Conjugacy

We have seen the beta-binomial model, i.e.,

- Beta prior $p(\theta) \sim Beta(a, b)$,
- Binomial sampling model $X \sim Bin(n, \theta)$, gives
- Beta posterior $p(\theta \mid x) \sim Beta(x + a, n - x + b)$.

We say the Beta distribution is **conjugate** for the Binomial sampling model.

Formally, a class $\mathcal{P}$ of prior distributions for θ is called **conjugate** for a sampling model $p(x|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}.$$

Conjugacy

Advantages:

- computational convenience;
- interpretability.

Limitations:

- inflexible, limited applicability to complex problems.

Other conjugate prior and sampling models (see this link):

- Gamma prior for Poisson model
- Dirichlet prior for Multinomial model
- Normal prior for Normal model
- ...

Normal Model - Exercise

Suppose our model is $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$ with σ^2 known. Our prior belief for θ is

$$p(\theta) \sim N(\mu, \tau^2),$$

with known μ and τ^2. Find the posterior distribution of θ based on observations $x_1, \ldots, x_n$.

Hint: follow the steps for the beta-binomial model

- find $p(x_1, \ldots, x_n \mid \theta)$,
- find $p(\theta)$, and
- identify the kernel of $p(\theta \mid x_1, \ldots, x_n)$ (recall a trick called *completing the squares*).

Multivariate Normal Model - Exercise

Now consider p-dimensional random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known. Our prior belief about $\boldsymbol{\theta}$ is encoded as

$$p(\boldsymbol{\theta}) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}),$$

with known $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$. Find the posterior distribution of $\boldsymbol{\theta}$ based on observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$.

Bonus question: like the uni-variate normal example, can you identify the posterior distribution without completing the squares?

A Peek into the Bayesian Course

What we have seen:

- one-parameter model
- conjugacy

What you'll learn:

- more flexible models with multiple parameters
- semi-conjugacy
- Gibbs sampling
- Metropolis-Hasting algorithm
- ...

Bayesian Inference

Coin Toss Example

Recall the earlier example where we are interested in the probability of a coin landing on heads.

- Goal: learn about θ, probability of landing on heads
- Parameter space: $\Theta = [0, 1]$
- Data: x, total number of heads in a sample of $n = 10$ tosses
- Sample space: $\Omega = \{0, \ldots, 10\}$
- Sampling model: $X \mid \theta \sim Bin(n, \theta)$
- Prior: $p(\theta) \sim Beta(a, b)$ with $a = 2$, $b = 2$
- Posterior: $p(\theta|x) \sim Beta(x + a, n - x + b)$

Classical Frequentist Inference

We have gone through the key elements of classical inference:

- The MLE is $\hat{\theta} = X/n$, unbiased, with MSE $= \theta(1-\theta)/n$
- $\hat{\theta}$ is asymptotically normal,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{\theta(1-\theta)}{n}\right).$$

- An asymptotic level (1-α) confidence interval for θ is

$$\left(\hat{\theta} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right).$$

How does Bayesian inference compare to this?

Bayesian Estimation

Suppose we observe $x = 4$, i.e., 4 out of 10 tosses land on heads.

Under classical inference,

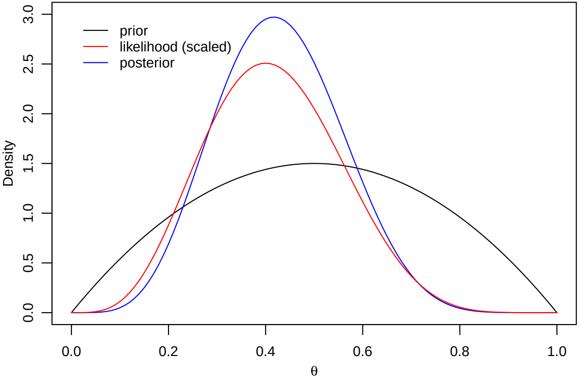
- the maximum likelihood estimate of θ is $4/10 = 0.4$.
- It's a "best" guess of the *true value* of θ based on this sample.

Under Bayesian inference,

- beliefs about θ are updated through Bayes rule to the posterior distribution: $Beta(6, 8)$.
- One estimator of θ is the posterior mean:

$$\hat{\theta}_B = \mathbb{E}[\theta \mid x] = \frac{a + x}{a + b + n} = \frac{6}{6 + 8} \approx 0.43.$$

Bayesian Updating



Posterior Mean Estimator

Let's take a closer look at the posterior mean estimator.

$$\begin{aligned}\hat{\theta}_B &= \frac{a+X}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{X}{n} \\ &= (1-\omega)\mathbb{E}(\theta) + \omega\hat{\theta}, \quad \omega = \frac{n}{a+b+n} \approx 0.71.\end{aligned}$$

It's a weighted average of the prior mean and the MLE.

