Week 3 Lecture Video Explanations Part I

Duke University

In this file, students are assumed to have watched all videos of Week 3, have gone over the Summary file for this Week. Therefore, we will not repeat some of the definitions of the key concepts. Instead, we will dive into more depth and will present the calculations and R codes used in these lecture videos.

Videos about **Losses and Decision Making** are mainly covered in the Summary file. Students can refer to the Summary file for more details. This explanation file, however, will mainly focus on the next two parts of Week 3.

1 Why Bayes Factor

Before we jump into the details about how to use Bayes factor to test hypotheses, let us talk briefly explain the reason why we care about Bayes factor.

Recall in Week 1, when we had two "hypotheses" about a coin, either it is fair (with probability $p = 0.5$ to get heads) or it is loaded (with probability $p = 0.8$ to get heads). In order to update our beliefs, we calculated the posterior probability using Bayes' Rule:

$$\mathbb{P}(p = 0.5 \mid \text{data}) = \frac{\mathbb{P}(\text{data} \mid p = 0.5)\mathbb{P}(p = 0.5)}{\mathbb{P}(\text{data} \mid p = 0.5)\mathbb{P}(p = 0.5) + \mathbb{P}(\text{data} \mid p = 0.8)\mathbb{P}(p = 0.8)}$$

$$\mathbb{P}(p = 0.8 \mid \text{data}) = \frac{\mathbb{P}(\text{data} \mid p = 0.8)\mathbb{P}(p = 0.8)}{\mathbb{P}(\text{data} \mid p = 0.5)\mathbb{P}(p = 0.5) + \mathbb{P}(\text{data} \mid p = 0.8)\mathbb{P}(p = 0.8)}$$

We see that the posterior probabilities $\mathbb{P}(p = 0.5 \mid \text{data})$ and $\mathbb{P}(p = 0.8 \mid \text{data})$ depend on not only the likelihood from the observed data, but also the prior probabilities of the two "hypotheses": $\mathbb{P}(p = 0.5)$ and $\mathbb{P}(p = 0.8)$. Suppose two people argue about the probability of this coin getting heads. Even if they have agreed that there are only 2 possible values of p, either $p = 0.5$ or $p = 0.8$, they may have different beliefs about $\mathbb{P}(p = 0.5)$ and $\mathbb{P}(p = 0.8)$. One may think there is higher chance for the coin is fair, so this person places $\mathbb{P}(p = 0.5) = 0.9$ and $\mathbb{P}(p = 0.8) = 0.1$. Another may strongly feel the coin is loaded, so this person places $\mathbb{P}(p = 0.8) = 0.9$ and $\mathbb{P}(p = 0.5) = 0.1$. We have discussed the effect of placing different beliefs. We say if our prior beliefs are far from the true situation, we need to have **enough** data to "correct" our priors. However, if we only have limited amount of data to compare two hypotheses, different people having different beliefs may end up getting different posterior probabilities of the hypotheses, which lead to different conclusion.

We try to eliminate the effect and make sure when comparing about hypotheses, everybody is on the same ground. That is, we do not talk about the posterior probabilities, instead, we emphasize on how much the prior probabilities get "amplified" by the observed data. This comes to the discussion about Bayes factors. Bayes factor measures how much the prior odds between the two hypotheses get "scaled" up or down by the data. For example, in the above coin case, we define

$$\text{BF}[H_1 : H_2] = \frac{\text{PO}[H_1 : H_2]}{\text{O}[H_1 : H_2]} = \frac{\mathbb{P}(\text{data} \mid p = 0.5)}{\mathbb{P}(\text{data} \mid p = 0.8)}.$$

This Bayes factor does not depend on any prior assumptions of the two hypotheses. So no matter how much one leans on one hypothesis than another, as long as people agree with the probability distribution of the parameter p, they can compare the two hypotheses on the same scale.

2 Comparing Two Proportions

We will cover the two videos **Comparing Two Proportions Using Bayes Factors: Assumptions** and **Comparing Two Proportions Using Bayes Factors** together. In the two videos, we use the **Parent's Perception of Bullying** as an example to compare the difference between frequentist approach and Bayesian approach. Our goal is to compare the two hypotheses:

$$\begin{aligned} H_1 &: p_{\text{male}} = p_{\text{female}}, \\ H_2 &: p_{\text{male}} \neq p_{\text{female}}. \end{aligned}$$

2.1 Review: Frequentist Approach

Recall that in the frequentist approach, we have the **null hypothesis**, saying that "nothing is happening. Under this context, it will be the H_1 hypothesis we presented above. And the **alternative hypothesis** is a different hypothesis that we may draw conclusion of based on the data we have obtained. So to restate the question under the frequentist approach, we are comparing:

$$\begin{aligned} H_0 &: p_{\text{male}} = p_{\text{female}}, \\ H_A &: p_{\text{male}} \neq p_{\text{female}}. \end{aligned}$$

In order the compare the two, we calculate the ***p*-value**, which is defined to be

$$p\text{-value } = \mathbb{P}(\text{the observed situation or more extreme situation happens} \mid \text{null hypothesis is true})$$

We rely on the p-value to make decision whether to reject H_0 or not. Intuitively speaking, the frequentist approach argues against null hypothesis by assuming the population behaves according to what null hypothesis states. Therefore, the p-value here describes the probability that the real world data "would follow" the null hypothesis assumption. One can see that the alternative hypothesis never enters into the calculation of the p-value in the frequentist approach. It only affects whether we should pick a two-sided p-value or one-sided p-value for comparison. Moreover, we do not have much flexibility for hypotheses. For example, the frequentist approach would not handle hypotheses as follow:

$$\begin{aligned} H_0 &: p = 0.5, \\ H_A &: p = 0.8. \end{aligned}$$

Assumptions to Conduct Hypothesis Test

We present the assumptions under the frequentist approach for comparison purpose:

