Week 3 Optional Supplementary Materials

Lizzy Huang
Duke University

Week 3

In this week, we mainly focus on the situation when the data follow the Normal distribution. We have seen in Week 2 that, if the variance σ^2 of the data is known, we can use the Normal-Normal conjugate family. When the variance σ^2 of the data is also unknown, we need to set up a joint prior distribution $p(\mu, \sigma^2)$ for both the mean μ and the variance σ^2. This leads to the Normal-Gamma conjugate family, its limiting case, the reference prior, and other mixtures of priors such as the Jeffrey-Zellner-Siow prior.

After we have introduced these conjugate families and priors, we can apply them to do hypothesis testing. In this week, we have discussed the Bayes factor, a ratio between likelihoods for comparing two competing hypotheses, provided the formulas when we make inference for means, compare two paired means, and compare two independent means. We have emphasized that Bayes factor is sensitive to prior choice, by showing the paradoxes when the Bayesian approach and the frequentist approach do not agree. All the Gamma distribution we use since Week 3 follows the $\mathsf{Gamma}(\alpha, \beta)$ definition.

In this file, we have chosen several concepts and provided the mathematic deviations of these concepts. These deviations are out of the scope of the course and only for advanced learners who are comfortable with integral calculation.

Normal-Gamma Conjugate Family

The Normal-Gamma conjugate family is used when the data is Normal, and when the data variance σ^2 is unknown. Therefore, we need to find a nice joint prior distribution $p(\mu, \sigma^2)$ of both μ and σ^2, which will provide conjugacy with the Normal likelihood from the data. To get the joint prior distribution, we start with a hierarchical model, that is, we first decide the prior distribution of μ conditioning on σ^2, $p(\mu \mid \sigma^2)$, pretending that we already have the information of the variance. Then we look for a nice prior $p(\sigma^2)$ for σ^2, so that finally

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2)$$

will provide conjugacy with the Normal distribution.

For convenience, we instead use the precision ϕ, which is the inverse of the variance, $\phi = \dfrac{1}{\sigma^2}$ as one of the hyperparameters[1].

It turns out that, when μ is Normal, conditioning on σ^2 (or ϕ), with ϕ to be Gamma (i.e., σ^2 is inverse Gamma), the joint prior distribution is conjugate with the Normal distribution.

$$\mu \mid \sigma^2 \ \sim \ \mathsf{N}(m_0, \frac{\sigma^2}{n_0}) = \mathsf{N}(m_0, n_0\phi),$$

$$\phi \ \sim \ \mathsf{Gamma}(\frac{v_0}{2}, \frac{s_0^2 v_0}{2}).$$

Here, in order to provide flexibility for the variance, we use n_0 as one of the hyperparameters to scale the variance of μ.

[1] Recall that the prior sample size is proportional to $1/\sigma^2$, which is the precision ϕ.

This is,

$$p(\mu \mid \phi) = p(\mu \mid \sigma^2) = \frac{1}{\sigma\sqrt{2\pi/n_0}} \exp\left(-\frac{1}{2}\frac{(\mu - m_0)^2}{\sigma^2/n_0}\right) = \sqrt{\frac{n_0}{2\pi}}\phi^{1/2} \exp\left(-\frac{n_0\phi}{2}(\mu - m_0)^2\right),$$

$$p(\phi) = \frac{(s_0^2 v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\phi^{\frac{v_0}{2}-1} \exp\left(-\frac{s_0^2 v_0}{2}\phi\right).$$

Multiply the two together, we get the joint prior distribution

$$p(\mu, \phi) = \sqrt{\frac{n_0}{2\pi}}\frac{(s_0^2 v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\phi^{(v_0-1)/2} \exp\left(-\frac{\phi}{2}\left[n_0(\mu - m_0)^2 + s_0^2 v_0\right]\right).$$

