Single-parameter models

March 12, 2023

Overview

Normal distribution with known variance

- Normal distribution could fit lots of data, or as approximation with CLT.
- For complicated data, mixture normal distribution could be used.
- Normal distribution with known variance means in this part we only discuss μ as our parameter θ, while σ^2 has been treated as known.

Normal distribution with known variance

- Likelihood of one data point

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

- Conjugate prior and posterior distributions

$$p(\theta) = e^{A\theta^2 + B\theta + C}.$$

Normal distribution with known variance

- The prior distribution could be parameterized as

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

  which means the prior distribution of θ is $\theta \sim N(\mu_0, \tau_0^2)$ with known hyperparameters μ_0, τ_0
- Please deduct the posterior distribution with the one data point likelihood.

Normal distribution with known variance

- The posterior distribution could be represented as

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right).$$

- By collecting terms, we can have

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

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

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

Normal distribution with known variance

- Precisions of the prior and posterior distributions: in manipulating normal distributions, the inverse of the variance plays a prominent role and is called the precision.
- For normal data and normal prior distribution, the posterior precision equals the prior precision plus the data precision.
- The posterior mean is expressed as a weighted average of the prior mean and the observed value, y, with weights proportional to the precisions.

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

Normal distribution with known variance

- Alternatively, we can express μ_1 as the prior mean adjusted toward the observed y

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2},$$

- or as the data 'shrunk' toward the prior mean,

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}.$$

Posterior predictive distribution with known variance

- The prediction could be calculated directly by integration

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\
&\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)d\theta.
\end{aligned}
$$

- With the posterior distribution, we can see the expectation of prediction is

$$
\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\mathrm{E}(\tilde{y}|\theta, y)|y) = \mathrm{E}(\theta|y) = \mu_1,
$$

- The variance is

$$
\begin{aligned}
\mathrm{var}(\tilde{y}|y) &= \mathrm{E}(\mathrm{var}(\tilde{y}|\theta, y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\theta, y)|y) \\
&= \mathrm{E}(\sigma^2|y) + \mathrm{var}(\theta|y) \\
&= \sigma^2 + \tau_1^2.
\end{aligned}
$$

Normal model with multiple observations

- For more than one observation, $y = (y_1, \ldots, y_n)$, the likelihood function is the product of one data point

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

- Please deduct the posterior mean and variance.

Normal model with multiple observations

- For the normal distribution with known variance, the posterior distribution depends on y only through the sample mean, $\bar{y}$ is a sufficient statistic in this model.
- The posterior distribution is

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

- where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

- As the sample size $n \to \infty$ with τ_0 fixed, or as $\tau_0 \to \infty$ with fixed n, we have

$$p(\theta|y) \approx \mathrm{N}(\theta|\overline{y}, \sigma^2/n),$$

Normal distribution with known mean but unknown variance

- The square of a normal distribution? What is in your mind?
- The easiest part is the likelihood, in which we just need to recognize the variance as our parameter.

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

Normal distribution with known mean but unknown variance

- The corresponding conjugate prior density is the inverse-gamma,

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2},$$

which has hyperparameters (α, β). A convenient parameterization is as a scaled inverse-χ^2 distribution with scale σ_0^2 and v_0 degrees of freedom. (the textbook has a small mistake)

Inverse-chi-square	$\theta \sim \text{Inv-}\chi^2_\nu$ $p(\theta) = \text{Inv-}\chi^2_\nu(\theta)$	degrees of freedom $\nu > 0$

$$p(\theta) = \frac{2^{-v/2}}{\Gamma(v/2)} \theta^{v/2+1} e^{-\theta/2}, \theta > 0$$

$$\text{same as Gamma } (\alpha = \frac{v}{2}, \beta = \frac{1}{2})$$

Normal distribution with known mean but unknown variance

- The prior distribution of σ^2 is taken to be the distribution of $\sigma_0^2 v_0/X$, where X is a $\chi^2_{v_0}$ random variable.
- We use the convenient but nonstandard notation, $\sigma^2 \sim Inv_{\chi^2}(v_0, \sigma_0^2)$
- The resulting posterior density for σ^2 is (we replace θ by $\sigma_0^2 v_0/\sigma^2$)

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

Normal distribution with known mean but unknown variance

- The posterior distribution of σ^2 is

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

  which is a scaled inverse-χ^2 distribution with scale equal to degree-of-freedom-weighted average of the prior and data scales and degrees of freedom equal to the sum of the prior and data degrees of freedom.
- The prior distribution can be thought of as providing the information equivalent to v_0 observations with average squared deviation σ_0^2.

Poisson model

- The likelihood of n independent and identically distributed observations $y = (y_1, \ldots, y_n)$ is

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &\propto \theta^{t(y)} e^{-n\theta}, \end{aligned}$$

  where $t(y) = \sum_{i=1}^{n} y_i$ is the sufficient statistic.
- Please rewrite the likelihood function with exponential family form.