- Responses are independence (both within groups and between groups).
- Within group, the rate of reporting is the same, so that we can use just one p_{male} to represent the male group and one p_{female} to represent the female group.
- We also require normality in order to use the z-score. This is achieved by requiring both the number of successes and the number of failures are at least 10 **under the null hypothesis**. If you recall, in the Inferential Statistics course, we used the **pooled** rate

  $$p_{\text{pooled}} = \frac{\#\text{ of successes in group 1} + \#\text{ of successes in group 2}}{\#\text{ of group 1 responses} + \#\text{ of group 2 responses}}.$$

  (In the Bayesian approach, however, we will no longer require the sample size to be large enough. This shows that the Bayesian approach is more flexible than the frequentist approach.)

2.2 Bayesian Approach

The Bayesian approach is quite different from the frequentist approach, mainly in the following aspects:

- We no longer require normality in the sample. Therefore, we are more flexible about sample size. This is because we are using a completely different mechanism, that is, to update the probability distribution of the parameters of interest. This do not require any approximation from the normal distribution or t-distribution.
- We need to obtain results for both hypotheses, and we do not assume one hypothesis is more important than another. The 2 competing hypotheses are viewed equally.

But there are also some drawbacks for Bayesian approach, which we will mention in the later videos:

- Different priors will lead to different conclusions, and it is hard to obtain a suitable prior from just one sample of observed data.
- It is way harder to calculate, especially when we do not have conjugate families. Therefore, we can only use simulation to get approximate results most of the time.
- While credible intervals give exactly the probability of the parameter falling into the interval, they are not unique. We also have different metrics to choose, which may result in different point estimates (mean/median/mode) of the parameter. And these will affect our decision making. Due to subject complexity, we do not spend too much time on point estimates in this course.

Let us go over the example in the video step by step to see how to apply Bayes factor for hypothesis testing. Recall that Bayes factor compares how one hypothesis is **relative stronger** than another one. It does not provide any absolute conclusion of the posterior probabilities of the two hypotheses.

2.3 Parent's Perception of Bullying Example

In this example, we provide two hypotheses:

$$H_1 : p_{\text{male}} = p_{\text{female}} \qquad\qquad H_2 : p_{\text{male}} \neq p_{\text{female}}.$$

To calculate the Bayes factor $\text{BF}[H_1 : H_2] = \dfrac{\mathbb{P}(\text{data} \mid H_1)}{\mathbb{P}(\text{data} \mid H_2)}$ directly, we need to get $\mathbb{P}(\text{data} \mid H_1)$ and $\mathbb{P}(\text{data} \mid H_2)$. But are we going to use the observed $\hat{p}_{\text{male}}$ and $\hat{p}_{\text{female}}$ to calculate the these probabilities? The major difference between frequentist approach and Bayesian approach is, data only provide the updating process based on our belief of the entire population. We do not use the observed statistics of the data directly to test our hypotheses. Recall that the Bayesian process that

$$\text{posterior of } p_{\text{male}} \propto \mathbb{P}(\text{data} \mid p_{\text{male}}) \times \text{prior of } p_{\text{male}}.$$

We know that we cannot easily draw conclusion about the values of p_{male} and p_{female} **without providing prior distributions** of the two proportions.

To calculate Bayes factor for comparison, we first need to set up a common "agreement", the prior distributions for each hypothesis.

Our Priors

Since we can view the data as a Binomial process. We have decided to use the **Beta-Binomial conjugate family** for the update. Here we use the following set of notations:

a_m : prior hyperparameter of p_{male}	b_m : another prior hyperparameter of p_{male}
n_m : total # of male responses in data	R_m : # of male reporting bullying in the data
a_f : prior hyperparameter of p_{female}	b_f : another prior hyperparameter of p_{female}
n_f : total # of female responses in data	R_f : # of female reporting bullying in the data

A question will arise: how to set up $a_m,\ b_m$, and likewise, a_f and b_f?

Recall from Week 2 Conjugacy file, we have mentioned that the sum of the prior hyperparameters $a_m + b_m$ represents the **effective sample size** (recall the video also mentions "imaginary sample"). And the ratio $\dfrac{a_m}{a_m + b_m}$ represents the success rate in our belief. The larger the effective sample size is, the more strong that we think our belief represents the true success rate. In this example, we pick $a_m = b_m = \dfrac{1}{2}$, so that the effective sample size is just 1 (this is the Jeffrey's prior, which the lecture will cover more in Week 4). And our belief of success rate is $\dfrac{a_m}{a_m + b_m} = \dfrac{1/2}{(1/2 + 1/2)} = \dfrac{1}{2}$. It is implied that we do not want our belief to have a strong effect on the posterior results, and would like to be very "neutral" of the success rate. The analysis of the values of a_f and b_f is similar.

When we consider the situation H_1, since we need to keep our belief consistent so that we can make comparison based on the data, we have to set the same effective sample size and same prior ratio. Previously, we assume the effective sample sizes are 1 respectively, with ratio $\dfrac{1}{2}$. To keep consistency, when we pool the data together, we assume the effective sample size of the total parents is 2 ($(a_m + b_m) + (a_f + b_f) = 1 + 1 = 2$), with prior ratio still $\dfrac{1}{2}$. Therefore, we assign the prior $\text{Beta}(a_m + a_f = 1, b_m + b_f = 1)$ to the prior of H_1. This is exactly the formula we see that

$$P(p) = \text{Beta}(1, 1) \propto p^{a_m+a_f-1}(1-p)^{b_m+b_f-1}$$

(Here $P(p)$ means the prior probability of p under hypothesis H_1. Since hypothesis H_1 is about a point estimate statement of p, $P(p)$ has in fact become the prior probability distribution of p, i.e., $\pi(p)$.)