Exercise: compute the bias, variance and MSE for $\hat{\theta}_B$.
Hint: recall the linear shrinkage estimator?

Bayesian Credible Interval

We can obtain (1-α) credible intervals based on the posterior distribution of θ.

An interval $[L(x), U(x)]$, based on the observed data $X = x$ has (1-α) **Bayesian coverage** for θ if

$$\mathbb{P}(L(x) < \theta < U(x) \mid X = x) = 1 - \alpha.$$

Recall, a random interval $[L(X), U(X)]$ has (1-α) **frequentist coverage** for θ if, before the data are gathered,

$$\mathbb{P}(L(X) < \theta < U(X) \mid \theta) = 1 - \alpha.$$

Bayesian Credible Interval

One easy way to obtain a credible interval is to use the posterior quantiles.

Find $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ such that

$$\mathbb{P}(\theta < \theta_{\alpha/2} \mid x) = \alpha/2$$
$$\mathbb{P}(\theta < \theta_{1-\alpha/2} \mid x) = 1 - \alpha/2.$$

Then $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$ has $(1-\alpha)$ Bayesian coverage.

Other Bayesian Inference Tools

Hypothesis testing

- $H : \theta \in \Theta_H$, vs $K : \theta \in \Theta_K$
- Sampling model: $p(x \mid \theta)$
- Prior beliefs on null and alternative $p(H)$ and $p(K)$
- Posterior beliefs after observing data $X = x$

$$p(H \mid x) = \frac{p(x \mid H)p(H)}{p(x)}, \quad p(K \mid x) = \frac{p(x \mid K)p(K)}{p(x)}$$

- Test rule: reject if the *Bayes Factor*

$$\frac{p(x \mid K)}{p(x \mid H)} = \frac{p(K \mid x)}{p(H \mid x)} \frac{p(H)}{p(K)}.$$

is large.

Other Bayesian Inference Tools

Posterior p-value

Recall the classical frequentist p-value for $H : \theta = \theta_0$ is a statistic

$$P(x, \theta_0) = \mathbb{P}(T(X) \geq T(x) | \theta_0).$$

The posterior p-value is defined as

$$\int P(x, \theta) \pi(\theta \mid x) d\theta,$$

where $\pi(\theta \mid x)$ denotes the posterior distribution for θ.

Other Bayesian Inference Tools

Posterior prediction

After observing some data, we can make predictions via the posterior predictive distribution, which reflects our updated beliefs/ uncertainties about the parameter.

Let x^{obs} denote the observed data, and x denote any potential new data, we are interested in

$$p(x \mid x^{obs}).$$

Other Bayesian Inference Tools

Posterior prediction

$$\begin{aligned}
p(x \mid x^{obs}) &= \int p(x, \theta \mid x^{obs}) d\theta \quad \text{(joint to marginal)} \\
&= \int p(x \mid \theta, x^{obs}) p(\theta \mid x^{obs}) d\theta \quad \text{(conditional distribution)} \\
&= \int p(x \mid \theta) p(\theta \mid x^{obs}) d\theta \quad \text{(conditional independence)}.
\end{aligned}$$

More generally,

$$p(x \mid \boldsymbol{a}) = \int p(x \mid \boldsymbol{b}) p(\boldsymbol{b} \mid \boldsymbol{a}) d\boldsymbol{b},$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ can be any vectors.

Summary

Frequentist vs Bayesian

Though procedures vary by projects, typical inference pipelines are:

Frequentist

1. identify sampling model, parameter(s) of interest
2. obtain point estimates/ intervals/ tests/ predictions via e.g.:
   - numerical optimization (e.g., EM algorithm) for MLE
   - bootstrapping for intervals/ tests
   - approximation with asymptotics

Bayesian

1. identify sampling model, parameter(s) of interest, priors
2. identify posterior and obtain posterior samples (typically via Markov Chain Monte Carlo)
3. all kinds of analysis can be done based on the posterior samples. E.g., credible intervals, predictions, $\mathbb{P}(3\theta_1 + \cos(\theta_2) - \exp(\theta_3) < 0.456)$, etc...

Frequentist vs Bayesian

Frequentist, or Bayesian?

- What would be a scenario where the Frequentist approach is more appropriate?
- What would be a scenario where a Bayesian approach works better?

Both are useful statistical tools to help solve problems, answer questions, and understand the world.

Reference Guide

- *A First Course in Bayesian Statistical Methods* - Hoff
- *Why isn't everyone a Bayesian* - Efron

Acknowledgement

Past contributors:

- Jordan Bryan, PhD student
- Brian Cozzi, MSS alumni
- Michael Valancius, MSS alumni
- Graham Tierney, PhD student
- Becky Tang, PhD student

This set of slides made reference to

- 2020F STA711 course materials