Week 2 Review and Examples

Duke University

R Functions for Distributions

R has all the distributions we use in this course, say `dbinom` calculates the probability of Binomial distribution at one value, while `pbinom` gives the accumulated probability of the Binomial distribution up to some value. To know more, simply type `?binom` in the RStudio Console and R will give you all the functions in R related to the Binomial distribution. This works the same as other distributions.

Moreover, in the Course **Resources** tab, we have summarized several useful websites, cheat sheets for R resources. They are under **Supplementary R Resources**.

Review

1 Discrete Distribution vs Continuous Distribution

1.1 Discrete Case

In week 1, we have introduced several examples on using Bayes' Rule to calculate posterior probability. We have the following relationship between the 3 "probabilities":

$$\underbrace{\mathbb{P}(\text{hypothesis} \mid \text{data})}_{\text{posterior probability of hypothesis}} \quad \propto \quad \underbrace{\mathbb{P}(\text{data} \mid \text{hypothesis})}_{\text{likelihood of data}} \quad \times \quad \underbrace{\mathbb{P}(\text{hypothesis})}_{\text{prior probability of hypothesis}}.$$

We can be more precise if our hypothesis is about the parameter of the data, say θ, and our data can be represented using a variable X. Then the above Bayes' Rule formula can be summarized as:

$$\underbrace{\mathbb{P}(\theta \mid X)}_{\text{posterior probability of }\theta} \quad \propto \quad \underbrace{\mathbb{P}(X \mid \theta)}_{\text{likelihood of }X} \quad \times \quad \underbrace{\mathbb{P}(\theta)}_{\text{prior probability of }\theta}.$$

Here $\propto$ means **proportional to**, which says the left-hand side is a scale multiple of the right-hand side. That is, left-hand side = some constant × right-hand side.

The above formulas only work when the θ and X takes discrete values. For example, suppose we have 1 coin, we do not know whether it is fair or not. We make assumptions that the chance of getting a head from this coin can either be 0.5 (fair) or 0.8 (biased towards heads). Then θ only takes values 0.5 and 0.8. We now start flipping the coin to update our beliefs. X represents the number of heads we get from all flips. Of course, X only also takes discrete values.

In Bayesian Statistics, our first focus is on the probability of θ, not the probability or likelihood when observed data happen. We can easily assign probability of the data according to the type of event happened. If an event is compose of steps that give binary results each time, it is likely to be a Binomial Process, which means X should follow the Binomial distribution. If an event involves counting the total occurrence within a period, X is likely to follow the Poisson distribution. However, we can never observe the parameter θ, and this is a course we use Bayes' Rule to update our beliefs or initial guesses of θ.

Therefore, when we discuss the difference between **discrete** and **continuous**, we are discussing whether the parameter θ takes discrete values or continuous values.

When θ takes discrete values:

We use the notation $\mathbb{P}(\theta = \text{some value})$ or simply $\mathbb{P}(\theta)$ to represent the probability of θ. This is called the **Probability Mass Function (pmf)** or the probability distribution of θ. Accordingly, when the data X also takes discrete values, we also have notations like $\mathbb{P}(X = \text{some value})$, $\mathbb{P}(X)$, $\mathbb{P}(X|\text{some condition})$ to denote the probability of X.

Examples of discrete distributions used in this course:

Binomial distribution (k successes out of n trials, with success rate p) : $\mathbb{P}(X = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$,

mean: np, variance: $np(1-p)$,

Poisson distribution (k counts occur when the mean count in general is λ) : $\mathbb{P}(X = k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$,

mean: λ, variance: λ,

Any other distribution you can impose for the parameter, such as : $\mathbb{P}(\theta = 0.5) = \frac{1}{3}$, $\mathbb{P}(\theta = 0.8) = \frac{2}{3}$.

1.2 Continuous Case

While both parameters and the data can take continuous values, in Week 2, we mainly focus on when parameters take continuous values and follow some continuous probability distribution (probability distribution function (pdf)).

Let us review the **definitions** of the 3 “ingredients” for Bayes' Rule again:

Prior Distribution: The probability distribution of the **parameter** θ. Often this distribution depends on other new parameters, which we call **hyperparameters** say α, β. We denote the function as $\pi(\theta)$ or $\pi(\theta; \alpha, \beta)$.

Likelihood: If X is discrete, this is the probability of **data/result** we observe given paramter(s). If X is continuous, this becomes the probability distribution function of X, under given parameter(s). We denote this as $\mathbb{P}(X|\theta)$ (discrete) or $f(X|\theta)$ (continous).