This is called the Normal-Gamma distribution. It has 4 hyperparameters, m_0 (location), n_0 (scale), v_0, and s_0^2. They correspond to the prior sample mean, prior sample size, prior sample degrees of freedom, and prior sample variance.

$$p(\mu, \phi) = \mathsf{NormalGamma}(m_0, n_0, s_0^2, v_0).$$

We usually ignore the complicated constants from the Gamma function and other multiplications, and focus more on the form of this distribution,

$$p(\mu, \phi) \propto \phi^{(v_0-1)/2} \exp\left(-\frac{\phi}{2}\left[n_0(\mu - m_0)^2 + s_0^2 v_0\right]\right).$$

When update the distribution with one data point $y_i \sim \mathsf{N}(\mu, \sigma^2 = \frac{1}{\phi})$, we get the posterior distribution by the Bayes' Rule

$$p(\mu, \phi \mid y_i) \propto \left\{\sqrt{\frac{\phi}{2\pi}} \exp\left(-\frac{\phi}{2}(y_i - \mu)^2\right)\right\} \times \left\{\phi^{(v_0-1)/2} \exp\left(-\frac{\phi}{2}\left[n_0(\mu - m_0)^2 + s_0^2 v_0\right]\right)\right\}$$
$$\propto \phi^{\frac{(v_0+1)-1}{2}} \exp\left(-\frac{\phi}{2}\left[(n_0+1)\left(\mu - \frac{n_0 m_0 + y_i}{n_0+1}\right)^2 + s_0^2 v_0 + n_0(y_i - m_0)^2\right]\right)$$

Comparing this to the format of the Normal-Gamma family, we get

$$m_1 = \frac{n_0 m_0 + y_i}{n_0 + 1}, \qquad n_1 = n_0 + 1, \qquad v_1 = v_0 + 1, \qquad s_1^2 v_1 = s_0^2 v_0 + n_0(y_i - m_0)^2.$$

When we have n data point with sample mean $\bar{y}$ and sample variance s^2, the hyperparameters will get updated to be

$$m_n = \frac{n_0 m_0 + n\bar{y}}{n_0 + n}, \qquad n_n = n_0 + n, \qquad v_n = v_0 + n, \qquad s_n^2 v_n = s_0^2 v_0 + (n-1)s^2 + \frac{n_0 n}{n_n}(\bar{y} - m_0)^2.$$

So the joint posterior distribution is

$$p(\mu, \phi \mid y_1, \ldots, y_n) \propto \phi^{\frac{v_n-1}{2}} \exp\left(-\frac{\phi}{2}\left[n_n(\mu - m_n)^2 + s_n^2 v_n\right]\right). \tag{1}$$

Marginal Posterior Distribution of μ

Once we have the joint posterior distribution given by (1), we can integrate ϕ out to get the marginal posterior distribution of μ.

$$p(\mu \mid y_1 \ldots, y_n) \propto \int_0^\infty \phi^{\frac{v_n-1}{2}} \exp\left(-\frac{\phi}{2}\left[n_n(\mu - m_n)^2 + s_n^2 v_n\right]\right) d\phi.$$

The quantity $n_n(\mu - m_n)^2 + s_n^2 v_n$ is a constant with respect to the integral variable ϕ, so we denote this quantity as

$$A = n_n(\mu - m_n)^2 + s_n^2 v_n,$$

for the purpose of simplicity. Then the integral has a form

$$\int_0^\infty \phi^{(v_n-1)/2} \exp\left(-\frac{A}{2}\phi\right) d\phi,$$

which is like the definition of the Gamma function,

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)\, dx.$$

By a change of variable $x = \frac{A}{2}\phi$, we get

$$p(\mu \mid y_1, \ldots, y_n) \propto (A)^{-(v_n+1)/2} \Gamma(\frac{v_n+1}{2}) \propto A^{-(v_n+1)/2}.$$

