Week 2 Supplementary Materials

Lizzy Huang
Duke University

Continuous Variables and Eliciting Probability Distributions

From the Discrete to the Continuous

In the previous week, we already showed how to use Bayes' Rule to calculate posterior probability and to compare two competing hypotheses. Recall the example we used in the Video **Frequentist vs. Bayesian Inference** in Week 1, we set up two competing hypotheses about the proportion p of yellow M&M's:

$$H_1 : p = 0.1, \qquad H_2 : p = 0.2.$$

We calcualted their posterior probabilities $P(H_1 \mid \text{data}) = P(p = 0.1 \mid \text{data})$ and $P(H_2 \mid \text{data}) = P(p = 0.2 \mid \text{data})$ using Bayes' Rule.

In this example, we only compared the posterior probabilities of two different values of the parameter p. However, more often we would not have only 2 parameter values to compare with, instead, we would have the following question:

$$H_1 : p = 0.1, \qquad H_2 : p \neq 0.1.$$

For the first hypothesis, we have already seen how to get its posterior probability using Bayes' Rule. But for the second hypothesis H_2, we need to consider a continuum of values between 0 and 1, excluding 0.1. The current Bayes' Rule only works when the parameter takes a discrete set of values. Therefore, we need to develop a continuous version of Bayes' Rule, which will enable us to calculate posterior probability such as

$$P(p \neq 0.1 \mid \text{data}).$$

Continuous Version of Bayes' Rule

In the lectures, we have already introduced the continuous counterpart of probability mass function (pmf), the probability density function (pdf). Unlike pmf's that give us the probability directly, pdf's only give us the "density" of the probability. To calculate the probability, we need the cumulative distribution function $F(x)$, which is defined as

$$F(x) = \int_{-\infty}^{x} p(t)\, dt,$$

for its corresponding pdf $p(t)$.[1] Recall that the graphical meaning of the above integral is the area under the curve $p(t)$ from the left up to a given value x, this integral can also be intepreted as $P(X \leq x)$. So for the area under the pdf within an interval (L, U), which is the probability $P(L < X < U)$, we can calculate it using

$$F(U) - F(L) = P(L < X < U) = \int_{L}^{U} p(t)\, dt.$$

In Week 1, we have introduced the Bayes' Rule, which look like:

$$P(A_i \mid \text{data}) = \frac{P(\text{data} \mid A_i) \times P(A_i)}{\sum_{j=1}^{n} P(\text{data} \mid A_j) \times P(A_j)}. \tag{1}$$

[1] The reason that we use t as the independent variable of the pdf is because we do not want to confuse t with the given value x.

Here A_i represents a specific event such as, a US military recruit has HIV, and the entire sample sample can be partitioned by all these A_j's. "data", instead, if the observed data we receive, such as, a US military recruite is tested ELISA positive. For example, if we consider the two hypotheses $H_1 : p = 0.1$ and $H_2 : p = 0.2$ for the proportion of yellow M&M's in a bag of 5, and if we believe that *the only possible values for p are 0.1 and 0.2*, when we are interested in the posterior probability of H_1, Formula (1) can be applied as

$$P(p = 0.1 \mid \text{data}) = \frac{P(\text{data} \mid p = 0.1) \times P(p = 0.1)}{P(\text{data} \mid p = 0.1) \times P(p = 0.1) + P(\text{data} \mid p = 0.2) \times P(p = 0.2)}.$$

Since the summation can only include finitely many events, that means, Formula (1) only works with pmf's and cannot be applied in the situation when we believe the parameter can take continuously many values.

Here, we need to bring in the continuous version of Bayes' Rule that works with pdf's[2]:

When only the parameter θ is continuous

$$p(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta) \times p(\theta)}{\displaystyle\int_{-\infty}^{\infty} P(\text{data} \mid \theta) \times p(\theta)\, d\theta}. \tag{2}$$

When both the data (represented by the random variable X) and the parameter θ are continuous

$$p(\theta \mid x) = \frac{p(x \mid \theta) \times p(\theta)}{\displaystyle\int_{-\infty}^{\infty} p(x \mid \theta) \times p(\theta)\, d\theta}. \tag{3}$$

You may notice that, the summation has been replaced as the integral, which makes the calculation more complicated. In fact, most of the time we do not bother to "touch" the integral in the denominator, since we know it will give us just a constant. Therefore, sometimes we will write the Bayes' Rule as

$$p(\theta \mid x) \propto p(x \mid \theta) \times p(\theta),$$

where $\propto$ means "proportional to", that is, the left-hand side is a scalar multiple of the right-hand side. This format not only simplies the writing, but also provides more insight of the form that the posterior distribution $p(\theta \mid x)$. We will see that in the later discussion.

Three Conjugate Families

Beta-Binomial Conjugate Family

In the Binomial distribution, there are two parameters, n, the sample size; and p, the probability of success. We are often given the sample size when we have the data, but we seldom know exactly what p is. Therefore, we want to infer information about p, that is, we want to know the distribution of p after seeing the data, that is, the posterior distribution $\pi(p \mid \text{data}) = \pi^*(p)$[3]. We can use the Bayes' Rule (**??**) to calculate $\pi^*(p)$, since the data comes from a discrete distribution.

It turns out that, one good candidate, the family of Beta distribution, will give us a convenient updating scheme. That is:

[2] To use the unified formulas below, we assume that all pdf's are defined on $(-\infty, \infty)$. If a parameter or a random variable X only take a finite range of continuous values, we can always extend their pdf so that the pdf is 0 on the values that they do not take.

[3] we use $\pi(p)$ to avoid notation confusion. Usually we use an $*$ superscript to denote the posterior distribution.

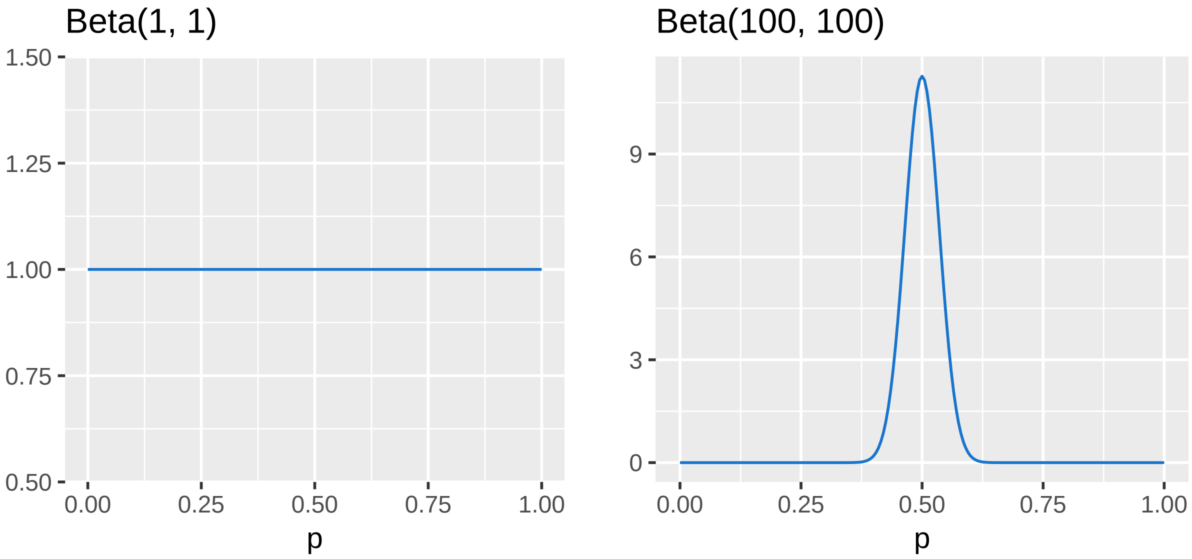
> If the prior distribution of the parameter p, the probability of success, is a Beta distribution, and the data follows a Binomial distribution, then the posterior distribution of p is still a Beta distribution.