Posterior Probability: The probability distribution of the **parameter** θ, updated after we have obtained the data. It is denoted as $\pi^*(\theta|X)$.

Examples of continuous distributions used in this course:

Note: The following formulas of the distributions are totally for your curiosity. Most of the time, you only need to memorize the form, the graph, and how the hyperparameters relate to statistics such as mean, variance, or standard deviation.

Beta distribution (with hyperparameters α, β) : $\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \propto p^{\alpha-1}(1-p)^{\beta-1}$,

mean: $\frac{\alpha}{\alpha+\beta}$, variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$(not required by course),

Gamma distribution (with hyperparameters k, θ) : $\pi(\lambda) = \frac{1}{\Gamma(k)\theta^k}\lambda^{k-1}e^{-\lambda/\theta} \propto \lambda^{k-1}e^{-\lambda/\theta}$

mean: $k\theta$, variance: $k\theta^2$, st dev: $\theta\sqrt{k}$,

Normal distribution (with hyperparameters ν, τ^2) : $\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-(\mu-\nu)^2/(2\tau^2)} \propto e^{-(\mu-\nu)^2/(2\sigma^2)}$

mean: ν, variance: τ^2, st dev: τ,

($\Gamma(\alpha)$ is called the gamma function, which has a complicated integral form. This function takes any positive number to turn it into another positive number. When α is an integer will $\Gamma(\alpha)$ take simple form. Feel free to look up the formula, but this is not required to know in this course.)

2 Continuous Bayes' Rule

$$X \text{ is discrete:} \quad \underbrace{\pi^*(\theta \mid X)}_{\text{posterior distribution of } \theta} = \frac{\mathbb{P}(X \mid \theta) \times \pi(\theta)}{\displaystyle\int \mathbb{P}(X \mid \theta) \times \pi(\theta)\, d\theta} \propto \underbrace{\mathbb{P}(X \mid \theta)}_{\text{likelihood/probability of } X} \times \underbrace{\pi(\theta).}_{\text{prior distribution of } \theta}$$

$$X \text{ is continous:} \quad \underbrace{\pi^*(\theta \mid X)}_{\text{posterior distribution of } \theta} = \frac{f(X \mid \theta) \times \pi(\theta)}{\displaystyle\int f(X \mid \theta) \times \pi(\theta)\, d\theta} \propto \underbrace{f(X \mid \theta)}_{\text{distribution of } X} \times \underbrace{\pi(\theta).}_{\text{prior distribution of } \theta}$$

We use "proportional to" because the denominators in the above 2 formulas will always give us some constants.

3 Example

In the following, we will work on a coin example by hand to illustrate the updating of posterior distribution when the parameter of getting heads takes continuous values:

> Suppose you have a coin without knowing whether it is biased or not. You have decided to place a uniform distribution between 0 and 1 ($\mathcal{U}([0,1])$) for the parameter p, the chance we will get a head each time we toss the coin. You toss the coin 3 times and get 1 head and 2 tails. What is the posterior distribution of p, the chance a coin toss will give a head?

3.1 Method 1: Use Bayes' Rule

This example is basically the same as the one we used in Week 1 Review file, except that this time, the parameter p can take continuous values between 0 and 1. Therefore, we will use the continuous version of Bayes' Rule to get the posterior.

We choose to view the 3 tosses as a whole. Can we update the posterior step by step? The answer is yes, but it will be too time consuming.

Our ingredients:

- Prior distribution: $\pi(p) = \mathcal{U}([0,1]) = \text{Beta}(1,1) = \dfrac{\Gamma(2)}{\Gamma(1)\Gamma(1)} p^{1-1}(1-p)^{1-1} = 1,$
- Likelihood (Binomial process with 1 head in 3 tosses): $\mathbb{P}(X|p) = \dbinom{3}{1} p^1(1-p)^{3-1} = 3p(1-p)^2.$

Our recipes:

- Bayes' Rule: $\pi^*(p|X) = \dfrac{\mathbb{P}(X|p) \times \pi(p)}{\displaystyle\int_0^1 \mathbb{P}(X|p) \times \pi(p)\, dp} = \dfrac{[3p(1-p)^2] \times [1]}{\displaystyle\int_0^1 [3p(1-p)^2] \times [1]\, dp} = \dfrac{3p(1-p)^2}{(1/4)} = 12p(1-p)^2$

3.2 Method 2: Use Conjugacy

Note: In this course, we only require you to know how to use conjugate families to solve problems. Non-conjugacy is optional and if you are interested, you may refer to the Non-Conjugacy file for examples.