Let us summarize. We have priors:

- Under H_1: $\pi(p \mid H_1) \sim \text{Beta}(1, 1)$
- Under H_2: $\pi(p \mid H_2) = \pi(p_{\text{male}}) \times \pi(p_{\text{female}}) \sim \text{Beta}(0.5, 0.5) \times \text{Beta}(0.5, 0.5)$

Calculate BF$[H_1 : H_2]$ (optional)

Since we assume the success rates in both hypotheses H_1 and H_2 take continuous values, we need to use the continuous version of the formula of Bayes factor

$$\text{BF}[H_1 : H_2] = \frac{\mathbb{P}(\text{data} \mid H_1)}{\mathbb{P}(\text{data} \mid H_2)} = \frac{\displaystyle\int \mathbb{P}(\text{data} \mid p_{\text{pool}}, H_1)\pi(p_{\text{pool}} \mid H_1)\, dp_{\text{pool}}}{\displaystyle\iint \mathbb{P}(\text{data} \mid p_{\text{male}}, H_2)\mathbb{P}(\text{data} \mid p_{\text{female}}, H_2)\pi(p_{\text{male}} \mid H_2)\pi(p_{\text{female}} \mid H_2)\, dp_{\text{male}}\, dp_{\text{female}}}$$

Since female and male responses are independent, we can spli the integral in the denominator into a product of integrals. Here, once we are clear of the priors of p_{pool}, p_{male}, and p_{female}, we can omit the notation H_1, H_2 just for simplification. We also just write the integral variables p when it is clear what prior each p takes.

$$\text{BF}[H_1 : H_2] = \frac{\displaystyle\int \mathbb{P}(\text{pooled data} \mid p)\text{Beta}(1, 1)\, dp}{\left(\displaystyle\int \mathbb{P}(\text{male data} \mid p)\text{Beta}(0.5, 0.5)\, dp\right) \times \left(\displaystyle\int \mathbb{P}(\text{female data} \mid p)\text{Beta}(0.5, 0.5)\, dp\right)}$$

It is possible to calculate the integrals. But once we recognize these integrals as the denominators in the Bayes' Rule, we may use quantities that we have known to replace these integrals, so that we can avoid integral calculations.

We have actually seen these integrals before, when we tried to apply the Bayes' Rule to derive the Beta-Binomial conjugacy and to update the prior of p. Recall that

$$\text{posterior of } p \;= \frac{\mathbb{P}(\text{data} \mid p) \times \text{Beta}(\alpha, \beta)}{\displaystyle\int \mathbb{P}(\text{data} \mid p) \times \text{Beta}(\alpha, \beta)\, dp}$$

We we know

$$\int \mathbb{P}(\text{data} \mid p) \times \text{Beta}(\alpha, \beta)\, dp = \frac{\mathbb{P}(\text{data} \mid p) \times \text{Beta}(\alpha, \beta)}{\text{posterior of } p} = \frac{\mathbb{P}(\text{data} \mid p) \times \text{Beta}(\alpha, \beta)}{\text{Beta}(\text{new } \alpha, \text{ new } \beta)}$$

Under the 3 updates, we have

$$\int \mathbb{P}(\text{pooled data} \mid p)\text{Beta}(1,1)\,dp = \frac{\mathbb{P}(\text{pooled data} \mid p) \qquad \overbrace{\text{Beta}(1,1)}^{\text{Beta}(a_m+a_f,\,b_m+b_f)}}{\underbrace{\text{Beta}(96,118)}_{\text{Beta}(a_m+a_f+(R_m+R_f),\,b_m+b_f+(n_m+n_f)-(R_m+R_f))}}$$

$$\int \mathbb{P}(\text{male data} \mid p)\text{Beta}(0.5,0.5)\,dp = \frac{\mathbb{P}(\text{male data} \mid p)\,\overbrace{\text{Beta}(0.5,0.5)}^{\text{Beta}(a_m,\,b_m)}}{\underbrace{\text{Beta}(34.5,56.5)}_{\text{Beta}(a_m+R_m,\,b_m+n_m-R_m)}}$$

$$\int \mathbb{P}(\text{female data} \mid p)\text{Beta}(0.5,0.5)\,dp = \frac{\mathbb{P}(\text{female data} \mid p)\,\overbrace{\text{Beta}(0.5,0.5)}^{\text{Beta}(a_f,\,b_f)}}{\underbrace{\text{Beta}(61.5,61.5)}_{\text{Beta}(a_f+R_f,\,b_f+n_f-R_f)}}$$

Combining all together, we have

$$\begin{aligned}\text{BF}[H_1 : H_2] =& \frac{\mathbb{P}(\text{pooled data} \mid p)\text{Beta}(1,1)}{\text{Beta}(96,118)} \div \left[\frac{\mathbb{P}(\text{male data} \mid p)\text{Beta}(0.5,0.5)}{\text{Beta}(34.5,56.5)} \times \frac{\mathbb{P}(\text{female data} \mid p)\text{Beta}(0.5,0.5)}{\text{Beta}(61.5,61.5)}\right] \\ =& \frac{\mathbb{P}(\text{pooled data} \mid p)}{\mathbb{P}(\text{male data} \mid p)\mathbb{P}(\text{female data} \mid p)} \times \frac{\text{Beta}(1,1) \times \text{Beta}(34.5,56.5) \times \text{Beta}(61.5,61.5)}{\text{Beta}(96,\ 118) \times \text{Beta}(0.5,0.5) \times \text{Beta}(0.5,0.5)}\end{aligned}$$