The family of Beta distribution is defined to be

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

The two parameters α and β are the shape parameters, that different sets of values of α and β give different Beta distributions. Since they are the parameters of the parameter of interest p, we also call them the hyperparameters. The Uniform distribution $\mathsf{Unif}(0,1)$ is a special Beta distribution with $\alpha=1$ and $\beta=1$[4]. Since we usually care more about what values of α and β we choose for the Beta distribution, we often suppress the independent variable p and denote the Beta distribution as $\mathsf{Beta}(\alpha,\beta)$.

The mean of $\mathsf{Beta}(\alpha,\beta)$ is $\dfrac{\alpha}{\alpha+\beta}$ and the variance is $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. $\alpha+\beta$ is the prior sample size of a Beta distribution. The choice of α and β depends on how much belief you have about the parameter p. Suppose p is the probability of coming head in a coin toss. If you are not sure about the value of p, you may choose the uniformly flat Beta distribution as its prior, i.e., the Uniform distribution $\mathsf{Beta}(1,1)$. But if you are very sure that p is close to 0.5, then you might choose $\mathsf{Beta}(100,100)$, which concentrates most of its density at $p=0.5$. The figure below shows the two Beta distributions. Notice that these two distributions have the same mean $\dfrac{1}{2}$, but their prior sample sizes are different.



We can formulate the updating rule as

$$\mathsf{Beta}(\alpha,\beta)\xrightarrow{\text{Binomial, } k \text{ successes in } n \text{ trials}}\mathsf{Beta}(\alpha^*=\alpha+k,\ \beta^*=\beta+n-k),$$

if there are k successes with n trials in the data.

Example 1

> In the RU-486 example, we have 4 pregnancy cases, with 0 success, since no pregnancy comes from taking RU-486. Let p be the probability of success, or, the probability that a pregnancy

[4] Beta distribution shares a very similar form as the Binomial distribution, however, it is not the Binomial distribution. In fact, Beta distributions can be conjugate priors for not only the Binomial distribution, but also the geometric distribution.

case comes from RU-486. Suppose the prior of p, $\pi(p)$, is the uniformly flat Beta distribution $\mathsf{Beta}(1, 1)$. What is the posterior distribution of p after seeing the data?

Method 1: Use Bayes' Rule Directly

Let us solve this problem using the Bayes' Rule (2). We have the prior $\pi(p) = 1$ (the uniform distribution over (0, 1)), and the likelihood

$$P(X = 0 \mid p) = \binom{4}{0} p^0 (1-p)^{4-0} = (1-p)^4.$$

By (2), the posterior distribution is

$$\pi(p \mid X = 0) = \pi^*(p) = \frac{P(X = 0 \mid p) \times \pi(p)}{\displaystyle\int_0^1 P(X = 0 \mid p) \times \pi(p)\, dp} = \frac{(1-p)^4}{\displaystyle\int_0^1 (1-p)^4\, dp}.$$

The integral of the denominator can be explicitly calculated as

$$\int_0^1 (1-p)^4\, dp = \left[-\frac{1}{5}(1-p)^5 \right]_0^1 = \frac{1}{5}.$$

Therefore, the posterior distribution is

$$\pi^*(p) = 5(1-p)^4.$$

This method only works when the integral in the demonimator is relatively simple. However, most of the time, we will encounter complicated integral, so computing posterior distributions by hand is not ideal.

Method 2: Use Conjugate Property between Beta and Binomial

Using the Beta-Binomial conjugate family, we can easily obtain the posterior distribution to be

$$\pi^*(p) = \mathsf{Beta}(\alpha^* = 1 + 0,\ \beta^* = 1 + (4 - 0)) = \mathsf{Beta}(1, 5) = 5(1-p)^4.$$

We see that, using the conjugate property, we end up getting the same result.

The mean and standard deviation of p before and after seeing the data can be summarized as

$$\begin{array}{lll} \text{Before:} & \text{mean of } p = \dfrac{\alpha}{\alpha+\beta} = \dfrac{1}{2} & \text{standard deviation of } p = \sqrt{\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} =\approx 0.289 \\ \text{After:} & \text{mean of } p = \dfrac{\alpha^*}{\alpha^*+\beta^*} = \dfrac{1}{6} & \text{standard deviation of } p = \sqrt{\dfrac{\alpha^*\beta^*}{(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)}} \approx 0.141 \end{array}$$

Now that we have the posterior distribution of a p given the data, we can calculate the probability of p lying in a range. For example, suppose we want to compare the two hypotheses

$$H_1 : p = 0.5, \qquad H_2 : p < 0.5.$$

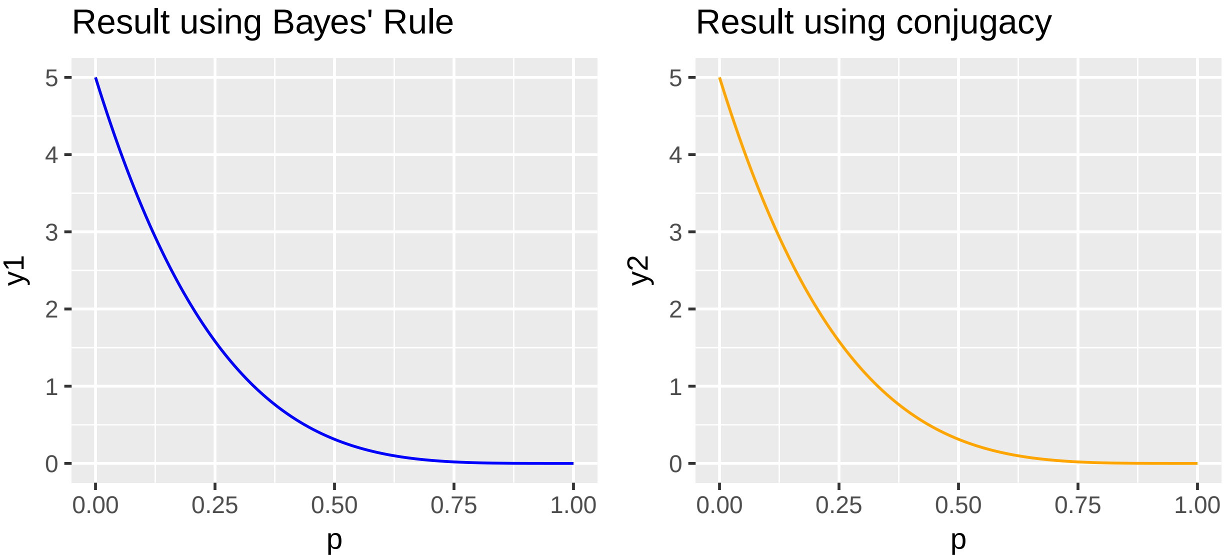
We can not calculate the posterior probability of H_2 by integrating the posterior distribution of p

$$P(H_2 \mid \text{data}) = P(p < 0.5 \mid \text{data}) = \int_0^{0.5} \pi^*(p)\, dp = \int_0^{0.5} 5(1-p)^4\, dp = 0.96875.$$