Since Bayes' Rule will require us to compute the denomination, which is an integral, it will be way more convenient if we can leverage conjugacy. Luckily, the example we have here follows the Beta-Binomial Conjugacy, so we can simply use

conjugacy to get the posterior distribution of p.

According to the Beta-Binomial Conjugacy, if we assume the success rate p follows the Beta distribution

$$\text{Beta}(\alpha, \beta)$$

as our prior, and the data follows a Binomial distribution with k successes in the total n trials, the posterior distribution of p is still a Beta distribution with new α^* and new β^*:

$$\text{Beta}(\alpha^* = \alpha + k, \beta^* = \beta + (n - k))$$

As we know, the prior distribution of p in this example, is the uniform distribution between 0 and 1, $\mathcal{U}([0, 1])$, which is a special Beta distribution Beta(1, 1). Under this example, we have 1 head (1 success) and 2 tails (2 failures) in 3 tosses (3 trials). Therefore, using the **Beta-Binomial Conjugacy**, we conclude that the posterior distribution of p must be

$$\text{Beta}(1 + 1, 1 + (3 - 1)) = \text{Beta}(2, 3) = \frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} p^{2-1}(1 - p)^{3-1} = 12p(1 - p)^2$$

Our result matches the one in Method 1!

4 Major Takeaway: Leverage Conjugacy Families

4.1 Rule of Thumb

If the prior distribution of the parameter and the distribution of the data form a **conjugate** family, the posterior distribution of the parameter will share the **same form** (or in the **same family**) of the prior distribution, and we simply **update the hyperparameters of the prior** to get the posterior distribution.

Only when we do not have conjugacy, we would consider using the continuous version of Bayes' Rule.

4.2 Conjugate Families in This Course

- **Beta-Binomial Conjugacy:** When Beta(α, β) meets Binomial distribution with k successes in total n trials

$$\text{Beta}(\alpha, \beta) \longrightarrow \text{Beta}(\alpha^* = \alpha + k,\ \beta^* = \beta + (n - k))$$

- **Gamma-Poisson Conjugacy:** When Gamma(k, θ) meets Poisson distribution with counts $x_1, x_2, \cdots, x_n$ in n periods

$$\text{Gamma}(k, \theta) \longrightarrow \text{Gamma}\left(k^* = k + \sum_{i=1}^{n} x_i,\ \theta^* = \frac{\theta}{n\theta + 1}\right)$$

Note: In Week 2 Lab, we use a different definition of Gamma distribution:

$$\text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

So the conjugacy rule will be

$$\text{Gamma}(\alpha, \beta) \longrightarrow \text{Gamma}(\alpha^* = \alpha + \sum_{i=1}^{n} x_i,\ \beta^* = n + \beta)$$

- **Normal-Normal Conjugacy (with known σ of the data):** When Normal(ν, τ^2) meets normal distribution of data $\{x_1, x_2, \cdots, x_n\}$ with **known** standard deviation σ

$$\text{Normal}(\nu, \tau^2) \longrightarrow \text{Normal}\left(\nu^* = \frac{\nu\sigma^2 + (\sum_{i=1}^{n} x_i)\tau^2}{\sigma^2 + n\tau^2},\ (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

(here I follow the usual convention of normal distributions and use the variance instead of the standard deviation as the hyperparameter.)

5 Posterior Probability of Parameter

Let us continue with the previous coin toss example.

> Suppose you have a coin without knowing whether it is biased or not. You have decided to place a uniform distribution between 0 and 1 for the parameter p, the chance we will get a head each time we toss the coin. You toss the coin 3 times and get 1 head and 2 tails. What is the posterior distribution of p, the chance a coin toss will give a head?

We have just solved the above example using 2 methods, and have obtained the posterior distribution of p to be the Beta distribution $\text{Beta}(2,3) = 12p(1-p)^2$. Now we can move on to more analysis of p. For example,

> **Question**: What is the posterior probability that p is smaller than 0.5, i.e, $\mathbb{P}(p \leq 0.5)$?

Reminder: Before updating, the probability of p less than 0.5 is 0.5.

Since we have observed just 1 head out of 3 tosses, it is very likely that we are concern about that p being smaller than 0.5. We have already updated the distribution of p, to calculate the probability, we need to get the **area** under the distribution of p. Mathematically speaking, we compute:

$$\mathbb{P}(p \leq 0.5) = \int_0^{0.5} \text{Beta}(2,3)\, dp = \int_0^{0.5} 12p(1-p)^2\, dp = \frac{11}{16} = 0.6875$$

R Codes

Since we know the posterior distribution is also a Beta distribution, You can calculate this probability using R without computing the integral:

```
> pbeta(0.5, shape1 = 2, shape2 = 3)  # probability of Beta up to 0.5, with alpha = 2 beta = 3
[1] 0.6875
```

Compared with the probability of p being less than 0.5, and the probability after we have updated the distribution based on the data, we can see that we have a strong belief that p, the chance of getting heads, should be more likely to be less than 0.5.