Finally, we argue that

$$\frac{\mathbb{P}(\text{pooled data} \mid p)}{\mathbb{P}(\text{male data} \mid p) \times \mathbb{P}(\text{female data} \mid p)} = 1$$

To see this, recall that the pooled data is just a combination of all responses from male parents and female parents. Since we assume the survey responses are independent, we can view the **pooled** process as sequential update. So the probability of pooling 212 responses between 90 men and 122 women, with 34 men saying ”yes” and 61 women saying ”yes” is:

$$\mathbb{P}(\text{pooled data} \mid p) = \binom{90}{34}\binom{122}{61}p^{34+61}(1-p)^{56+61}.$$

This number is exactly

$$\mathbb{P}(\text{male data} \mid p) \times \mathbb{P}(\text{female data} \mid p) = \left[\binom{90}{34}p^{34}(1-p)^{56}\right]\left[\binom{122}{61}p^{61}(1-p)^{61}\right].$$

Now we have

$$\text{BF}[H_1 : H_2] = \frac{\text{Beta}(1,1) \times \text{Beta}(34.5,56.5) \times \text{Beta}(61.5,61.5)}{\text{Beta}(96,\ 118) \times \text{Beta}(0.5,0.5) \times \text{Beta}(0.5,0.5)}.$$

It seems the right-hand side will give us a function. However, from the definition of the Beta distribution

$$\text{Beta}(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1},$$

we are lucky enough have all the powers α’s and β’s cancel each other through the fraction (due to the fact that we keep consistency when we set our priors). What is left are only the Γ functions.

To further simply our notation, we introduce the Beta function (not the Beta distribution), where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

which is the reciprocal of the constant in front of the Beta distribution. With this notation, we can rewrite the formula for the Bayes factor:

$$\text{BF}[H_1 : H_2] = \frac{\overbrace{B(96, 118)}^{B(a_m+a_f+R_m+R_f, b_m+b_f+(n_m+n_f)-(R_m+R_f))}}{\underbrace{B(1, 1)}_{B(a_m+a_f, b_m+b_f)}} \div \left[\frac{\overbrace{B(34.5, 56.5)}^{B(a_m+R_m, b_m+n_m-R_m)}}{\underbrace{B(0.5, 0.5)}_{B(a_m, b_m)}} \times \frac{\overbrace{B(61.5, 61.5)}^{B(a_f+R_f, b_f+n_f-R_f)}}{\underbrace{B(0.5, 0.5)}_{B(a_f, b_f)}} \right] \approx 2.927814.$$

This is the formula you see at 1:43 in the 2nd video. You may use `beta` in R to calculate the value of $B(a, b)$.

The Bayes factor 2.93 is definitely “not worth a bare mention”. This Bayes factor just tells us how much the odd of H_1 over H_2 will get amplified after observing the data. However, Bayes factor does not tell us how much the probability of H_1 or H_2 will increase after observing the data. Hence, in the video, we also mentioned the posterior probability of H_1.

Calculate Posterior Probability of H_1

We will explain how to use the above tricks and the value of $\text{BF}[H_1 : H_2]$ to get an easy formula of $\mathbb{P}(H_1 \mid \text{data})$ without using Bayes’ Rule and calculating integrals. You will also see the purpose of deriving formulas on the slide at 1:28 in the 2nd video under this topic.

Recall that

$$\text{PO}[H_1 : H_2] = \frac{\mathbb{P}(H_1 \mid \text{data})}{\mathbb{P}(H_2 \mid \text{data})}.$$

We are lucky in this example, that H_1 and H_2 are two complementary events:

$$\mathbb{P}(H_1 \mid \text{data}) + \mathbb{P}(H_2 \mid \text{data}) = 1.$$

Therefore,

$$\text{PO}[H_1 : H_2] = \frac{\mathbb{P}(H_1 \mid \text{data})}{1 - \mathbb{P}(H_1 \mid \text{data})}.$$

Solving for $\mathbb{P}(H_1 \mid \text{data})$ in terms of the posterior odd $\text{PO}[H_1 : H_2]$, we get

$$\mathbb{P}(H_1 \mid \text{data}) = \frac{\text{PO}[H_1 : H_2]}{\text{PO}[H_1 : H_2] + 1},$$

which is the first formula in the slide at 1:28.

Since prior odd and posterior odd are similar (one is before data, another is after data), we should be able to convince ourselves that the following is also true:

$$\text{O}[H_1 : H_2] = \frac{\mathbb{P}(H_1)}{1 - \mathbb{P}(H_1)},$$

which is the last formula in the same slide.

Now we use $\text{PO}[H_1 : H_2] = \text{BF}[H_1 : H_2] \times \text{O}[H_1 : H_2]$ again, we can derive

$$\mathbb{P}(H_1 \mid \text{data}) = \frac{\text{PO}[H_1 : H_2]}{\text{PO}[H_1 : H_2] + 1} = \frac{\text{BF}[H_1 : H_2] \times \text{O}[H_1 : H_2]}{\text{BF}[H_1 : H_2] \times \text{O}[H_1 : H_2] + 1}.$$

Suppose we think there is a 50-50 chance for either H_1 or H_2 to happen before seeing the data, that is, we assume equal prior odd, i.e., $\text{O}[H_1 : H_2] = 1$, then we have

$$\mathbb{P}(H_1 \mid \text{data}) = \frac{\text{BF}[H_1 : H_2]}{\text{BF}[H_1 : H_2] + 1} = \frac{2.93}{2.93 + 1} \approx 0.75.$$

(The above formula is only true when we impose equal prior odd. In this way, we can examine the effect of Bayes factor on the posterior probability of H_1 without worrying the relative size of the prior probabilities between H_1 and H_2. However, in general $\mathbb{P}(H_1 \mid \text{data})$ also depends on $\text{O}[H_1 : H_2]$.)

