Data 582 - Bayesian Inference

Normal-Normal Model

Contents

1 Introduction **1**
1.1 Completing the Square 1

2 Known variance **2**
2.1 The Likelihood 2
2.2 The Prior 2
2.3 The Posterior 3
2.4 Simplifying the likelihood 4
2.5 Combining Information 6

1 Introduction

In lecture we saw how we could perform Bayesian Inference on the parameters of a Normal distribution. Inference for this two-parameter model can be broken down into two one-parameter problems. Our first considerations involved the one-parameter model (inference on μ assuming σ^2 known, then inference on σ^2 assuming μ known).

We denote a normal variable Y with mean μ and variance σ^2 (and thus standard deviation σ) by $Y \sim \mathcal{N}(\mu, \sigma^2)$ having pdf given by:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

1.1 Completing the Square

We will be "completing the square" at multiple stages in this document. If you do not recall the details for this method, you can review at Khan Academy or mathisfun.com, or at various other places on the internet. Recall to complete the square we take the following steps:

$$ax^2 + bx + c$$
$$a\left(x^2 + \frac{b}{a}x\right) + c \qquad \text{Factor out a}$$

Add and subtract the second term halved and squared

$$a\left(x^2+\frac{b}{a}x+\left(\frac{b}{2a}\right)^2-\left(\frac{b}{2a}\right)^2\right)+c$$

Pull the term you subtracted out of the parenthesis

$$=a\left(x^2+\frac{b}{a}x+\left(\frac{b}{2a}\right)^2\right)+c-a\left(\frac{b}{2a}\right)^2$$

Convert the terms in the parentheses into a perfect square

$$=a\left(x+\frac{b}{2a}\right)^2+c-a\left(\frac{b}{2a}\right)^2$$

Simplify the constant term

$$=a\left(x+\frac{b}{2a}\right)^2+c-\frac{b^2}{4a}$$

2 Known variance

For now, we will consider inference on μ assuming σ^2 is known. We will denote the population mean by θ rather than μ to emphasize this is the parameter of interest and use "$\mid \sigma^2$" to denote that σ^2 is "given", or known, in the equations below. We need to identify the likelihood and define a prior distribution so we can calculate

$$p(\theta \mid \sigma^2, y) \propto p(y \mid \theta, \sigma^2) \times p(\theta \mid \sigma^2)$$

2.1 The Likelihood

The likelihood for normal samples $y = (y_1, \ldots, y_n)$ can be written,

$$\begin{aligned}\mathcal{L}(\theta) = p(y \mid \theta, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \theta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\end{aligned} \tag{1}$$

2.2 The Prior

Let the prior distribution be given[1] as

$$\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Removing any constants that do not depend on θ we get:

$$\begin{aligned} p(\theta \mid \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right) \end{aligned}$$

2.3 The Posterior

For $p(\theta \mid \sigma^2)$ to be a conjugate prior, the posterior needs to have the form $\exp(c_1(\theta - c_2)^2)$. The following was edited from Jesse Mu's notes available on his GitHub repo.

$$\begin{aligned} p(\theta \mid \sigma^2, y) &\propto p(y \mid \theta, \sigma^2) \times p(\theta \mid \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \times \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right\} \end{aligned}$$

expand the polynomials we get ...

$$\begin{aligned} &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\theta^2\right)\right\} \times \exp\left\{-\frac{1}{2\sigma_0^2}\left(\theta^2 - 2\theta\mu_0 + \mu_0^2\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\sum y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right) + \frac{1}{\sigma_0^2}\left(\theta^2 - 2\theta\mu_0 + \mu_0^2\right)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\sum y_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + \left(\frac{\mu_0^2}{\sigma_0^2}+\frac{1}{\sigma^2}\sum y_i^2\right)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\sum y_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right]\right\} \times \exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2}+\frac{1}{\sigma^2}\sum y_i^2\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum y_i}{\sigma^2}\right)\right]\right\} \\ &:= \exp\left\{-\frac{1}{2}\left[\theta^2(a) - 2\theta\,(b)\right]\right\} \end{aligned}$$

[1] in lecture we used m and s^2 for hyperparmeter notation instead of μ_0 and σ_0^2, respectively.