Poisson model

- The exponential family form of the likelihood is

$$p(y|\theta) \propto e^{-n\theta} e^{t(y)\log\theta},$$

- The conjugate prior distribution is

$$p(\theta) \propto (e^{-\theta})^{\eta} e^{\nu \log\theta},$$

  indexed by hyperparameters (η, v)
- We know it could be written as the exponential family form, but is this convenient?

Poisson model

- Or we just write the prior distribution based on the likelihood form: $p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$, which is the form of a Gamma distribution with parameters α and β, Gamma(α, β)
- With the conjugate prior distribution, the posterior distribution is

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

- With the updated posterior Gamma distribution, the posterior expectation of θ is $\frac{\alpha + n\bar{y}}{\beta + n}$

The negative binomial distribution

- With conjugate families, the known form of the prior and posterior densities can be used to find the marginal distribution, $p(y)$, using the formula
$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$
- the marginal distribution of a single observation, y, has prior predictive distribution

$$\begin{aligned} p(y) &= \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)} \\ &= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}}, \end{aligned}$$

The negative binomial distribution

- The marginal distribution of y is known as the negative binomial density:

$$y \sim \text{Neg-bin}(\alpha, \beta)$$

$$p(y) = \binom{\alpha + y - 1}{y} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \left(\frac{1}{\beta + 1}\right)^{y},$$

- the negative binomial distribution is a mixture of Poisson distributions with rates, θ, that follow the gamma distribution:

$$\text{Neg-bin}(y|\alpha, \beta) = \int \text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha, \beta)d\theta.$$

Poisson model parameterized in terms of rate and exposure

- Suppose we have explanatory variable x and response variable y to form the conditional Poisson distribution

$$y_i \sim \mathrm{Poisson}(x_i\theta),$$

- Please write the likelihood function.

Poisson model parameterized in terms of rate and exposure

- Suppose we have explanatory variable x and response variable y to form the conditional Poisson distribution

$$y_i \sim \text{Poisson}(x_i\theta),$$

- The likelihood function is

$$p(y|\theta) \propto \theta^{\left(\sum_{i=1}^{n} y_i\right)} e^{-\left(\sum_{i=1}^{n} x_i\right)\theta}$$

- By choosing the gamma distribution, Gamma(α, β), as prior distribution, please write the posterior distribution.

Poisson model parameterized in terms of rate and exposure

- Suppose we have explanatory variable x and response variable y to form the conditional Poisson distribution

$$y_i \sim \mathrm{Poisson}(x_i\theta),$$

- The likelihood function is

$$p(y|\theta) \propto \theta^{\left(\sum_{i=1}^n y_i\right)} e^{-\left(\sum_{i=1}^n x_i\right)\theta}$$

- By choosing the gamma distribution, Gamma(α, β), as prior distribution, the posterior distribution should be

$$\theta|y \sim \mathrm{Gamma}\left(\alpha + \sum_{i=1}^n y_i,\ \beta + \sum_{i=1}^n x_i\right).$$

Estimating a rate from Poisson data: an idealized example

- Observation: A city in the US, for a single year, 3 persons out of 200,000 died of asthma (mortality rate: 1.5 per 100,000).
- A Poisson mode for this case: the number of deaths in a city of 200,000 in one year y may follow *Poisson*(2.0θ), where θ represents the true underlying long-term asthma mortality rate in the city per 100,000 persons per year and 2.0 is the exposure x (with unit 100,000 persons).
- We can use knowledge about asthma death rates around the world to construct a prior distribution for θ and then combine the observation $y = 3$ to obtain a posterior distribution.

Example: set up prior and posterior distribution

- Prior: the typical asthma mortality rate is around 0.6 per 100,000 in Western countries. By choosing the conjugate prior with mean 0.6, a $Gamma(3.0, 5.0)$ is used with 97.5% of the mass of the density lies below 1.44.
- Posterior: the posterior distribution is $\Gamma(\alpha + y, \beta + x)$. In this case it is $Gamma(6.0, 7.0)$, which has mean 0.86
- Posterior with additional data: Suppose 10 years of data are obtained and the mortality rate is 1.5 per 100,000 ($y = 30$). Assuming the population is constant and the outcomes in the ten years are independent with constant θ. The posterior distribution is $Gamma(33.0, 25.0)$, the posterior mean of θ is 1.32, and the posterior probability that θ exceeds 1.0 is 0.93.

Example: set up prior and posterior distribution graphs

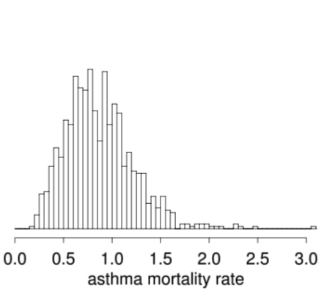

Figure 2.5 *Posterior density for* θ*, the asthma mortality rate in cases per* 100,000 *persons per year, with a* Gamma$(3.0, 5.0)$ *prior distribution: (a) given* $y = 3$ *deaths out of* 200,000 *persons; (b) given* $y = 30$ *deaths in* 10 *years for a constant population of* 200,000*. The histograms appear jagged because they are constructed from only 1000 random draws from the posterior distribution in each case.*