Complete all **Exercises**, and submit answers to **Questions** in the **Quiz: Week 2 Lab** on Coursera.

## Getting Started

### Load packages

In this lab we will explore some basic Bayesian inference using conjugate priors and credible intervals to examine some categorical and count data from the [CDC’s Behavioral Risk Factor Surveillance System](http://www.cdc.gov/brfss/) (BRFSS). A subset of these data from 2013 have been made available in the statsr package, as usual we will first load the package and then the data set.

Let’s load the package,

library(statsr)  
data(brfss)

This data set contains 5000 observations of 6 variables:

|  |  |
| --- | --- |
| variable | description |
| weight | Respondent’s weight in pounds. |
| height | Respondent’s height in inches. |
| sex | Respondent’s sex |
| exercise | Has the respondent exercised in the last 30 days |
| fruit\_per\_day | How many servings of fruit does the respondent consume per day |
| vege\_per\_day | How many servings of dark green vegetables does the respondent consume per day |

## Credible Interval Calculator

Recall that probability distribution (prior/posterior distribution) of a parameter that describes the distribution of the data is given by:

Here, , , and are the variables of their own distributions (the values of them define the distributions of the data), and other parameters such as , and are the parameters of the distributions of , , and .

(**Note:** In this lab, we use the following definition of Gamma distribution:

This definition of the Gamma distribution is different from the one introduced in the video lecture.)

Below is an interactive app for visualizing posterior distributions and credible intervals of , , and given different values of parameters. We will use this app to explore how both our choice of prior distribution, as well as our data, affect the posterior distribution and the credible interval for , , and .

Note that this app assumes you now the posterior distribution as well as the parameters of this distribution. In the remainder of the lab we will walk you through how to calculate the posterior distribution in the Beta-Binomial Conjugacy and the Gamma-Poisson Conjugacy cases based on real world data from BRFSS. Then you will be asked to calculate the credible interval using codes similar to the one shown under the graph of the app.

First, let us do some exercises to learn how to use this app.

credible\_interval\_app()

## Error in credible\_interval\_app(): Shiny app will only run when built within RStudio.

Suppose the posterior distribution of follows a Normal distribution with mean 10 and variance 5. Which of the following are the bounds of a 95% credible interval for ? Answer this question using the app.

* (-1.96, 1.96)
* (0.419, 0.872)
* (0.959, 3.417)
* (5.618, 14.382)

Confirm your answer by running the code given below the distribution plot in the app.

# Type your code for Exercise 1 here.

Suppose the posterior distribution of follows a Beta distribution with and . Which of the following are the bounds of a 90% credible interval for ? Answer this question using the app.

* (-1.678, 5.678)
* (0.043, 0.641)
* (0.063, 0.582)
* (0.071, 0.949)

Confirm your answer by running the code given below the distribution plot in the app.

# Type your code for the Exercise 2 here.

Suppose the posterior distribution of follows a Gamma distribution with and . Which of the following are the bounds of a 99% credible interval for ? Answer this question using the app.

* (-3.284, 11.284)
* (0.069, 0.693)
* (0.084, 1.372)
* (0.171, 0.969)

Confirm your answer by running the code given below the distribution plot in the app.

# Type your code for the Exercise 3 here.

## Beta-Binomial Conjugacy

As we discussed in the videos, the Beta distribution is conjugate to the Binomial distribution - meaning that if we use a Beta prior for the parameter of the Binomial distribution then the posterior distribution of after observing data will be another Beta distribution.

Our goal with inference in general is to take specific observations (data) and use them to make useful statements about unknown population parameters of interest. The Beta-Binomial Conjugacy is a Bayesian approach for inference about a single population proportion . Whereas with the frequentist approach we used we will now just use and directly with being the number of successes obtained from identical Bernoulli trials. (A Bernoulli trial is a random experiment with exactly two possible outcomes, “success” and “failure”, in which the probability of success is the same every time the experiment is conducted.) As such, we can view as a Binomial random variable with the number of trials, and the probability of success.

To complete our Bayesian approach of inference, all we need is to define our prior beliefs for by defining a prior distribution. Our choice of the prior hyperparameters ( and ) should reflect our prior beliefs about . In the following, we will use the term **hyperparameter** to define parameters of prior/posterior distributions, and the term **parameter** to define the unknown parameters of the likelihood, such as . For most conjugate distributions there is usually a straight forward interpretation of these hyperparameters as the previously observed data – in the case of the Beta-Binomial Conjugacy, we can think of our hyperparameters as representing previous successes and previous failures.

### Data and the updating rule

We will start by performing inference on the sex ratio of respondents to BRFSS, we can define success as being Female and we would like to make some statement about the overall sex ratio of American adults based on our sample from BRFSS. We will do this by estimating , the true proportion of females in the American population, using credible intervals. For each credible interval you compute, always check back in with your intuition, which hopefully says that should be around 0.5 since we would expect roughly 50% females and 50% males in the population.

Here is the observed sex distribution in the data:

table(brfss$sex)

##   
## Male Female   
## 2414 2586

Let’s store the relevant, total sample size and number of females, for use in later calculations:

n <- length(brfss$sex)  
x <- sum(brfss$sex == "Female")

For each observed data point from a Binomial ( and ) we can calculate the values of the posterior parameters using the following updating rule:

From the data we now have (the number of females), and (the number of males). We’ll start with a Beta prior where and . Remember that this is equivalent to a Uniform distribution. By combining the data with the prior, we arrive at a posterior where

What is the 95% credible interval for , the proportion of females in the population, based on the posterior distribution obtained with the updating rule shown above. Use the credible interval app to answer this question.