R Code

We can use dbeta to obtain the Beta distribution value and pbeta to obtain the probability of p up to a given value if p follows the Beta distribution. The hyperparameters α and β can be specified using shape1 and shape2 in the functions. We still provide you the R code of both methods, just in case you are curious.

```
# Method 1
integrand = function(p) {  # Handle integration
  (1 - p) ** 4
}
denominator = integrate(integrand, lower = 0, upper = 1)
p = seq(0, 1, by = 0.005)  # set a vector of values of p for plotting
y1 = (1 - p) ** 4 / denominator$value
plot1 = qplot(p, y1, geom = "line", col = I("blue"), main = "Result using Bayes' Rule")

# Method 2
y2 = dbeta(p, shape1 = 1, shape2 = 5)  # shape1 = alpha, shape2 = beta
plot2 = qplot(p, y2, geom = "line", col = I("orange"), main = "Result using conjugacy")
grid.arrange(plot1, plot2, ncol = 2)
```



```
# Probability of p<0.5 after seeing the data
pbeta(0.5, shape1 = 1, shape2 = 5)
```

```
## [1] 0.96875
```

Poisson Distribution

The second conjugate family is the Gamma-Poisson conjugate family, which involves the Poisson distribution (as the likelihood for the data), and the Gamma distribution (as the prior distribution for the parameter). We first talk about the Poisson distribution. This distribution is used to describe some number of occurrences within a give time period. For example, the number of phone calls in one day, the number of buses that come in one hour, or the number of Academy Award winners who die in a month. The random variable X is usually used as the number of counts that happen, with a parameter λ, which describes the *average* occurring rate of the counts. The Poisson distribution is defined as

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

This is a discrete distribution, therefore, we already have the pmf to help us to calculate the probability that k counts happen. Since λ describes the *rate*, its unit must be "something/unit".

Example 2

> Assume that the number of Academy Award winners who die in a month has the Poisson distribution with mean 1.5. What is the probability that 2 or more Academy Award winners die next month?

In this example, we have the mean of "deaths/month", which is the mean of the death rate, is 1.5. So $\lambda = 1.5$. Then we can set up the Poisson distribution as

$$P(X = k) = \frac{1.5^k}{k!}e^{-1.5}.$$

Remember, X here, represents the number of deaths. Now the question asks for the probability of 2 or more deaths in the next month. Mathematically, that is the same as

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + \cdots = \sum_{k=2}^{\infty} P(X = k).$$

This is an infinite sum and we do not know how to do it. So we consider the complementary event, that is, 1 or fewer death. The probability of 1 or fewer death in the next month is

$$P(X = 0) + P(X = 1) = \frac{1.5^0}{0!}e^{-1.5} + \frac{1.5^1}{1!}e^{-1.5} = 2.5e^{-1.5}.$$

So the probability of 2 or more deaths would be

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - 2.5e^{-1.5} \approx 0.442.$$

R Code

The value of Poisson distribution can be obtained by the `dpois` function. If you type `?dpois` in the R Console, you would see there are other Poisson related functions such as the `ppois` (for cumulative probability), `qpois` (for quantiles), `rpois` for generating random sequence that follows the Poisson distribution. We can use either `dpois` or `ppois` to solve the above Academy Award winner example.

```r
# Method 1: use ‘dpois‘ and complementary event
prob.comp = dpois(0, lambda = 1.5) + dpois(1, lambda = 1.5)
prob = 1 - prob.comp
prob
```

```
## [1] 0.4421746
```

```r
# Method 2: use ‘ppois‘ directly. Set lower.tail = FALSE to calculate P(X > 1)
prob = ppois(1, lambda = 1.5, lower.tail = FALSE)
prob
```

```
## [1] 0.4421746
```

Gamma-Poisson Conjugate Family

Suppose we do not have the exact information of the *average rate* λ, we would like to elicit the prior distribution of λ, then update the distribution using the data we have observed. It turns out that the family of Gamma distribution forms conjugacy with the Poisson distribution. There are two ways to define the Gamma distribution (and they are equivalent):

$$p(\lambda) = \mathsf{Gamma}(\lambda;\ k, \theta) = \frac{1}{\Gamma(k)\theta^k}\lambda^{k-1}e^{-\lambda/\theta},$$

or

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

k with θ, or α with β, are the hyperparameters for this family of distribution. Comparing the two definitions and you will see that $k = \alpha$, and $\beta = \frac{1}{\theta}$. The mean and the standard deviation of the Gamma distribution, i.e., the mean and the standard deviation of the independent variable λ, are calculated as

$$\begin{aligned} &\text{mean of } \lambda = k\theta, && \text{standard deviation of } \lambda = \theta\sqrt{k} \\ &\text{or,} && \\ &\text{mean of } \lambda = \frac{\alpha}{\beta}, && \text{standard deviation of } \lambda = \frac{\sqrt{\alpha}}{\beta}. \end{aligned}$$

The Gamma distribution is for continuous random variable which take positive values.

The updating rules under the two definitions:

$$\mathsf{Gamma}(k, \theta) \xrightarrow{\text{Poisson, } \sum x_i \text{ total counts within } n \text{ period}} \mathsf{Gamma}\left(k^* = k + \sum x_i,\ \theta^* = \frac{\theta}{n\theta + 1}\right)$$

or,

$$\mathsf{Gamma}(\alpha, \beta) \xrightarrow{\text{Poisson, } \sum x_i \text{ total counts within } n \text{ period}} \mathsf{Gamma}\left(\alpha^* = \alpha + \sum x_i,\ \beta^* = \beta + n\right).$$

Example 2 Continued

> Assume the number of Academy Award winners who die in a month has a Poisson distribution with mean λ. We don't know λ, but our expertise tells us that the prior distribution of λ is the Gamma distribution with $k = 0.25$ and $\theta = 6$. Next month 2 Academy Award winners die. What is the posterior mean of λ?

In this question, the data we have are, 2 total counts in 1 period (recall that we use 1 month as the unit for the rate λ). We only need to use the data to update the hyperparameters k and θ:

$$k^* = k + \text{total count} = 0.25 + 2 = 2.25,$$

$$\theta^* = \frac{\theta}{n\theta + 1} = \frac{6}{1 \times 6 + 1} = \frac{6}{7}.$$

Hence, the posterior mean of λ is

$$\text{posterior mean} = k^*\theta^* = 2.25 \times \frac{6}{7} \approx 1.93 \text{ deaths/month}.$$

Compared to the prior mean $k\theta = 0.25 \times 6 = 1.5$, this is a higher average death rate.