6 Credible Interval

Credible interval is a range for the values of the **parameter**, not the data. That said, we need to first obtain the posterior distribution of the parameter θ, then we can find such an interval. This is very different from the frequentist approach.

6.1 Definition

That

> The 95% credible interval of a parameter θ is $[L, U]$

means

> Given the observed data, there is 95% probability that θ is inside $[L, U]$

6.2 How to Obtain Credible Interval

Credible intervals are not unique. There are several ways to define a credible interval with a given percentage.

- The **major** method we use in this course is to find the **shortest** interval of θ, such that the area under the curve of the posterior distribution of θ within this interval is this given percentage. This interval sometimes is also called the **highest posterior interval**.
- Another method you will see in this course is called the **equal-tailed credible interval**. Such interval is easy to obtain using quantiles.

Example: RU-486 Posterior Example

Shortest Credible Interval

In course video **Credible Intervals**, we have discussed how to obtain the 95% credible interval after we have obtained the posterior distribution of p, the chance of getting pregnant after taking RU-486. The posterior distribution of p is

$$\pi^*(p \mid \text{data}) = \text{Beta}(1,5) = 5(1-p)^4$$

We hope to find two points L and U so that the area under $\pi^*(p)$ is 95% and the length of the interval $[L, U]$ is the shortest.

This means, we require

$$\int_L^U \pi^*(p \mid \text{data})\, dp = \int_L^U 5(1-p)^4\, dp = 0.95 \qquad \text{(This is equivalent to } F(U) - F(L) = 0.95 \text{ in the video.)}$$

and find the shortest interval of all such $[L, U]$'s.

This seems hard to solve at the first glance. But we can leverage some nice property of $\pi^*(p \mid \text{data}) = 5(1-p)^4$ to argue where the shortest interval should occur. Its graph has been shown in the video. As we can see, this posterior distribution is a decreasing function. Therefore, the closer the interval is to 0, the more area we can get with a fixed range. Equivalently, the closer the interval is to 0, we only need a shorter range to get the desired area (95%).

This is why we set $L = 0$ in the video. Solving for U using $0.95 = \int_L^U 5(1-p)^4\, dp = \int_0^U 5(1-p)^4\, dp$, we get $U = 1 - (0.05)^{1/5} \approx 0.451$. Therefore, the **shortest credible interval** is

$$0 \leq p \leq 0.451$$

Equal-Tailed Credible Interval

Another interval we will use a lot in the labs is the equal-tailed interval. In this case, we simply hope to get L and U so that the area within $[0, L]$ is the same as the area within $[U, 1]$.

That said, if we want to get the 95% credible interval, the area within $[0, L]$ should be 0.025, the same as the area within $[U, 1]$. This means L is the 2.5% quantile of this posterior distribution, and U is the 97.5% quantile. Using R, we can easily obtain L and U.

```
# Compute the quantiles of Beta(1,5)
> qbeta(c(0.025, 0.975), shape1 = 1, shape2 = 5)
[1] 0.005050763 0.521823750
```

Therefore, the **equal-tailed credible interval** is

$$0.005 \leq p \leq 0.522$$

this interval is different from the shortest credible interval.

7 Different Beliefs (Different Prior Hyperparameters)

We have mentioned in Week 2's videos, that the hyperparameters α and β represent the level of our belief of the distribution of parameter θ. In the coin toss example, we places the Beta(1, 1) prior for p, which is a relatively weak prior, because the effective sample size is small. (For effective sample size, please refer to Resources of this course.) Therefore, after 3 tosses with just 1 head, we have significantly shifted our belief of p, weighing more probability for p being less than 0.5.

What if we have a much stronger belief that p is close to 0.5?

7.1 Use Beta(100, 100) as Prior

Suppose we assign the prior distribution of p to be Beta(100, 100). This prior still maintain the same ratio between α and β, the same mean $\dfrac{\alpha}{\alpha+\beta} = \dfrac{100}{200} = \dfrac{1}{2}$. But this prior has much larger **effective sample size**. Intuitively, you may think of this prior reflects that you have previously tossed 200 times and get 100 heads with this coin. After observing 1 head in 3 tosses, we update the distribution of p to be Beta(100+1, 100+2) = Beta(101, 102).