Recall that, $A = n_n(\mu - m_n)^2 + s_n^2 v_n$, so

$$A^{-\frac{v_n+1}{2}} = \left(s_n^2 v_n + n_n(\mu - m_n)^2\right)^{-\frac{v_n+1}{2}} \propto \left(1 + \frac{\left(\frac{\mu - m_n}{s_n/\sqrt{n_n}}\right)^2}{v_n}\right)^{-\frac{v_n+1}{2}},$$

which shares the same form as the Student's t-distribution

$$p(t) \propto \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

with $t = \dfrac{\mu - m_n}{s_n/\sqrt{n_n}}$. That is to say, the marginal posterior distribution of μ is a non-standardized Student's t-distribution, with location m_n, scale $s_n/\sqrt{n_n}$, and degrees of freedom v_n

$$\mu \mid \text{data} \;\sim\; \mathsf{t}(v_n, m_n, \frac{s_n^2}{n_n}).$$

While we have a nice closed form for the marginal posterior distribution for μ, we do not have such luck for the marginal posterior distribution for the variance σ^2. We will need simulation to get the distribution of σ^2.

Mixtures of Conjugate Priors

Recall that the prior sample size in the Normal-Normal conjugate family is proportional to n_0, if $\mu \sim \mathsf{N}(m_0, \sigma^2/n_0)$. If we are uncertain about what prior sample size we should choose to match our prior belief, we might "insert one more layer" into the hierachical model and impose a Gamma distribution as the prior of n_0. Here we have

$$\begin{aligned} \mu \mid \sigma^2, n_0 &\sim \mathsf{N}(m_0, \frac{\sigma^2}{n_0}), \\ n_0 \mid \sigma^2 &\sim \mathsf{Gamma}(\frac{1}{2}, \frac{r^2}{2}). \end{aligned}$$

We can obtain the prior distribution of μ conditioning on σ^2 by integrating the product of the above two distributions over n_0:

$$\begin{aligned} p(\mu \mid \sigma^2) &= \int_0^\infty \left[\frac{\sqrt{n_0}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{n_0}{2\sigma^2}(\mu - m_0)^2\right)\right] \times \left[\frac{\sqrt{r^2/2}}{\Gamma(1/2)} n_0^{-1/2} \exp\left(-\frac{r^2}{2} n_0\right)\right] dn_0 \\ &\propto \int_0^\infty \exp\left[-\frac{n_0}{2}\left(\frac{1}{\sigma^2}(\mu - m_0)^2 + r^2\right)\right] dn_0 \\ &\propto \left(r^2 + \frac{1}{\sigma^2}(\mu - m_0)^2\right)^{-1} \\ &\propto \left(1 + \frac{(\mu - m_0)^2}{\sigma^2 r^2}\right)^{-1}. \end{aligned}$$

This is a special t-distribution, with degree of freedom $\nu = 1$, that is,

$$p(\mu \mid \sigma^2) = \frac{1}{\pi\sigma r}\left(1 + \frac{(\mu - m_0)^2}{\sigma^2 r^2}\right)^{-1},$$

the Cauchy distribution with location m_0 and scale σr.

With the Jeffrey reference prior for σ^2

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

the joint prior distribution is proportional to

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2) \propto \frac{1}{\pi\sigma^3 r}\left(1 + \frac{(\mu - m_0)^2}{\sigma^2 r^2}\right)^{-1}.$$

This is the Jeffrey-Zellner-Siow prior. This prior does not form any conjugacy with any distribution, so we need to use simulation method to get the posterior distribution of μ and σ^2.

Bayes Factors: Hypothesis Testing for Means with Known σ^2

In this file, we only demonstrate the calculation of Bayes factor for the hypothesis testing of one sample mean. The calculation of Bayes factor for other hypothesis testing is similar, but with a more complicated calculation steps. So we will not include them here.