3 Comparing Two Paired Means

3.1 Normal-Gamma Joint Distribution for μ and σ: When σ Is Unknown

Recall from last week, we have discussed the **Normal-Normal conjugate family** for the situation when the **data** standard deviation σ is known. In that case, we only need to focus on the distribution of the **data** mean μ and discuss how to update the distribution using conjugacy. However, in most real world cases, we would not have previous information of σ, and fixing σ for the problem would not be appropriate. Therefore, we still need to come up with a way to update the joint probability distribution for both μ and σ. Let us denote this joint probability distribution

$$\pi(\mu, \sigma^2).$$

(As usual, I follow the convention to use the variance σ^2.)

What would be a good distribution for this joint distribution so that we could possibly get a conjugate family? Well, if σ^2 is known, then we are good to go. So one way to think about this problem is, suppose we have the distribution of σ^2, denoted as $\pi(\sigma^2)$, then use the usual Normal distribution of μ conditioning on $\pi(\sigma^2)$, we will return to the usual Normal-Normal conjugacy. To execute this idea, we apply the conditional probability

$$\pi(\mu, \sigma^2) = \pi(\mu \mid \sigma^2) \times \pi(\sigma^2).$$

So now the key point has become, what would be a good distribution of σ^2 so that we could still maintain the Normal-Normal conjugacy? The answer is, the **Inverse Gamma** distribution. For notional convenience and the reason that we hate to use "inverse", let us denote the inverse of σ^2 as

$$\phi = \frac{1}{\sigma^2}.$$

This is called **precision** of the data in statistics. If we impose σ^2 to follow the inverse Gamma distribution, then the precision ϕ, which is the inverse of σ^2, should follow the Gamma distribution. We will see in the later calculation, using ϕ is more convenient than keeping track of σ^2. Hence, from now on, we will use $\pi(\mu, \phi)$ to denote the prior joint distribution and use μ and ϕ as our parameters of interest. If we want to get information of the variance σ^2, we will only need to invert ϕ.

Priors

We assign the prior distribution of the data precision as

$$\pi(\phi) = \text{Gamma}\left(\frac{\nu_0}{2},\ \sigma_0^2\frac{\nu_0}{2}\right), \qquad \text{with hyperparameters } \nu_0 \text{ and } \sigma_0^2.$$

Here we use the definition of Gamma distribution defined in Week 2 Lab:

$$\Gamma(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}.$$

So to correspond the 2 sets of hyperparameters, we have

$$\alpha = \frac{\nu_0}{2} \Longrightarrow \nu_0 = 2\alpha, \qquad \beta = \sigma_0^2\frac{\nu_0}{2} \Longrightarrow \sigma_0^2 = \frac{\beta}{\alpha}.$$

(Later we will talk about why we chose ν_0 and σ_0^2 instead of the usual α and β.)

Since $\phi = \frac{1}{\sigma^2}$, the prior distribution of σ^2 is

$$\pi(\sigma^2) = \frac{1}{\text{Gamma}(\nu_0/2,\ \sigma_0^2\nu_0/2)}.$$

However, we will not use this formula since using the precision ϕ is more convenient. We will see this in the definition of the Normal distribution for μ.

Now assuming we have the distribution of σ^2 (which is given by the prior distribution of ϕ), we may "pretend" that we are back to the situation when σ^2 is known, and we adopt the Normal distribution as the prior distribution for μ:

$$\pi(\mu \mid \phi) = \pi(\mu \mid 1/\sigma^2) = \mathcal{N}\left(m_0,\ \frac{\sigma^2}{n_0}\right), \qquad \text{with hyperparameters } m_0 \text{ and } n_0.$$

We understand the above conditional distribution as follows: suppose we have obtained information of the data variance σ^2 (which is given by ϕ), we assume the mean μ of the data follows the Normal distribution with mean m_0 (this is the mean of the parameter μ_0, not the mean of the data), and variance $\dfrac{\sigma^2}{n_0}$.

Recall from the file Conjugacy of Week 2, we argue that if σ^2 is known, and μ follows the Normal distribution $\mathcal{N}(\nu,\ \tau^2)$, then the effective sample size of the prior distribution of μ is $\dfrac{\sigma^2}{\tau^2}$.

Now for $\pi(\mu \mid 1/\sigma^2)$, we can see the **effective sample size** of this conditional prior distribution given σ^2 is $\dfrac{\sigma^2}{(\sigma^2/n_0)} = n_0$. This is the meaning of n_0.

When we **put everything together**, we get the formula of the prior joint distribution

$$\pi(\mu,\phi) = \pi(\mu, 1/\sigma^2) = \pi(\mu \mid 1/\sigma^2)\pi(1/\sigma^2) = \pi(\mu \mid \phi)\pi(\phi).$$

To see that the last quantity is the better formula for $\pi(\mu,\sigma^2)$, let us write out the details of $\pi(\mu \mid \sigma^2)$. We see that using ϕ, we can avoid division of σ^2.