R Code

We can use `dgamma` to access the Gamma distribution.

```
# Before the data
k = 0.25; theta = 2

# Data
n = 1; counts = 2

# After the data
k_after = k + counts; theta_after = theta / (n * theta + 1)
k_after
```

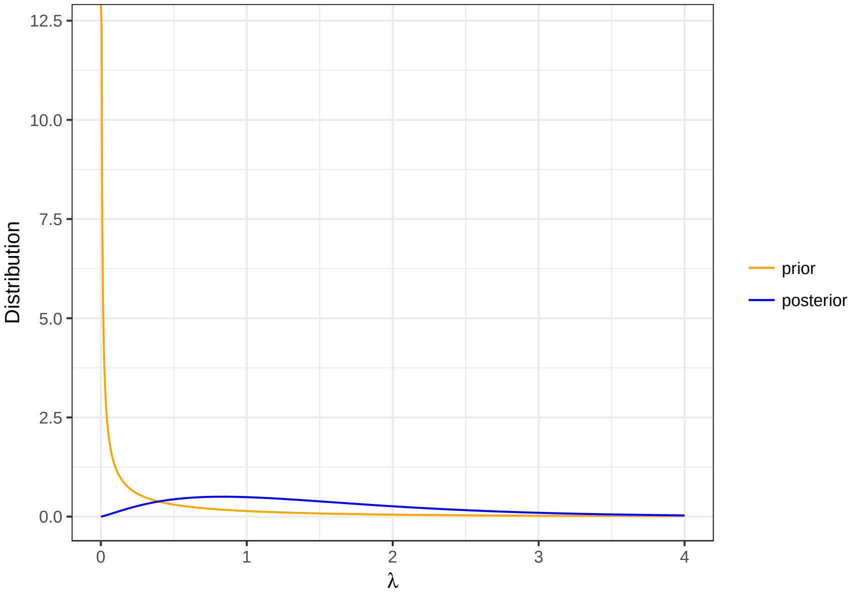
```
## [1] 2.25
```

```
theta_after
```

```
## [1] 0.6666667
```

The prior and posterior distributions of λ are as follows:

```
# Plots the prior and posterior distributions
x = seq(0, 4, by = 0.005)
y_prior = dgamma(x, shape = k, scale = theta)                  # prior
y_posterior = dgamma(x, shape = k_after, scale = theta_after) # posterior
data = data.frame(x = x, y_prior = y_prior, y_posterior = y_posterior)
ggplot(data = data) + geom_line(aes(x = x, y = y_prior, col = "a")) +
  geom_line(aes(x = x, y = y_posterior, col = "b")) +
  scale_color_manual(values = c("orange", "blue"), labels = c("prior", "posterior"), name = "") +
  xlab(expression(lambda)) + ylab("Distribution") + theme_bw()
```



As we can see, after the update, the distribution of λ is much flatterned and spread out than the original prior distribution.

(Optional) Unit of λ

The mean and the standard deviation of the Gamma distribution will be consistent under different units. For example, suppose we insist to set the *average rate* to be the number of deaths in 2 months, instead of 1 month. Then the prior mean of λ should be 3 deaths/2 months, and the prior standard deviation of λ would double too, which is 6 deaths/2 months. We can back solve the k and θ:

$$k_{\text{new}} = \frac{1}{4}, \qquad \theta_{\text{new}} = 12.$$

The data said that we observed 2 deaths the next month, which is 2 deaths in half ($n = 0.5$) of a 2 month period. We can calculate the new k and θ by

$$k^*_{\text{new}} = k_{\text{new}} + n = \frac{1}{4} + 2 = 2.25,$$

$$\theta^*_{\text{new}} = \frac{\theta_{\text{new}}}{n\theta_{\text{new}} + 1} = \frac{12}{7}.$$

One shall see that, the posterior mean and standard deviation also double. For example,

$$\text{posterior mean} = k^*_{\text{new}}\theta^*_{\text{new}} \approx 3.86 \text{ deaths/2 months} = 1.93 \text{ deaths/month}.$$

This result is consistent wich the previous result.

Normal-Normal Conjugate Family (σ Known)

Finally, we come to the Normal-Normal conjugate family. Recall that the Normal distribution has 2 parameters, the mean μ, and the standard deviation σ. The Normal-Normal conjugate family is used when σ is known and the distribution of μ is of interest.

Here, we "define"[5] the prior distribution of the mean μ to be Normal

$$p(\mu) = \mathsf{N}(\mu;\ \nu, \tau^2) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\mu-\nu)^2}{\tau^2}\right).$$

Here, ν is the mean of the data mean μ, and τ is the standard deviation of the data mean μ.

The updating rule is

$$\mathsf{N}(\nu, \tau^2) \xrightarrow{\text{Normal, } n \text{ data points with sample mean } \bar{x}} \mathsf{N}\left(\nu^* = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2},\ (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right).$$

Example 3

> In the "mass of ammonium nitrate" example, we assume the balance has a known standard deviation $\sigma = 0.2$. The chemist wants to know what the distribution of the average mass of her sample μ would be. The chemist thinks that the prior mean of μ is 10 milligrams and the prior standard deviation of μ is 2 milligrams, that is, $\mu \sim \mathsf{N}(10, 2^2)$. The chemist has 5 samples, with a sample mean $\bar{x} = 10.5$. What are the posterior mean of and the posterior standard deviation of μ?

We can simply plug the numbers we have in the formula

$$\nu^* = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2} = \frac{10 \times (0.2)^2 + 5 \times 10.5 \times 2^2}{(0.2)^2 + 5 \times (2^2)} \approx 10.499,$$

$$\tau^* = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}} = \sqrt{\frac{(0.2)^2(2)^2}{(0.2)^2 + 5 \times (2)^2}} \approx 0.089.$$

Hence, the posterior distribution of the mean of the mass μ is

$$\mu \mid \text{data} \sim \mathsf{N}(10.499, 0.089^2).$$

[5] same Normal distribution formula as before, but different notation to distinguish the hyperparameters

R Code

It is way easier to move the calculation to R. Here we calculate the posterior mean and posterior standard deviation of μ, which itself is the mean of the mass. Then we plot the two distributions for comparison.

