Bayesian Statistics

Anita Wang

2017 - 2018

Bayesian Theorem

$$\underbrace{p(\theta|y)}_{posterior\ density} = \frac{p(\theta,y)}{p(y)} = \frac{\overbrace{p(\theta)}^{prior\ distribution}\overbrace{p(y|\theta)}^{data\ distribution}}{p(y)} = \frac{p(\theta)p(y|\theta)}{\sum_\theta p(\theta)p(y|\theta)}$$

An Example: College Students Sleeping

- Parameter p: the proportion of American college students who sleep at least eight hours.
- A sample of 27 students is taken. In this group, 11 record that they had at least eight hours of sleep the previous night
- Discrete prior probability:

$$\begin{cases} \Pr(p = 0.2) = 0.6 \\ \Pr(p = 0.4) = 0.3 \\ \Pr(p = 0.7) = 0.1 \end{cases}$$

An Example: College Students Sleeping

The posterior probability:

$$Pr(p=0.2|y)=\frac{Pr(p=0.2)Pr(y|p=0.2)}{\sum Pr(p)\Pr(y|p)}$$

$$=\frac{0.6*\binom{27}{11}0.2^{11}0.8^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16}+0.3\binom{27}{11}0.4^{11}0.6^{16}+0.1\binom{27}{11}0.7^{11}0.3^{16}}=0.089$$

$$Pr(p=0.4|y)=\frac{Pr(p=0.4)Pr(y|p=0.4)}{\sum Pr(p)\Pr(y|p)}$$

$$=\frac{0.3*\binom{27}{11}0.4^{11}0.6^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16}+0.3\binom{27}{11}0.4^{11}0.6^{16}+0.1\binom{27}{11}0.7^{11}0.3^{16}}=0.909$$

$$Pr(p=0.7|y)=\frac{Pr(p=0.7)Pr(y|p=0.7)}{\sum Pr(p)\Pr(y|p)}$$

$$=\frac{0.1*\binom{27}{11}0.7^{11}0.3^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16}+0.3\binom{27}{11}0.4^{11}0.6^{16}+0.1\binom{27}{11}0.7^{11}0.3^{16}}=0.002$$

Bayesian Thinking

- Parameter θ is unknown and to be estimated
- Previously, we use sample data information to estimate θ (For example, sample proportion $\hat{p}$ to estimate population proportion p)
- Bayesian thinking:
  1) Prior information of the parameter: the subject prior opinion of the distribution of the parameter
  2) Sample data information
  3) Posterior distribution: combine the information in the data with the prior distribution

Statistical Inference

Two main approaches

- Frequentist

Model parameters are fixed unknown quantities.
Randomness only in data.

  - Estimation - Maximum likelihood, method of moments
  - Confidence intervals
  - Significance testing - p-values
  - Hypothesis testing - Reject/Don't Reject H_0

Statistical Inference

- Bayesian

Model parameters are random variables. Inference is based on $P(\theta|\mathrm{Data})$, the posterior distribution given the data.

  - Estimation - Posterior means, modes
  - Credible intervals/sets
  - Posterior probabilities

Bayes' Rule

An equivalent form omits the factor $p(y)$, which does not depend on θ and, with fixed y, can thus be considered a constant, yielding the **unnormalized posterior density**,

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The second term in this expression, $p(y|\theta)$, is taken here as a function of θ, not of y.

Prediction

- **Prior predictive distribution** (also called marginal distribution of y)

$$p(y) = \int p(y,\theta)d\theta = \int p(\theta)p(y|\theta)d\theta$$

prior because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

Prediction

- Posterior predictive distribution

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y},\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \end{aligned}$$

Once the data y have been observed, the unknown observable $\tilde{y}$ can be predicted. For example, $y = (y_1, y_2, \ldots, y_n)$ may be the vector of recorded weights of an object weighed n times on a scale, $\theta = (\mu, \sigma^2)$ is the prior, and $\tilde{y}$ may be the yet to be recorded weight of the object in a planned new weighing.

Simulate the Posterior Predictive Distribution

- Assuming that you can simulate from the posterior distribution of the parameter, which is usually feasible.
- To simulate the posterior predictive distribution involves two steps:

1. Simulate θ_i from $\theta|y$; $i = 1, \dots, m$
2. Simulate $\tilde{y}_i$ from $\tilde{y}|\theta_i$ $(= \tilde{y}|\theta_i, y)$

The pairs $(\theta_i, \tilde{y}_i)$ are draws from the joint distribution $\theta, \tilde{y}|y$. Therefore the $\tilde{y}_i$ are draws from $\tilde{y}|y$.