The last line removes the constant of proportionality $c = \exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2} + \frac{1}{\sigma^2}\sum y_i^2\right)\right\}$. To simplify this, let $a = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$ and $b = \frac{\mu_0}{\sigma_0^2} + \frac{\sum y_i}{\sigma^2}$. By completing the square and removing any constants of proportionality we get:

$$p(\theta \mid \sigma^2, y_1, \ldots, y_n) \propto \exp\left[-\frac{1}{2}(a\theta^2 - 2b\theta)\right]$$

Complete the square...

$$\begin{aligned}
&\propto \exp\left[-\frac{1}{2}\left\{a\left(\theta^2 - \frac{2b}{a}\theta + \left(\frac{2b}{2a}\right)^2\right) - a\left(\frac{2b}{2a}\right)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{a\left(\theta - \frac{b}{a}\right)^2 - a\left(\frac{2b}{2a}\right)^2\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{a\left(\theta - \frac{b}{a}\right)^2\right\}\right] && \text{Remove constants} \\
&\propto \exp\left[-\frac{(\theta - b/a)^2}{2(1/a)}\right] && \text{Express as a normal kernel}
\end{aligned}$$

We recognize the last line as having the functional form of a normal distribution having mean equal to b/a and variance equal to $1/a$. Denoting these posterior parameters by μ_n and σ_n^2, we write $\theta \mid \sigma^2, y_1, \ldots, y_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ where,

$$\mu_n = \frac{b}{a} = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \qquad \sigma_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}. \tag{2}$$

We can summarize the above as follows:

$$\begin{aligned}
Y_i &\sim \mathcal{N}(\mu, \sigma^2) \ (\sigma^2 \text{ known}) \\
\theta \mid \sigma^2 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
\theta \mid \sigma^2, y_1, \ldots, y_n &\sim \mathcal{N}(\mu_n, \sigma_n^2)
\end{aligned}$$

2.4 Simplifying the likelihood

Expanding the polynomial and completing the square, we can express the likelihood as follows:

$$
\begin{aligned}
\mathcal{L}(\theta) = \propto \exp & \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right] \\
\propto \exp & \left[-\frac{1}{2\sigma^2} \left(\sum y_i^2 - 2\theta \sum y_i + n\theta^2 \right) \right] && \text{expand the polynomial} \\
\propto \exp & \left[-\frac{n}{2\sigma^2} \left(\frac{\sum y_i^2}{n} - 2\theta \frac{\sum y_i}{n} + \theta^2 \right) \right] && \text{factor out the n} \\
\propto \exp & \left[-\frac{n}{2\sigma^2} \left(\frac{\sum y_i^2}{n} - 2\theta\overline{y} + \theta^2 \right) \right] && \text{sub in } \overline{y} = \sum \frac{y_i}{n} \\
\propto \exp & \left[-\frac{n}{2\sigma^2} \left(\frac{\sum y_i^2}{n} - 2\theta\overline{y} + \theta^2 + \overline{y}^2 - \overline{y}^2 \right) \right] && \text{add 0} \\
\propto \exp & \left[-\frac{n}{2\sigma^2} \left(\frac{\sum y_i^2}{n} + (\theta - \overline{y})^2 - \overline{y}^2 \right) \right] && \text{complete the square} \\
\propto \exp & \left[-\frac{n}{2\sigma^2} (\theta - \overline{y})^2 \right] && \text{remove constants}
\end{aligned}
$$

The last line tells us that **the likelihood of the normal random sample $y_1, \ldots y_n$ is proportional to the likelihood of the sample mean of $\overline{y}$ and variance σ^2/n**, i.e.

$$y_1, \ldots, y_n \mid \theta \sim \mathcal{N}(\overline{y}, \sigma^2/n) \tag{3}$$

We can think of this as drawing a single value, $\overline{y}$, the sample mean, from the normal distribution with mean θ and variance σ^2/n.

$$\mu_{post} = \frac{a\mu_{prior} + b\overline{y}}{a+b} \qquad \sigma^2_{post} = \frac{1}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2/n}} = \frac{1}{a+b} \tag{4}$$

where $a = 1/\sigma^2_{prior}$ and $b = \frac{1}{\sigma^2/n} = n/\sigma^2$. More generally, if we only have a single observation x from a $\mathcal{N}(\theta, \sigma^2)$ with a $N(\mu_{post}, \sigma^2_{post})$ prior distribution on θ then the posterior pdf is $N(\mu_{post}, \sigma^2_{post})$ where $a = 1/\sigma^2_{prior}$, $b = 1/\sigma^2$ and

$$\mu_{post} = \frac{a\mu_{prior} + bx}{a+b} \qquad \sigma^2_{post} = \frac{1}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2}} = \frac{1}{a+b} \tag{5}$$

Notice that the likelihood only depends on $\overline{y}$ rather than the individual observations $y_1, y_2, \ldots, y_n$. Informally, a sufficient statistic contains as much information about θ as the entire sample does. To put another way, the individual observations $y_1, y_2, \ldots, y_n$ do not provide any additional information about θ than known only $\overline{y}$. $\overline{y} = \sum_{i=1}^{n} y_i$ is said to be a *sufficient statistic* in this model. This can be understoon mathematically as follows:

Day 1 Observe data $x_1, \ldots, x_n$ and update inference on an unknown parameter θ.