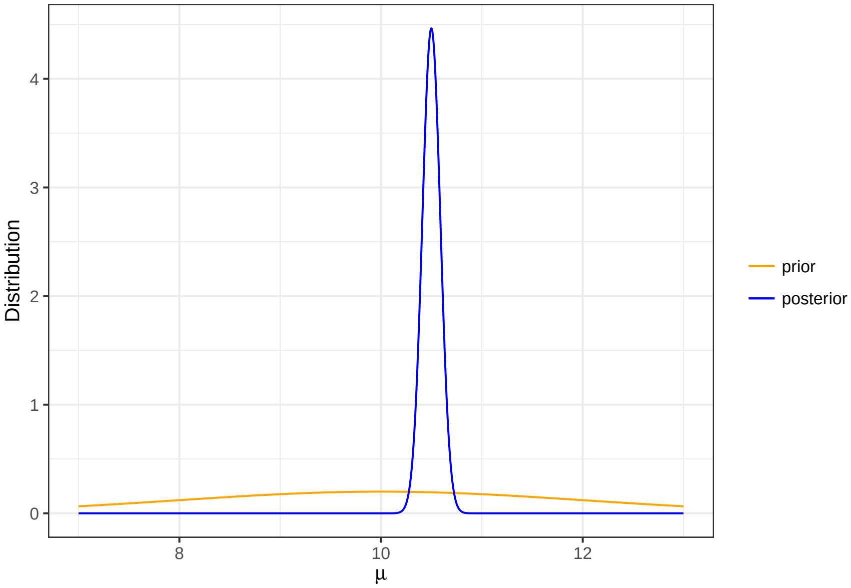
```
# Before the data
nu = 10; tau = 2

# Data
n = 5; x_bar = 10.5; sigma = 0.2

# After the data
nu_after = (nu * sigma^2 + n * x_bar * tau^2) / (sigma^2 + n * tau^2)
tau_after = sqrt(sigma^2 * tau^2 / (sigma^2 + n * tau^2))
nu_after
```

```
## [1] 10.499
```

```
tau_after
```

```
## [1] 0.08935341
```

```
# Plot prior and posterior distribution
x = seq(7, 13, by = 0.001)
y_prior = dnorm(x, mean = nu, sd = tau)
y_posterior = dnorm(x, mean = nu_after, sd = tau_after)
data = data.frame(x = x, y_prior = y_prior, y_posterior = y_posterior)
ggplot(data = data) + geom_line(aes(x = x, y = y_prior, col = "a")) +
  geom_line(aes(x = x, y = y_posterior, col = "b")) +
  scale_color_manual(values = c("orange", "blue"),
                     labels = c("prior", "posterior"), name = "") +
  xlab(expression(mu)) + ylab("Distribution") + theme_bw()
```



The distribution of μ has changed from a very flat Normal distribution to a very “concentrated” Normal distribution, after the input of the data.

Credible Intervals

Recall in Course 2 **Inferential Statistics**, we are able to construct the confidence interval of the parameter of interest by using the formula

$$(L, U) = (\text{point est.} - \text{critical value} \times \text{std. error},\ \text{point est.} + \text{critical value} \times \text{std. error}).$$

We looked up the critical value using either the z-table or t-table, and found the standard error using the sample standard deviation and the sample size. However, due to the fact that the hypothesis testing only provide the probability of seeing "some or more extreme cases" given the null hypothesis is true (i.e., $P(\text{some or more extreme cases} \mid H_0 \text{ is true})$), the interpretation of a 95% confidence interval is

> 95% of similarly constructed intervals will contain the parameter.

But more often, what we want is

> The probability that the parameter lies between L and U is 95%.

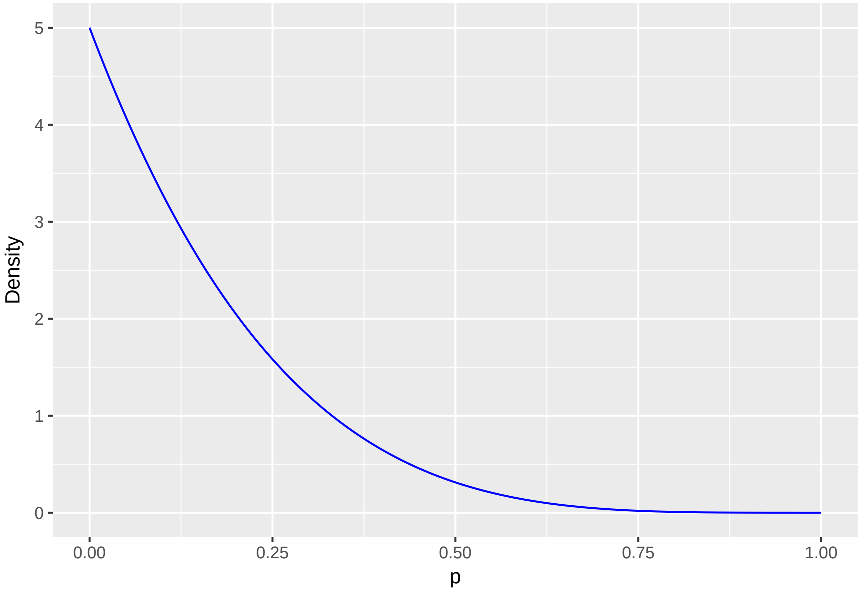
Now that we have the posterior distribution of the parameter, we can use this distribution to find such an interval. These intervals that actually "give us the probability" are called credible intervals.

Credible intervals are not unique, since we can use any interval such that the area under the distribution curve within this interval is 0.95. But we are more interested in those intervals that provide the same amount of area, but are the shortest. These intervals are called the **highest posterior density credible interval (HPD interval)**.

Example 1 Revisit

Recall in the RU-486 example, we have obtained the posterior distribution of the proportion p to be the Beta distribution

$$p \mid \text{data} \;\sim\; \mathsf{Beta}(1, 5).$$



Apparently, this probability density function "concentrates" most of the "mass" near $p = 0$ and is decreasing. That means, there are more areas near the point $p = 0$ than anywhere else. This implies that the higest probability density credible interval must start from $L = 0$.

Since $\mathsf{Beta}(1,5) = 5(1-p)^4$ is not a complicated function, we can easily obtain its cumulative distribution function (which calculate the cumulative area under the distribution)

$$F(x) = \int_0^x \mathsf{Beta}(1,5)\,dp = \int_0^x 5(1-p)^4\,dp = 1-(1-x)^5.$$

We have determined the lower bound of the interval $L = 0$. Now we only need to find the upper bound U. By setting

$$0.95 = F(U) - F(L) = 1-(1-U)^5 - F(0) = 1-(1-U)^5,$$

we can back solve U and get $U \approx 0.451$.

R Code

If you have noticed that the key to this example is the find U, which is a 95% quantile of the $\mathsf{Beta}(1,5)$ distribution, it is not hard to realize we should use the `qbeta` function to find U in R.

```
U = qbeta(0.95, shape1 = 1, shape2 = 5)
U
```

```
## [1] 0.4507197
```

However, suppose the posterior distribution does not share a nice shape as the $\mathsf{Beta}(1,5)$, we can also use the `HPDinterval` function from the `coda` package to get a simulation result.

```
library(coda)

# Draw random sample to approximate Beta(1, 5)
y = rbeta(10000, shape1 = 1, shape2 = 5)

# Convert vector y into an ‘mcmc‘ object
y_mcmc = as.mcmc(y)

# Obtain HPD
HPDinterval(y_mcmc, prob = 0.95)
```

```
##             lower     upper
## var1 0.0001100117 0.4569976
## attr(,"Probability")
## [1] 0.95
```

Predictive Inference on New Data

Posterior distributions can also be used to make predictive inference on new data. In this week, we will only focus on the discrete situation. We provide you the continuous version formula, but will not go into details. In Week 3 when we talk about more predictive inference on continuous cases, we will perform the inference using simulation methods.