Single parameter model

- Single parameter model is statistical models where only a single scalar parameter is to be estimated; that is, the estimand θ is **one-dimensional**

In this chapter:

- **Binomial**
- **Normal**
- **Poisson**
- **Exponential**

Binomial

- In the simple binomial model, the aim is to estimate an **unknown population proportion** from the results of a sequence of 'Bernoulli trials'; that is, data $y_1, \ldots, y_n$.
- Because of the exchangeability, the data can be summarized by the total number of successes in the n trials, which we denote here by y.
- The binomial sampling distribution is

$$p(y|\theta) = \mathrm{Bin}(y|n, \theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y}$$

where on the left side we suppress the dependence on n because it is regarded as part of the experimental design that is considered fixed

Example

- We consider the estimation of the sex ratio within a population of human births. The currently accepted value of the proportion of female births in large European-race populations is 0.485.
- Let y be the number of girls in n recorded births. we are assuming that the n births are conditionally independent given θ, with the probability of a female birth equal to θ for all cases.
- For simplicity, we assume that the prior distribution for θ is uniform on the interval [0, 1].
- The posterior density,

$$p(\theta|y) \propto \theta^{y}(1-\theta)^{n-y}.$$

Different prior densities

- We consider a parametric family of prior distributions that includes the uniform as a special case and construct a family of prior densities that lead to simple posterior densities.
- $\theta \sim \text{Beta}(\alpha, \beta)$:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

- this prior density is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures.
- The posterior density,

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y, \beta+n-y). \end{aligned}$$

Conjugate prior

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**; the beta prior distribution is a **conjugate family** for the binomial likelihood.
- If F is a class of sampling distributions $p(y|\theta)$, and P is a class of prior distributions for θ, then the class P is conjugate for F if

$$p(\theta|y) \in P \text{ for all } p(\cdot|\theta) \in F \text{ and } p(\cdot) \in P.$$

- This definition is formally vague since if we choose P as the class of all distributions, then P is always conjugate no matter what class of sampling distributions is used.

Normal mean with known variance: a single observation

- Consider a single scalar observation y from a normal distribution parameterized by a mean θ and variance σ^2, where for this initial development we assume that σ^2 is known.

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

The conjugate prior $\theta \sim N(\mu_0, \tau_0^2)$

$$p(\theta) \propto \exp(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2)$$

hyperparameters μ_0 and τ_0^2.

Posterior distribution

$$p(\theta|y) \propto \exp(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2)$$

$$\theta|y \sim N(\mu_1, \tau_1^2)$$

where

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \text{ and } \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- the posterior **precision** equals the prior precision plus the data precision.

Normal mean with known variance: more observations

- more realistic situation:

a sample of independent and identically distributed observations y = $(y_1, \ldots, y_n)$ is available.

- Posterior density:

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) \\
&= p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\right).
\end{aligned}
$$

Posterior distribution

The posterior distribution is also a normal distribution:

$$p(\theta|y_1, \ldots, y_n) = p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2),$$

where

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \text{and } \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Incidentally, the same result is obtained by adding information for the data $y_1, \ldots, y_n$ one point at a time, using the posterior distribution at each step as the prior distribution for the next

Normal distribution with known mean but unknown variance

- For $p(y|\theta, \sigma^2) = N(y|\theta, \sigma^2)$, with θ known and σ^2 unknown, the likelihood for a vector y of n i.i.d observations is

$$p(y|\sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right)$$

$$= (\sigma^2)^{-n/2} \exp(-\frac{n}{2\sigma^2} v)$$

The sufficient statistics is

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

Prior

- The conjugate prior density is a scaled inverse-χ^2 distribution with scale σ_0^2 and degrees of freedom ν_0.

$$p(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

- Posterior density

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-((n+\nu_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + nv)\right). \end{aligned}$$

- Thus, $\sigma^2|y \sim Inv-\chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2+nv}{\nu_0+n})$

Poisson distribution

- Observations: $y = (y_1, y_2, \dots, y_n)$
- Likelihood:

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{n\bar{y}} e^{-n\theta}$$

- Prior density: Gamma(α, β)

$$p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$$

- Posterior density:

$$p(\theta|y) \propto e^{-(n+\beta)\theta} \theta^{n\bar{y}+\alpha-1}$$

$$\theta|y \sim \text{Gamma}(n\bar{y} + \alpha, \text{n} + \beta)$$

Exponential Distribution

- Observations: $y = (y_1, y_2, \dots, y_n)$
- Likelihood:

$$p(y|\theta) = \theta^n \exp(-n\bar{y}\theta)$$

- Prior density: Gamma(α, β)

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$$

- Posterior density:

$$p(\theta|y) \propto \theta^{\alpha+n-1} \exp(-(n\bar{y} + \beta)\theta)$$

$$\theta|y \sim \text{Gamma}(\alpha + n, n\bar{y} + \beta)$$

The sampling distribution when viewed as the likelihood of θ, for fixed y, is proportional to a Gamma(n+1, ny) density. Thus the Gamma(α, β) prior distribution for θ can be viewed as α−1 exponential observations with total waiting time β

Jeffreys' Priors

- Jeffreys' principle leads to defining the noninformative prior density

$$p(\theta) = [J(\theta)]^{1/2}$$

where $J(\theta)$ is the *Fisher information* for θ

$$J(\theta) = E\left[\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\frac{d^2\log p(y|\theta)}{d\theta^2}|\theta\right]$$