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

Now the multiplication seems easier to manipulate:

$$\begin{aligned}\pi(\mu,\phi) =& \pi(\mu \mid \phi)\times\pi(\phi)\\ =& \frac{\sqrt{n_0}}{\sqrt{2\pi}}\phi^{1/2}\exp\left(-\frac{n_0\phi(\mu-m_0)^2}{2}\right)\times\frac{\left(\sigma_0^2\frac{\nu_0}{2}\right)^{\nu_0/2}}{\Gamma(\nu_0/2)}\phi^{\nu_0/2-1}\exp\left(-\sigma_0^2(\nu_0/2)\phi\right)\\ \propto& \phi^{(\nu_0-1)/2}\exp\left(-\frac{n_0(\mu-m_0)^2+\sigma_0^2\nu_0}{2}\phi\right).\end{aligned}$$

This prior distribution is called the **Normal-Gamma** distribution, with 4 hyperparameters:

$$\pi(\mu,\phi) \sim \text{NormalGamma}(m_0, n_0, \sigma_0^2, \nu_0).$$

The definition of Normal-Gamma distribution is given by:

$$\text{NormalGamma}(\mu,\phi;\ m_0,\ n_0,\ \sigma_0^2,\ \nu_0) = \sqrt{\frac{n_0}{2\pi}}\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\phi^{(\nu_0-1)/2}\exp\left(-\frac{n_0(\mu-m_0)^2+\sigma_0^2\nu_0}{2}\phi\right).$$

The **marginal distribution** of μ if μ,ϕ is under the Normal-Gamma distribution can be computed integrating all over ϕ:

$$\pi(\mu) = \int_0^\infty \pi(\mu,\phi)\,d\phi = \int_0^\infty \sqrt{\frac{n_0}{2\pi}}\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\phi^{(\nu_0-1)/2}\exp\left(-\frac{n_0(\mu-m_0)^2+\sigma_0^2\nu_0}{2}\phi\right)\,d\phi.$$

The resulting distribution is the Student's t-distribution with degree of freedom ν_0 (that is why we stick with ν_0 but not α), center at m_0, and variance $\dfrac{\sigma_0^2}{n_0}$ (that is why we stick with σ_0^2 but not β)

$$\mu \sim t_{\nu_0}\left(m_0,\ \frac{\sigma_0^2}{n_0}\right)$$

3.2 NormalGamma-Normal Conjugacy

We illustrate the conjugacy by updating with 1 data point x_1 under the Normal distribution with mean μ and precision ϕ. From the Bayes' Rule, we have

$$\begin{aligned}
\pi^*(\mu,\phi) \propto & f(x_1 \mid \mu,\phi) \times \pi(\mu,\phi) \\
\propto & \left[\frac{1}{\sqrt{2\pi}}\phi^{1/2}\exp\left(-\frac{\phi(x_1-\mu)^2}{2}\right)\right] \times \phi^{(\nu_0-1)/2}\exp\left(-\frac{n_0(\mu-m_0)^2+\sigma_0^2\nu_0}{2}\phi\right) \\
\propto & \phi^{\frac{(\nu_0+1)-1}{2}}\exp\left[-\frac{(n_0+1)\left(\mu-\frac{n_0m_0+x_1}{n_0+1}\right)^2+\sigma_1^2(\nu_0+1)}{2}\phi\right] \\
= & \text{NormalGamma}(m_1=\frac{n_0m_0+x_1}{n_0+1},\ n_1=n_0+1,\ \sigma_1^2,\ \nu_1=\nu_0+1),
\end{aligned}$$

where

$$\sigma_1^2=\frac{\sigma_0^2\nu_0+n_0m_0^2+x_1^2-\frac{(n_0m_0+x_1)^2}{n_0+1}}{\nu_0+1}=\frac{\sigma_0^2\nu_0+\frac{n_0}{n_0+1}(x_1-m_0)^2}{v_0+1}.$$

In general, when we have a set of n data points $\{x_1,\cdots,x_n\}$ which are independent and each follows the Normal distribution with mean μ and precision ϕ. Then we have

$$\pi(\mu,\phi)=\text{NormalGamma}(m_0,\ n_0,\ \sigma_0^2,\ \nu_0)\xrightarrow{n\text{ data points independently}\sim\mathcal{N}(\mu,\ \phi)}\text{NormalGamma}(m_n,\ n_n,\ \sigma_n^2,\ \nu_n)=\pi^*(\mu,\phi\mid\text{data}),$$

where

$$\begin{aligned}
m_n =& \frac{\sum x_i+n_0m_0}{n+n_0}=\frac{n\bar{D}+n_0m_0}{n+n_0} \\
n_n =& n_0+n \\
\nu_n =& \nu_0+n \\
\sigma_n^2 =& \frac{1}{\nu_0+n}\left(\sigma_0^2\nu_0+\sum_i\left(x_i-\bar{D}\right)^2+\frac{n_0n}{n_0+n}\left(\bar{D}-m_0\right)^2\right), \qquad \bar{D}=\frac{1}{n}\sum_i x_i \quad \text{is the sample mean.}
\end{aligned}$$

The marginal distribution of μ after observing data is the t-distribution with degree of freedom ν_n, center at m_n and variance σ_n^2/n_n:

$$\mu\mid\text{data}\sim t_{\nu_n}\left(m_n,\ \frac{\sigma_n^2}{n_n}\right) \qquad \Longrightarrow \qquad \frac{\mu-m_n}{\sigma_n/\sqrt{n_n}}\mid\text{data} \quad \text{follows the standardized } t\text{-distribution with degree of freedom } \nu_n$$

Marginal Distribution of Sample Mean (optional)

We can also calculate the marginal distribution of the sample mean $\bar{D}$ of the data $\{x_1,\cdots,x_n\}$ if we assume the data follow the Normal distribution, under the assumption, $\bar{D}$ will follow a Normal distribution with mean m_0, and variance $\frac{\sigma_0^2}{n_0}+\sigma^2$. Then

$$\pi(\bar{D})=\iint\pi(\bar{D}|\mu,\sigma^2)\pi(\mu,\sigma^2)\,d\mu\,d\sigma^2,$$

We will not show the detailed derivation here, but if we assume that $\mu\sin\mathcal{N}(0,\sigma_0^2/n_0)$, we can show that $\bar{D}$ follows the Student's t-distribution shown at 4:18 on the slide of the first video. This part is not required either in the course. The purpose of mentioning this is to show you the connection between the Normal-Gamma distribution and the Student's t-distribution, which we are very familiar with when calculating probability and confidence intervals.