* (0.500, 0.536)
* (0.503, 0.531)
* (0.507, 0.530)
* (0.468, 0.496)

Which of the following is the correct Bayesian interpretation of this interval?

* The probability that the true proportion of females lies in this interval is either 0 or 1.
* The probability that the true proportion of females lies in this interval is 0.95.
* 95% of the time the true proportion of females is in this interval.
* 95% of true proportions of females are in this interval.

Let’s now use a more informative prior that reflects a **stronger** belief that the sex ratio should be 50-50. For this, we use a Beta prior with hyperparameters and .

Confirm by plotting the following two Beta distributions and using the app above to show that the distribution is centered around 0.5 and much more narrow than the uniform distribution, i.e. .

What is the 95% credible interval for , the proportion of females in the population, based on a prior distribution of . **Hint:** You need to determine the hyperparameters of the posterior distribution, then use the app to construct the credible interval.

* (0.498, 0.531)
* (0.500, 0.528)
* (0.504, 0.532)
* (0.502, 0.527)

Let’s consider one other prior distribution: .

Which is of the following is the center of the distribution? **Hint:** modify the code under the distribution plot to get the center.

* approximately 0.03
* approximately 0.15
* approximately 0.50
* approximately 0.97

# Type your code for Question 7 here.

What is the 95% credible interval for , the proportion of females in the population, based on a prior distribution of . **Hint:** You need to determine the posterior distribution first, then use the app to construct the credible interval.

* (0.503, 0.531)
* (0.499, 0.535)
* (0.486, 0.509)
* (0.484, 0.511)

In summary, when we used a prior distribution that was centered around a realistic value for (the center is around 0.5), the credible interval we obtained was also more realistic. However when we used a strong prior distribution that was centered around a clearly unrealistic value for (say the prior), the credible interval we obtained did not match the distribution of the data (with the proportion of female respondents ). Hence, a good prior helps, however a bad prior can hurt your results.

Next, let’s turn our attention to the exercise variable, which indicates whether the respondent exercised in the last 30 days. While for the sex variable we had some intuition about the true proportion of females (we would expect it to be around 0.5), many of us probably do not have a strong prior belief about the proportion of Americans who exercise. In this case we would be more inclined to use a non-informative prior, e.g. a uniform distribution, which says that is equally likely to be anywhere between 0 and 1.

Here is the observed exercise distribution in the data:

table(brfss$exercise)

##   
## Yes No   
## 3868 1132

What is the 90% credible interval for , the proportion of Americans who exercise, based on a uniform prior distribution?

* (0.762, 0.785)
* (0.764, 0.783)
* (0.718, 0.737)
* (0.758, 0.789)

## Gamma-Poisson Conjugacy

Since the Poisson distribution describes the number of counts in a given interval, we will use this distribution to model the fruit\_per\_day variable which records the servings of fruit the respondents consume per day. The Poisson distribution has a single parameter, , which is the expected number of counts per time period.

The Gamma-Poisson conjugacy is another example of conjugate families where we use the Gamma distribution as the prior for the count parameter . In this lab, we use the following definition of Gamma distribution:

With Bayes’ Rule and the likelihood which is given by the Poisson distribution, we will get a Gamma posterior for .

Once again, our choice of the prior parameters ( and ) should reflect our prior beliefs about the parameter . In the case of the Gamma-Poisson conjugacy, we can view as the number of total counts and as the prior number of observations. For example, setting and reflects a belief based on data that 3 respondents on average consume a total of 12 fruits per day. At a first glance, this might sound equivalent to setting and or and , however these three distributions, , , and , while they all have the same expected value 4, differ in their spreads which indicates a different degree of belief about the parameter .

Use the app to plot the following three prior Gamma distributions, , , and . Confirm that they all have the same center but different spreads. Order them in ascending order of spreads, from least to most variable.

### Data and the updating rule

For each observed data point from the Poisson distribution () we can calculate the values of the posterior parameters using the following updating rule:

However in this case we have 5000 observations and we would like to avoid updates every single count individually. As we saw last week, we can use our subsequentially updated posterior as a new prior. As such, a more general multi-observation updating rule is

Using the multi-observation updating rule, what should the posterior distribution be when the hyperparameters of the Gamma prior are and , and we have observed the data .

* Gamma(, )
* Gamma(, )
* Gamma(, )
* Gamma(, )

# Type your code for Question 10 here.

The government recommends that Americans consume approximately 5 servings of fruits per day. Which of the following represents a weak prior that Americans on average follow this recommendation?

* Gamma(, )
* Gamma(, )
* Gamma(, )
* Gamma(, )

Using the correct prior distribution from the previous question and the data of fruit\_per\_day in the BRFSS dataset, calculate the hyperparameters of the posterior distribution.

* Gamma(, )
* Gamma(, )
* Gamma(, )
* Gamma(, )

# Type your code for Question 12 here.

Using the correct posterior distribution from the previous question, calculate the 90% credible interval for , the expected number of servings of fruit Americans consume per day.

* (1.575, 1.668)
* (1.588, 1.659)
* (1.592, 1.651)
* (1.594, 1.653)

Based on this result, do Americans appear to follow the government guidelines which recommend consuming 5 servings of fruits per day?

* Yes
* No

Repeat the preceding analysis for number of servings of vegetables per day (vege\_per\_day), and evaluate whether Americans follow the government guidelines which recommend consuming 5 servings of vegetables per day.

# Type your code for the Exercise 6 here.

This work is licensed under [GNU General Public License v3.0](https://www.gnu.org/licenses/quick-guide-gplv3.html).