Example 4

> Consider we have a coin with unknow probability of getting heads. Our prior knowledge is that, the probability of getting heads is either $H_1 : p = 0.7$ or $H_2 : p = 0.4$. We also believe that these two situations share the same prior probability, that is, $P(H_1) = P(p = 0.7) = 0.5$, and $P(H_2) = P(p = 0.4) = 0.5$. We toss the coin twice and receive 2 heads.

We can update the posterior probability of H_1 and H_2 by:

$$\begin{aligned}
P(H_1 \mid \text{data}) =& P(p = 0.7 \mid \text{data}) = \frac{P(2 \text{ heads}) \times P(H_1)}{P(2 \text{ heads}) \times P(H_1) + P(2 \text{ heads}) \times P(H_2)} = \frac{(0.7)^2 \times 0.5}{(0.7)^2 \times 0.5 + (0.4)^2 \times 0.5} \approx 0.754, \\
P(H_2 \mid \text{data}) =& P(p = 0.4 \mid \text{data}) = 1 - P(H_1 \mid \text{data}) \approx 0.246.
\end{aligned}$$

With the two posterior probabilities, we can use the law of total probability to predict the posterior probability of getting the next head:

$$\begin{aligned}
P(\text{next head} \mid \text{data}) =& P(\text{next head} \mid H_1, \text{data}) P(H_1 \mid \text{data}) + P(\text{next head} \mid H_2, \text{data}) P(H_2 \mid \text{data}) \\
=& P(\text{next head} \mid p = 0.7, \text{data}) P(p = 0.7 \mid \text{data}) + P(\text{next head} \mid p = 0.4, \text{data}) P(p = 0.4 \mid \text{data}) \\
=& 0.7 \times 0.754 + 0.4 \times 0.246 \approx 0.626.
\end{aligned}$$

R Code

We can use R again to perform the calculation

```r
prior = c(0.5, 0.5)
likelihood = c(dbinom(2, 2, 0.7), dbinom(2, 2, 0.4))
posterior = prior * likelihood / sum(prior * likelihood)

# Posterior conditional probability of next head
post.cond.head = c(dbinom(1, 1, 0.7), dbinom(1, 1, 0.4))

# Posterior total probability of next head
post.head = sum(post.cond.head * posterior)
post.head
```

```
## [1] 0.6261538
```

When the parameter of interest θ can take a continuum of values, we need the continuous version of the above formula to calculate the posterior probability for the data. Let X denote the random variable representing the "new data". Suppose $p(t \mid \theta, \text{data})$ is the probability distribution conditioning on the parameter θ and the observed data (similar to $P(\text{next head} \mid p = 0.7, \text{data})$ in the discrete example). $p^*(\theta) = p(\theta \mid \text{data})$ is the posterior probability of θ after seeing the data. Then the formal calculation to obtain the posterior probability $P(X \leq x \mid \text{data})$ is given by

$$P(X \leq x \mid \text{data}) = \int_{-\infty}^{\infty} P(X \leq x \mid \theta, \text{data}) p(\theta \mid \text{data})\, d\theta = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} p(t \mid \theta, \text{data}) p(\theta \mid \text{data})\, dt \right) d\theta.$$

This calculation is beyond the scope of this course so we only provide the formula here.

(Optional) Non-Conjugate Priors

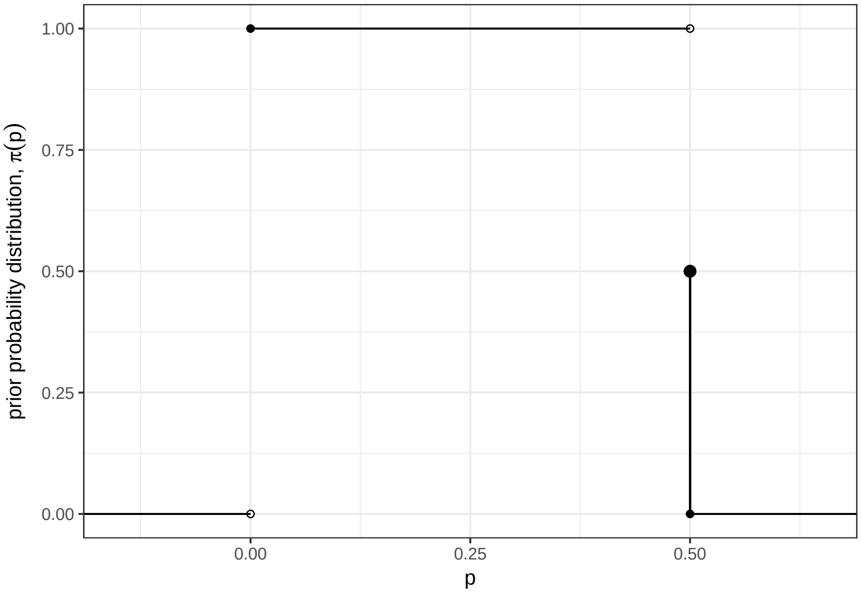
In many applications, the prior distribution and the likelihood of the data just do not form any conjugacy. In this case, Bayes' Rule still works, but we may not be able to obtain a nice closed form of the posterior distribution. Even we are able to calculate the posterior distribution by hands, this posterior distribution may no longer belong to any family of probability distributions. Often, we will use simulation to help us to get the numerical results.

Example 5

> Suppose in the RU-486 that the prior distribution of the proportion of pregnancy cases caused by RU-486 p is given by the following distribution.

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



This prior distribution has a uniformly flat distribution within $0 \leq p < 0.5$, and then a point mass at $p = 0.5$. It can be formulated as

$$\pi(p) = \begin{cases} 1, & 0 \leq p < 0.5 \\ 0.5, & p = 0.5 \\ 0, & \text{otherwise} \end{cases}$$

According to the data, we have 0 pregnancy from RU-486 out of 4 total pregnancy cases. We can update the posterior distribution using the Bayes' Rule

When $0 \leq p < 0.5$**,**

$$\pi^*(p) = \frac{(1-p)^4 \times (1)}{(1-p)^4 \times (1) + (1-0.5)^4 \times (0.5)} = \frac{40}{9}(1-p)^4.$$

When $p = 0.5$**,**

$$\pi^*(p) = \frac{(1-0.5)^4 \times (0.5)}{(1-p)^4 \times (1) + (1-0.5)^4 \times (0.5)} = \frac{5}{36}.$$

So combining together, we have the posterior distribution

$$\pi^*(p) = \begin{cases} \frac{40}{9}(1-p)^4, & 0 \leq p < 0.5 \\ \frac{5}{36}, & p = 0.5 \\ 0, & \text{otherwise} \end{cases}$$