3.3 Bayes Factor

We will walk through the **Zinc Water** example. Our goal is to compare the two hypotheses:

$$H_1:\mu_{\text{diff}}=0, \qquad H_2:\mu_{\text{diff}}\neq 0$$

For H_1, we have provided the value of μ_{diff}, so we do not need to get a joint prior distribution of both μ and σ^2. For H_2, we place the following prior:

$$\mu_{\text{diff}} \mid \sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{n_0}), \qquad \text{and} \qquad \phi = \frac{1}{\sigma^2} \sim \text{Gamma}(\frac{\nu_0}{2}, \frac{\sigma_0^2\nu_0}{2})$$

Similar to the **Bullying Report Survey** example, we would like to place a good prior for μ and σ^2 (or ϕ). Since the prior of μ mainly depends on σ^2, the key here is to find a good prior of σ^2 so that our beliefs of the prior distribution will not affect too much of the analysis.

The prior we chose here is a non-informative reference prior which is in the limiting case when $\nu_0 \to 0$. If we get back to the Gamma distribution of ϕ, when $\nu_0 \to 0$, we have

$$\text{Gamma}(\phi;\ \nu_0, \sigma_0^2\nu_0/2) \to (\text{some constant }) \times \frac{1}{\phi}, \qquad \text{i.e.,Gamma}(\phi;\ \nu_0, \sigma_0^2\nu_0/2) \propto \frac{1}{\phi}.$$

Recall that the effective sample size of Gamma distribution is given by $\beta = \sigma_0^2\nu_0/2$, and the prior mean is $\dfrac{\alpha}{\beta} = \dfrac{\nu_0/2}{\sigma_0^2\nu_0/2} = \dfrac{1}{\sigma_0^2}$. By setting $\nu_0 \to 0$, we are restricting the information of the prior while keeping the prior mean unchanged.

Using the formula of Bayes factor

$$\begin{aligned}
\text{BF}[H_1 : H_2] =& \frac{\displaystyle\int f(\text{data} \mid \mu = 0,\ \phi)\pi(\phi)\, d\phi}{\displaystyle\iint f(\text{data} \mid \mu, \phi)\pi(\mu \mid \phi)\pi(\phi)\, d\mu\, d\phi} \\
=& \frac{\displaystyle\int f(\text{data} \mid \mu = 0,\ \phi)\frac{1}{\phi}\, d\phi}{\displaystyle\iint f(\text{data} \mid \mu,\ \phi)\pi(\mu \mid \phi)\frac{1}{\phi}\, d\phi} = \frac{\displaystyle\int f(\text{data} \mid \mu = 0\, \sigma^2)\frac{1}{\sigma^2}\, d\sigma^2}{\displaystyle\iint f(\text{data} \mid \mu, \sigma^2)\pi(\mu \mid \sigma^2)\frac{1}{\sigma^2}\, d\mu\, d\sigma^2}
\end{aligned}$$

(One can demonstrate that if $\text{Gamma}(\phi)d\phi \propto \dfrac{1}{\phi}d\phi$, then $\text{Inverse-Gamma}(\sigma^2)d\sigma^2 \propto \dfrac{1}{\sigma^2}d\sigma^2$, given that $\phi = \dfrac{1}{\sigma^2}$.)

Similar to the previous two proportion example, we do not need to calculate this integral by hands. We could view the two integrals as the normalizing constant appearing in the Bayes' Rule when we update the distribution of the parameters, and replace the integrals by those formulas. We will adopt the final result from the slide without showing the details.

$$\text{BF}[H_1 : H_2] = \left(\frac{n + n_0}{n_0}\right)^{1/2} \left(\frac{t^2 \frac{n_0}{n+n_0} + \nu}{t^2 + \nu}\right)^{(\nu+1)/2},$$

where

$$t = \frac{|\bar{D}|}{s/\sqrt{n}} \text{ (the usual } t\text{-statistics)}, \qquad \nu = n - 1 \text{ (the usual degree of freedom of the sample)}$$

The Bayes factor here depends on the parameter n_0. Recall that n_0 is the effective sample size of the prior distribution of μ. To get the least information from our prior, we set $n_0 = 1$. Using the dataset given by the zinc water example, we will get

$$\text{BF}[H_1 : H_2] \approx 0.0154, \qquad \text{which implies} \qquad \text{BF}[H_2 : H_1] = \frac{1}{\text{BF}[H_1 : H_2]} \approx 64.9351.$$

This provides strong evidence **against** H_1.

Again, assuming **equal prior odd**, we can calculate the posterior probability of H_1 as

$$\mathbb{P}(H_1 \mid \text{data}) = \frac{\text{BF}[H_1 : H_2]}{\text{BF}[H_1 : H_2] + 1} = \frac{0.0154}{0.0154 + 1} \approx 0.0152$$

3.4 What To Report

We have obtained the Bayes factor, which implies strong evidence to support $H_2 : \mu_{\text{diff}} \neq 0$. However, we would like to ask questions such as

- What would be a good estimate of μ_{diff} if it was not 0?
- What would be the most likely range that μ_{diff} would fall into?