Now we consider the two competing hypotheses of a mean parameter μ, under the situation when the variance σ^2 is known

$$H_1 : \mu = m_0, \qquad H_2 : \mu \neq m_0.$$

We use the Bayes factor to compare the two hypothese. Since in H_1, μ is given as m_0 (and σ^2 is always given), the likelihood under H_1 is purely

$$p(\text{data} \mid H_1) = p(\text{data} \mid \mu = m_0, \sigma^2).$$

For H_2, we impose the prior of μ as the Normal distribution (since we use the Normal-Normal conjugate family) with mean m_0 (we believe μ is around m_n) and variance σ^2/n_0. We keep another hyperparameter n_0 to ensure flexibility of the spread. Then the likelihood of H_2 can be calculated as

$$p(\text{data} \mid H_2) = \int p(\text{data} \mid \mu, \sigma^2) p(\mu \mid m_0, n_0, \sigma^2)\, d\mu.$$

Suppose we have n data points $y_1, \ldots, y_n$, each is independent and identically normally distributed, the likelihood of H_1 is simply the product of n Normal distributions with mean m_0 and variance σ^2

$$p(\text{data} \mid H_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_i - m_0)^2}{\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m_0)^2\right).$$

The likelihood of H_2, however, looks a little complicated

$$p(\text{data} \mid H_2) = \int \left[\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)\right] \times \left[\frac{\sqrt{n_0}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{n_0}{2\sigma^2}(\mu - m_0)^2\right)\right] d\mu$$

It requires some algebra to combine terms, and change of variables to normalize. Here, we only present the final result

$$p(\text{data} \mid H_2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \left(\frac{n_0}{n_0 + n}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 + \frac{nn_0}{n_0 + n}(\bar{y} - m_0)^2\right]\right).$$

Here $\bar{y}$ is the sample mean.

Therefore, the Bayes factor is

$$BF[H_1 : H_2] = \frac{p(\text{data} \mid H_1)}{p(\text{data} \mid H_2)} = \left(\frac{n_0 + n}{n_0}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\frac{n^2}{n_0 + n}(\bar{y} - m_0)^2\right) = \left(\frac{n_0 + n}{n_0}\right)^{1/2} \exp\left(-\frac{1}{2}\frac{n}{n_0 + n}Z^2\right),$$

where Z is the z-score

$$Z = \frac{\bar{y} - m_0}{\sigma/\sqrt{n}}.$$

R Code to Compute Bayes Factor

The package `BayesFactor` provide a function `ttestBF` for one sample mean, two paired means and two independent means. This function only provide the simulation result from the Jeffrey-Zellner-Siow prior. For more prior results and construction of credible interval, we provide the `bayes_inference` function in the `statsr` package.

We will demonstrate the use of `ttestBF` on the `tthm` variable in the `tapwater` data set. The sample mean of the variable is about 55.5. And we will like to see whether the parameter mean is 50

$$H_1 : \mu = 50, \qquad H_2 : \mu \neq 50.$$

```r
# Load library
library(BayesFactor)

# Load data from `statsr` package
library(statsr)
data(tapwater)

# Inference for `tthm` variable
bf = ttestBF(tapwater$tthm, mu = 50)  # default `mu` value is 0
bf
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 0.4079531 ±0.02%
##
## Against denominator:
##   Null, mu = 50
## ---
## Bayes factor type: BFoneSample, JZS
```

The analysis shows that, the Bayes factor of the althernative hypothesis H_2 against the null hypothesis H_1 is about $BF[H_2 : H_1] = 0.408$. That means, the Bayes factor of the null hypothesis against the alternative is

$$BF[H_1 : H_2] = \frac{1}{BF[H_2 : H_1]} = \frac{1}{0.408} \approx 2.45,$$

which can be done using

```r
1/bf
```

```
## Bayes factor analysis
## --------------
## [1] Null, mu=50 : 2.451262 ±0.02%
##
## Against denominator:
##   Alternative, r = 0.707106781186548, mu =/= 50
## ---
## Bayes factor type: BFoneSample, JZS
```

According to the Jeffrey's scale, the difference between the two hypotheses are not worth a bare mention.