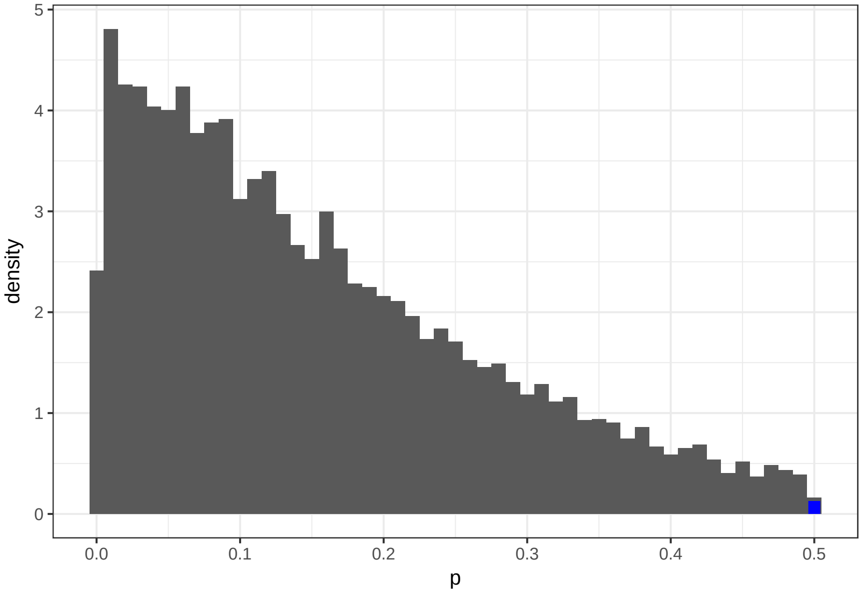
This is a piecewise function. To be able to use the data generated by this distribution, we often perform simulation.

Here we provide the R code that generates the histogram of the sampling data using Gibbs sampling in the `rjags` package. However, Gibbs sampling is beyond the scope of this course.

```
library(rjags)

# Set up model
str = "model {
  dummy ~ dunif(0, 1)  # this step should be fine
  p = ifelse(dummy > 0.5, 0.5, dummy)
  y ~ dbin(p, 4)
}"

set.seed(1234)

# Initialize data
data_jags = list(y = 0)
params = c(’dummy’, ’p’)
init = function(){
  init = list(’dummy’ = 0.2)   # just a random number
}

mod = jags.model(textConnection(str), data = data_jags, inits = init)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 13
##
## Initializing model
```

```
# Run Gibbs sampling
mod_sim = coda.samples(model = mod, variable.names = params, n.iter = 10000)

# Obtain sample data
p = mod_sim[[1]][, 2]

# Split p into 2 parts

# The point mass piece
p.pointMass = p[p == 0.5]

# The continuous piece
p.nonPointMass = p[p != 0.5]

# Plot the histogram of the continuous piece, then add the point mass
plot1 = qplot(x = p.nonPointMass, y = ..density.., geom = "histogram", binwidth = 0.01)
plot2 = plot1 + geom_segment(aes(x = 0.5, y = 0, xend = 0.5,
                                 yend = length(p.pointMass) / length(p)),
                     col = "blue", lwd = 3) + xlab("p") + theme_bw()
plot2
```



(Optional) Deviation of Conjugacy

In this section, we will derive the 3 conjugate families and discuss briefly about how the update from prior distribution to posterior distribution affects the effective sample size.

Beta-Binomial Conjugacy

In this conjugate family, the conjugate prior is the Beta distribution

$$\pi(p) = \mathsf{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

And the data follow the Binomial distribution with parameter p, the probability of success

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We apply the Baye's Rule to derive the posterior distribution of p:

$$\pi(p \mid X=k) = \frac{P(X=k)\times\pi(p)}{\displaystyle\int_0^1 P(X=k)\times\pi(p)\,dp} \tag{4}$$

$$= \frac{\left[\binom{n}{k} p^k(1-p)^{n-k}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]}{\displaystyle\int_0^1 \left[\binom{n}{k} p^k(1-p)^{n-k}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] dp}. \tag{5}$$

The denominator on the right-hand side of the last equality seems extremely complicated, however, we can use some trick to get around this. Since the denominator is an integral, it is just some constant normalizing the distribution so that the area under the curve of the posterior distribution $\pi(p \mid X = k)$ is 1. Therefore, we can rewrite (5) to be

$$\pi(p \mid X = k) \propto P(X = k) \times \pi(p) \propto \left[p^k(1-p)^{n-k}\right] \times \left[p^{\alpha-1}(1-p)^{\beta-1}\right] = p^{(\alpha+k)-1}(1-p)^{(\beta+n-k)-1}.$$

The last quantity shares a format as

$$p^{\text{some power}-1}(1-p)^{\text{another power}-1},$$

which again belongs to some distribution in the Beta family. Therefore, we argue that, the posterior distribution is still Beta, with $\alpha^* = \alpha + k$, and $\beta^* = \beta + n - k$.

The constant terms together, in fact, can be shown to be

$$\frac{\binom{n}{k}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\displaystyle\int_0^1 \left[\binom{n}{k}p^k(1-p)^{n-k}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\right]\, dp} = \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}.$$

Effective Sample Size

We may also compare the change from the prior mean to the posterior mean of the distributions. The posterior mean of the new Beta distribution, $\mathsf{Beta}(\alpha^* = \alpha + k,\ \beta^* = \beta + n - k)$, is

$$\begin{aligned}\text{posterior mean} = \frac{\alpha^*}{\alpha^*+\beta^*} &= \frac{\alpha+k}{(\alpha+k)+(\beta+n-k)} = \frac{\alpha+k}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n} + \frac{k}{\alpha+\beta+n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\cdot\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\cdot\frac{k}{n}\end{aligned}$$

In the last equality, we decompose the posterior mean as a weighted average between the prior mean $\dfrac{\alpha}{\alpha+\beta}$ and the "mean" (average success rate) from the data $\dfrac{k}{n}$, with weights $\dfrac{\alpha+\beta}{\alpha+\beta+n}$ and $\dfrac{n}{\alpha+\beta+n}$. n is the actual sample size from the data. We call $\alpha+\beta$, and $\alpha+\beta+n = \alpha^*+\beta^*$ the corresponding effective sample size of the prior and posterior Beta distributions respectively. Then this decomposition can be interpreted as

$$\underbrace{\frac{\alpha^*}{\alpha^*+\beta^*}}_{\text{posterior mean}} = \underbrace{\left(\frac{\overbrace{\alpha+\beta}^{\text{prior effective sample size}}}{\underbrace{\alpha+\beta+n}_{\text{posterior effective sample size}}}\right)}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\alpha+\beta+n}_{\text{posterior effective sample size}}}\right)}_{\text{data weight}} \cdot \underbrace{\frac{k}{n}}_{\text{data mean}},$$

with the two weights as the ratio between the prior sample size and the posterior sample size, and the ratio between the actual data sample size, and the posterior sample size.

Since we need to impose the prior to derive the posterior, we normally do not know exactly what the prior effective sample size should be. Sometimes we call the prior effective sample size the "imaginary sample size", which implies we might need to use expert opinion or some trials to get the right sample size so that the prior is compatible with the data.

With the notion of effective sample size, we can compare the difference between the uniformly flat Beta distribution $\mathsf{Beta}(1, 1)$, and the one that really concentrates in the middel $\mathsf{Beta}(100, 100)$. We can see that, both distributions have the same mean, $\dfrac{1}{2}$, but with different effective sample size, 2 and 200. Suppose the data sample size is n. The prior weights of the two different prior would be $\dfrac{2}{2+n}$ compared to $\dfrac{200}{200+n}$. Since $\dfrac{200}{200+n}$ is closer to 1, this means, imposing a prior distribution with a larger prior sample size would give a posterior mean that is closer to the prior mean. That is to say, the larger the prior sample size, the stronger effect the prior has on the overall results, and the harder the data can "move" the prior and affect the prior belief. Therefore, it is very important to consider the prior sample size when imposing prior.