Univariate Normal with a Noninformative Prior

- Consider a vector y of n independent observations from a univariate normal distribution, $N(\mu, \sigma^2)$
- Assuming prior independence of location and scale parameters, is uniform on (μ, log σ) or,

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

- The joint posterior distribution

$$p(\mu, \sigma^2 | y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

The marginal posterior distribution

- The marginal posterior distribution, $p(\sigma^2|y)$

$$p(\sigma^2|y) \propto \int \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu$$

$$\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{\frac{2\pi\sigma^2}{n}}$$

$$\propto (\sigma^2)^{-\frac{n+1}{2}} \exp(-\frac{(n-1)s^2}{2\sigma^2})$$

which is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

Joint Posterior Density

- $p(\sigma^2|y)$ is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \mathrm{Inv}-\chi^2(n-1, s^2)$$

Therefore,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that this result agrees with the standard frequentist result on the sample variance.

- As we know before,

$$\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$$

The joint posterior density,

$$p(\mu, \sigma^2|y) \propto p(\mu|\sigma^2, y)p(\sigma^2|y)$$

Sampling from the joint posterior distribution

- Now that we have $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, inference on μ isn't difficult.
- One method is to use the Monte Carlo approach discussed earlier
    1. Sample σ_i^2 from $p(\sigma^2|y)$
    2. Sample μ_i from $p\left(\mu|\sigma_i^2, y\right)$

Then $\mu_1, \ldots, \mu_m$ is a sample from $p(\mu|y)$.

- Note that in this case, it is actually possible to derive the exact density of $p(\mu|y)$.

Marginal Posterior Distribution for μ

- The marginal posterior distribution

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

Therefore,

$$\frac{\mu - \bar{y}}{s/\sqrt{n}}|y \sim t_{n-1}$$

which corresponds to the standard result used for inference on a population mean

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}|\mu, \sigma^2 \sim t_{n-1}$$

- The sampling distribution of the pivotal quantity $(\bar{y} - \mu)/(s/\sqrt{n})$ does not depend on the nuisance parameter σ^2, and its posterior distribution does not depend on data.

Conjugate Prior

- This has been labelled as $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2)$ distribution
- its four parameters can be identified as the location and scale of μ and the degrees of freedom and scale of σ^2
- One important thing to note is that with this prior, μ and σ^2 are dependent (i.e. $p(\mu|\sigma^2)$ is a function of σ^2, for example, if σ^2 is large, then a high-variance prior distribution is induced on μ
- This has a different feel from the standard frequentist analysis where $\bar{y}$ and s^2 are independent.

The Posterior Density

- The posterior density satisfies

$$
\begin{aligned}
p(\mu, \sigma^2 | y) &\propto \frac{1}{\sigma} \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2\right]\right) \\
&\times \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right) \\
&\propto \frac{1}{\sigma} \frac{1}{(\sigma^2)^{\frac{\nu_n}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2 + \kappa_n(\mu - \mu_n)^2\right]\right)
\end{aligned}
$$

The posterior distribution is $\mathrm{N-Inv-}\chi^2(\mu_n, \frac{\sigma_n^2}{\kappa_n}; \nu_n, \sigma_n^2)$

The Posterior Density

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2$$

The parameters of the posterior distribution combine the prior information and the information contained in the data. For example μ_n is a weighted average of the prior mean and the sample mean, with weights determined by the relative precision of the two pieces of information.

The Conditional Posterior Distribution $p(\mu|\sigma^2, y)$

- By using that $p(\mu|\sigma^2, y) \propto p(\mu, \sigma^2|y)$ with σ as a constant, we get

$$\mu|\sigma^2, y \sim \ N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

Note that the mean and variance can be written as

$$\mu_n = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} \qquad \sigma_n^2 = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$$

which matches with the fixed variance case discuss earlier.

The Marginal Posterior Distribution $p(\sigma^2|y)$

- $p(\sigma^2|y)$

$$\sigma^2|y \sim \mathrm{Inv}-\chi^2(\nu_n, \sigma_n^2)$$

This can be seen by the same way $p(\sigma^2|y)$ was shown in the non-informative prior case or by recognizing the $\mathrm{N}-\mathrm{Inv}-\chi^2$ form of the joint density.

- $p(\mu|y)$

As mentioned before, this can be determined by simulation (see in the next slide). In this case an exact answer can be determined by integrating out σ^2 from the joint density (as in the non-informative case), we get

$$\mu|y \sim t_{\nu_n}(\mu_n, \frac{\sigma_n^2}{\kappa_n})$$

Simulation of $p(\mu|y)$

- we first draw σ^2 from its marginal posterior distribution $p(\sigma^2|y)$,

$$\sigma^2|y \sim \mathrm{Inv}-\chi^2(\nu_n, \sigma_n^2)$$

- then draw μ from its normal conditional posterior distribution $p(\mu|\sigma^2, y)$

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

using the simulated value of σ^2.

The Prior and Posterior Distribution

- The conjugate prior distribution

Dirichlet: a multivariate generalization of the beta distribution

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

where $\theta_j \in (0,1)$ and $\sum \theta_j = 1$

- The posterior distribution

The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

Thanks