> **Same question**: What is the posterior probability of p being less than 0.5?

Using R, we get

```
> pbeta(0.5, shape1 = 101, shape2 = 102) # probability up to 0.5, with alpha = 101 beta = 102
[1] 0.5280348
```

The number is still very close to 0.5, the prior probability of p. This means, we are "not that convinced" the data from 3 tosses can easily change our belief of the distribution of p.

8 Predictive Probability of Data Using Conditional Probability and Posterior

8.1 Discrete Case

We use the Week 1 example to illustrate the idea.

> Suppose you have a coin, which may be biased (0.8 towards heads) or unbiased (50-50 chance). Since you do not have any previous information of this coin, you have decided that the prior probabilities of p to be 0.8 and 0.5 are both 1/2. You flip the coin 3 times, with 1 head and 2 tails. Now the question has changed:
>
> **New question**: What would the probability be when you flip this coin 2 more times and get 1 head and 1 tail?

Recall that the posterior probabilities we have calculated for p (please refer to Week 1 Review file) are

$$\text{new } \mathbb{P}(p = 0.8) = \mathbb{P}(\text{biased} \mid \text{data}) = \mathbb{P}(p = 0.8 \mid \text{data}) = \frac{32}{157}$$

Then

$$\text{new } \mathbb{P}(p = 0.5) = \mathbb{P}(\text{unbiased} \mid \text{data}) = \mathbb{P}(p = 0.5 \mid \text{data}) = 1 - \mathbb{P}(\text{biased} \mid \text{data}) = \frac{125}{157}$$

(I keep fractions to ensure accuracy. You can use R to help you with this part.)

Conditional Probability

Since we are actually not 100% sure that p must be 0.5, instead of 0.8, even we see a much smaller chance for p to be 0.8. Therefore, we need to take the 2 situations into account when we compute the predictive probability of the data. Using conditional probability, we have

$$\begin{aligned}\mathbb{P}(\text{2 flips with 1 head}) =& \mathbb{P}(\text{2 flips with 1 head} \mid p = 0.8)\mathbb{P}(p = 0.8) + \mathbb{P}(\text{2 flips with 1 head} \mid p = 0.5)\mathbb{P}(p = 0.5) \\ =& \left[\binom{2}{1}(0.8)^1(1-0.8)^1\right]\left[\frac{32}{157}\right] + \left[\binom{2}{1}(0.5)^1(1-0.5)^1\right]\left[\frac{125}{157}\right] = \frac{3637}{7850} \approx 0.4633121\end{aligned}$$

R Codes

```
> priors <- c(1/2, 1/2)                                       # Original priors
> likelihoods <- c(dbinom(1, 3, 0.8), dbinom(1, 3, 0.5))      # 3 flips with 1 head
> posteriors <- priors * likelihoods / sum(priors * likelihoods) # Bayes' Rule

> # Conditional probabilities for the new 2 flips with 1 head
> condProbs <- c(dbinom(1, 2, 0.8), dbinom(1, 2, 0.5))
> # Calculate total probability
> probability <- sum(condProbs * posteriors)

> probability
[1] 0.4633121
```

8.2 Continuous Case (Optional)

We use this Week's example to illustrate the idea when the distribution of the parameter is continuous.

> Suppose you have a coin without knowing whether it is biased or not. You have decided to place a uniform distribution between 0 and 1 ($\mathcal{U}([0,1])$) for the parameter p, the chance we will get a head each time we toss the coin. You toss the coin 3 times and get 1 head and 2 tails.
>
> **New question:** What would the probability be when you toss this coin 2 more times and get 1 head and 1 tail?

We have calculated the posterior probability of p, which is $\pi^*(p) = \text{Beta}(2,3) = 12p(1-p)^2$. To calculate the predictive probability of the new data (2 toss with 1 head), we use the **continuous conditional probability** formula

$$\mathbb{P}(\text{2 tosses with 1 head}) = \int_0^1 \mathbb{P}(\text{2 tosses with 1 head}|\ p) \times \pi^*(p)\, dp = \int_0^1 \left[\binom{2}{1} p^1(1-p)^1\right] \times [12p(1-p)^2]\, dp = 0.4$$

(To calculate the last integral, you may do integration by parts twice, or use the property that the function inside the integral is $24p^2(1-p)^3 = (0.4) \times \text{Beta}(3,4)$.)

When the distributions of parameters and the distributions of data are getting more complicated, there seldom exist nice forms for the integration. We usually would like to use R to do discrete simulation (like the one in the discrete case) to estimate the result.