- Define prior $p(\theta)$ and calculate the posterior

Day 2 Observe data $x_{n+1}, \ldots, x_{n+m}$ and update inference on an unknown parameter θ.

- Yesterday's posterior becomes today's prior from which we calculate our posterior.

More generally, our posterior on day i (AKA our prior on day $i+1$) contains all information we need going forward. To update our beliefs when confronted with new information x_{n+1}, we can ignore past observations $x_1, \ldots, x_n$ so long as the went into to updating our current prior beliefs. The posterior defined on **Day 2** can be computed as:

$$\begin{aligned}
\text{posterior}_{day2} &\propto \text{prior}_{day2} \times \text{likelihood}_{day2} \\
\text{posterior}_{day2} &\propto \text{posterior}_{day1} \times \mathcal{L}_{day2} \\
p(\theta \mid x_{n+1}, \ldots, x_{n+m}) &\propto p(\theta \mid x_1, \ldots, x_n) p(x_{n+1}, \ldots, x_{n+m} \mid \theta) \\
\text{posterior}_{day2} &\propto \text{prior}_{day1} \mathcal{L}_{day1} \times \mathcal{L}_{day2} \\
p(\theta \mid x_{n+1}, \ldots, x_{n+m}) &\propto p(\theta) p(x_1, \ldots, x_n \mid \theta) p(x_{n+1}, \ldots, x_{n+m} \mid \theta) \\
p(\theta \mid x_{n+1}, \ldots, x_{n+m}) &\propto p(\theta) \prod_{i=1}^{n} f(x_i \mid \theta) \prod_{i=n+1}^{n+m} f(x_i \mid \theta)
\end{aligned}$$

which can be expressed as:

$$p(\theta \mid x_1, \ldots x_{n+m}) \propto p(\theta) p(x_1, \ldots, x_{n+m} \mid \theta)$$

In other words, if we perform sequential updating every time we get new data, we will get the same answer (that is, we will get the same posterior distribution) as if we did a single batch update had we obtained all the data at once. This result will be explored in the example in Section **??**.

2.5 Combining Information

Recall that the precision is defined as the inverse of the variance. The more spread out a distribution is (larger variance), the less precise it is (smaller precision). Notice that the posterior mean is a weighted average of the prior mean and the sample mean; the weights are proportional to the respective precisions $a = 1/\sigma^2_{prior}$ and $b = \frac{1}{\sigma^2/n} = n/\sigma^2$:

$$\mu_{post} = \frac{\frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{n}{\sigma^2}} = \frac{a\mu_{prior} + b\overline{y}}{a+b} = \frac{n}{n + \frac{\sigma^2}{\sigma^2_{\text{prior}}}}\,\overline{y} + \frac{\frac{\sigma^2}{\sigma^2_{\text{prior}}}}{n + \frac{\sigma^2}{\sigma^2_{\text{prior}}}}\mu_{\text{prior}} \tag{6}$$

We discussed in class how equivalent prior sample size, or *effective sample size* is $\sigma^2/\sigma^2_{\text{prior}}$ (to see this, just compare the numerators from the right hand side of the weighted posterior mean equation above). Thus, the prior contains the same amount of information as $\sigma^2/\sigma^2_{\text{prior}}$ observations.

Notice that when the prior precision $1/\sigma^2_{\text{prior}}$ is is small relative to the data precision, n/σ^2, then the posterior distribution can be approximated by the limiting case when $\sigma^2_{\text{prior}} \to \infty$:

$$\begin{aligned} p(\mu \mid x) &\sim \mathcal{N}\left(\frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_{\text{prior}}}{\cancel{\sigma^2_{\text{prior}}}}}{\frac{n}{\sigma^2} + \frac{1}{\cancel{\sigma^2_{\text{prior}}}}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\cancel{\sigma^2_{\text{prior}}}}}\right) \\ &\sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right) \end{aligned} \tag{7}$$

Side note: The above result above can also be obtained from assuming the improper prior $p(\theta) \propto$ a constant for $\theta \in (-\infty, \infty)$; see Bolstad & Curran (2016) Chapter 11 page 219. Again we see that the posterior is a balance act between the prior and the likelihood. These two forces will be affected by the following:

1. Lots of data will have a big influence on the posterior.
2. High certainty (low variance) in the prior has a big influence on the posterior.