Gamma-Poisson Conjugacy

We will derive this conjugacy using the same trick from the Beta-Binomial conjugate family. In the Gamma-Poisson conjugate family, we have the conjugate prior as the Gamma distribution

$$p(\lambda) = \mathsf{Gamma}(k, \theta) = \frac{1}{\Gamma(k)\theta^k}\lambda^{k-1}e^{-\lambda/\theta},$$

or equivalently,

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

And we assume the data follow the Poisson distribution with parameter λ, the average rate of occurrences

$$P(X = x_i) = \frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}.$$

Here, x_i denotes the number of occurrences, or the count, **within 1 period of unit**.

Using the Bayes' Rule, **over 1 period of unit**, we get,

$$p(\lambda \mid X = x_i) \propto P(X = x_i) \times p(\lambda) \propto \left[\lambda^{x_i}e^{-\lambda}\right] \times \left[\lambda^{k-1}e^{-\lambda/\theta}\right] = \lambda^{x_i+k-1}e^{-(1+1/\theta)\lambda},$$

or equivalently,

$$p(\lambda \mid X = x_i) \propto P(X = x_i) \times p(\lambda) \propto \left[\lambda^{x_i}e^{-\lambda}\right] \times \left[\lambda^{\alpha-1}e^{-\beta\lambda}\right] = \lambda^{x_i+\alpha-1}e^{-(1+\beta)\lambda}$$

So the new hyperparameters are

$$k^* = x_i + k, \qquad \theta^* = \frac{1}{1+1/\theta} = \frac{\theta}{\theta+1},$$

or

$$\alpha^* = x_i + \alpha, \qquad \beta^* = 1 + \beta.$$

This is the result if we observe the data within 1 period of unit. We can use **sequential updating** to derive the hyperparameters after we have observed the data within n period of unit:

$$k^* = \sum_{i=1}^n x_i + k, \qquad \theta^* = \frac{1}{n+1/\theta} = \frac{\theta}{n\theta+1},$$

or

$$\alpha^* = \sum_{i=1}^n x_i + k, \qquad \beta^* = n + \beta.$$

Effective Sample Size

The effective sample size of a Gamma distribution is its rate parameter β or the inverse of the scale parameter $\frac{1}{\theta}$. Here, we analyze the decomposition of the posterior mean using the $\mathsf{Gamma}(\alpha, \beta)$ definition.

$$
\begin{aligned}
\text{posterior mean} =& \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum x_i}{\beta + n} \\
=& \underbrace{\left(\frac{\overbrace{\beta}^{\text{prior effective sample size}}}{\underbrace{\beta + n}_{\text{posterior effective sample size}}} \right)}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\beta}}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\beta + n}_{\text{posterior effective sample size}}} \right)}_{\text{data weight}} \cdot \underbrace{\frac{\sum x_i}{n}}_{\text{data mean}}
\end{aligned}
$$

When the prior effective sample size β is larger, the posterior mean will be closer to the prior mean, and it will take more period of unit n to change our prior belief.

Normal-Normal Conjugacy

When the variance of the data σ^2 is known, we can consider using the Normal-Normal conjugate family to update the distribution of the mean of the data μ. In this conjugate family, the conjugate prior is the Normal distribution with mean ν and variance τ^2

$$
p(\mu) = \mathsf{N}(\nu, \tau^2) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(\mu - \nu)^2}{\tau^2} \right).
$$

And we assume the data follow another Normal distribution, with parameter μ and known variance σ^2

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

We first derive the posterior distribution hyperparameters using only one data point x_i. We apply the Bayes' Rule, and use some algebra the manipulate the terms, and get

$$
\begin{aligned}
p(\mu \mid x_i) \propto p(x = x_i \mid \mu) \times p(\mu) \propto & \left[e^{-\dfrac{(x_i - \mu)^2}{2\sigma^2}} \right] \times \left[e^{-\dfrac{(\mu - \nu)^2}{2\tau^2}} \right] \\
&= \exp\left\{ -\frac{1}{2} \frac{(\mu - \frac{\nu\sigma^2 + x_i\tau^2}{\tau^2 + \sigma^2})^2}{\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}} + (1 - \frac{1}{\sigma^2 + \tau^2})(\sigma^2\nu^2 + (x_i)^2\tau^2) \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \frac{(\mu - \nu^*)^2}{(\tau^*)^2} \right\},
\end{aligned}
$$

where the new hyperparameters

$$
\nu^* = \frac{\nu\sigma^2 + x_i\tau^2}{\sigma^2 + \tau^2}, \qquad (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.
$$

Then under **sequential updating**, we can obtain the results if we have n data points $\{x_1, x_2, \ldots, x_n\}$ with sample mean $\bar{x}$

$$\nu^* = \frac{\nu\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}, \qquad (\tau^*)^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$$

Effective Sample Size

The effective sample size of a Normal distribution $\mathsf{N}(\nu, \tau^2)$ with known variance σ^2 under the **Normal-Normal conjugate family**[6] is $\dfrac{\sigma^2}{\tau^2}$, which is proportional to the precision of $\mathsf{N}(\nu, \tau^2)$.

We can decompose the posterior mean ν^* as

$$\begin{aligned}
\text{posterior mean } = \nu^* =& \frac{\nu\sigma^2 + (\sum x_i)\tau^2}{\sigma^2 + n\tau^2} = \frac{\nu\sigma^2}{\sigma^2 + n\tau^2} + \frac{(\sum x_i)\,\tau^2}{\sigma^2 + n\tau^2} \\
=& \frac{\sigma^2}{\sigma^2 + n\tau^2} \cdot \nu + \frac{n\tau^2}{\sigma^2 + n\tau^2} \cdot \frac{\sum x_i}{n} \\
=& \underbrace{\left(\frac{\overbrace{\dfrac{\sigma^2}{\tau^2}}^{\text{prior effective sample size}}}{\underbrace{\dfrac{\sigma^2}{\tau^2} + n}_{\text{posterior effective sample size}}} \right)}_{\text{prior weight}} \cdot \underbrace{\nu}_{\text{prior mean}} + \underbrace{\left(\frac{\overbrace{n}^{\text{data size}}}{\underbrace{\dfrac{\sigma^2}{\tau^2} + n}_{\text{posterior effective sample size}}} \right)}_{\text{data weight}} \cdot \underbrace{\frac{\sum x_i}{n}}_{\text{data mean}}
\end{aligned}$$

In the Normal-Normal conjugacy, it turns out that for a fixed data variance σ^2, the **smaller** τ^2 is, the **larger** the prior effective sample size is. This makes sense because the variance τ^2 represents our prior belief of the **spread** of the prior data. The less variance of our belief in the prior data, the more centered the prior will be, then the stronger we trust our prior.

[6] We will introduce another conjugate family the Normal-Gamma conjugate family in Week 3.