Point Estimate

The first question relates to the point estimate of μ_{diff}. We have discussed point estimate based on the given loss function in the first few videos of decision making. Here, we can also apply the same idea to pick the point estimate of the one that gives the smallest loss. We will not go into each individual situation, but rather, we only present an example when we choose the mean of μ_{diff} as its point estimate. (Point estimates do not have to be unique. This requires expertise towards specific questions.)

As we mentioned earlier, the updated posterior joint distribution of μ and ϕ (or σ^2) gives a Student t-distribution as the marginal distribution of μ, with degree of freedom ν_n, center at m_n, and variance σ_n^2/n_n. Recall from the previous video, we picked the non-informative reference prior (which was introduced later in this week) such that $\nu_0 = 0$. And we also assume under H_2, $m_0 = 0$. Therefore, the marginal distribution of μ under H_2 satisfies

$$\pi^*(\mu_{\text{diff}} \mid \text{data},\ H_2) \sim t_{\nu_n}\left(m_n, \frac{\sigma_n^2}{n_n}\right) = t_{\nu_n}\left(\frac{n}{n+n_0}\bar{D}, \frac{\sigma_n^2}{n_n}\right).$$

(Here we omit the details of ν_n, σ_n, and n_n since we focus on the mean of μ_{diff}.)

The Student's t-distribution is a symmetric unimodal distribution. Therefore, the mean of μ_{diff} under H_2 is

$$\mathbb{E}(\mu_{\text{diff}} \mid H_2,\ \text{data}) = \frac{n}{n+n_0}\bar{D} = \frac{10}{10+1}(0.0804) \approx 0.0731.$$

(We have already set $n_0 = 1$ in the last video to restrict information from the prior of μ.)
Since μ_{diff} is always 0 under H_1, we have $\mathbb{E}(\mu_{\text{diff}} \mid H_1,\ \text{data}) = 0$.

From the law of total expectation (a variation of conditional probability), we have the total expected value (or mean) of μ_{diff} is

$$\mathbb{E}(\mu_{\text{diff}} \mid \text{data}) = \mathbb{E}(\mu_{\text{diff}} \mid H_1,\ \text{data})\mathbb{P}(H_1 \mid \text{data}) + \mathbb{E}(\mu_{\text{diff}} \mid H_2,\ \text{data})\mathbb{P}(H_2 \mid \text{data}) = (0)(0.0152) + (0.0731)(0.9848) \approx 0.072.$$

(We keep "data" for each term to imply that all quantities are obtained using the posterior distribution/probability.)

Credible Interval

The point estimate would not be enough to argue that μ_{diff} is very likely to be away from 0. Hence we turn our attention to the credible interval. To get the credible interval, we need to consider the posterior probability obtained under both H_1 and H_2.

Under the assumption of equal odd, i.e., the prior probability $\mathbb{P}(H_1) = \mathbb{P}(H_2)$, we have obtained the posterior probabilities

$$\mathbb{P}(H_1 \mid \text{data}) = 0.0152, \qquad \mathbb{P}(H_2 \mid \text{data}) = 0.9848$$

When looking for the 95% credible interval, the probability under H_1 is surely not enough. So should we include H_1 or just focus on H_2? It depends on if we use the probability distribution of μ_{diff} obtained under H_2, whether the credible interval will cover 0, the value of μ_{diff} under H_1.

Therefore, we turn to the posterior distribution of μ_{diff} under H_2. The posterior distribution is a symmetric unimodal t-distribution, with the largest amount of area centered around the mean. If we were to look for the **highest posterior density interval**, we were surly focus on the center region. Moreover, since the distribution is symmetric, we can argue that the highest posterior density interval is the same as the equal-tailed interval. Since the total area (probability) under H_2 is only 0.9848, to get the true 95% credible interval, we need to rescale this posterior distribution

$$\text{new posterior distribution } = (\text{old } t\text{-distribution}) \times \mathbb{P}(H_2 \mid \text{data}),$$

so that the new posterior distribution only covers area of $0.9848 = \mathbb{P}(H_2 \mid \text{data})$. We have found that the interval $[0.0276, 0.1187]$ gives the desired probability and it does not include 0. Therefore, we are not including the probability obtained under H_1. However, since credible intervals are not unique (different from confidence interval), we need subject expertise to decide whether this interval is meaningful to the problem.

Posterior Probabilities of Directional Hypotheses

The posterior probability of H_2 only tells us the probability of $\mu_{\text{diff}} \neq 0$. We can also use this probability to obtain posterior probability for the more detailed hypotheses such as

$$H_3 : \mu_{\text{diff}} > 0, \qquad \text{and} \qquad H_4 : \mu_{\text{diff}} < 0$$

Since H_3 discusses a smaller subset of the event in H_2, we may obtain the conditional posterior probability of H_3 by

$$\mathbb{P}(H_3 \mid H_2,\ \text{data}) = \int_0^\infty t_{\nu_n}\left(\mu;\ \frac{n}{n+n_0}\bar{D}, \frac{\sigma_n^2}{n_n}\right)\, d\mu \approx 0.9984.$$

To obtain the unconditional probability, we use the property of conditional probability and get

$$\mathbb{P}(H_3 \mid \text{data}) = \mathbb{P}(H_3 \mid H_2,\ \text{data}) \times \mathbb{P}(H_2 \mid \text{data}) \approx 0.9832$$

To get the posterior probability of H_4 is similar. Since H_3 and H_4 are complementary events under H_2, we can easily get

$$\mathbb{P}(H_4 \mid \text{data}) = \mathbb{P}(H_2 \mid \text{data}) - \mathbb{P}(H_3 \mid \text{data}) \approx= 0.001597.$$

R Codes

The functions related to the t-distribution in R are `dt, qt, pt`, which gives the density, the quantile (for credible intervals), and the probability (for posterior probabilities) respectively.