Web Activity: Bayesian Analysis of Binary Data

Prerequisites:

- Be familiar with the probability concepts of a random variable and conditional probability.
- Know that the Bernoulli distribution (which corresponds to a Binomial distribution with $n = 1$) is used to model data that can only take two values, also known as binary data.
- Know that, when making inference for binary data, a random sample is used to estimate the value of the probability of success, p.
- Be able to calculate and interpret approximate confidence intervals for the probability of success of binary data.
- In the context of Bayesian statistics:
  - Be familiar with the concepts of prior distribution, hyperparameter, likelihood, and posterior distribution.
  - Know that the prior distribution should reflect your initial belief about the true value of the parameter.
  - Be able to describe the roles of the hyperparameters in shaping the prior distribution.

Learning Outcomes:
In the context of binary data with success probability of p:

- Given data and a Beta prior, calculate by hand the parameters of the posterior distribution of p.
- Given data and a Beta prior, order, from largest to smallest, the prior mean of p, the sample mean, and the posterior mean of p.
- Given data and a Beta prior, obtain the posterior mean and variance of p using an online interactive resource.
- Compare and contrast the interpretation of a frequentist $(1-\alpha)\times 100\%$ confidence interval for p with a Bayesian $(1-\alpha)\times 100\%$ credible interval for p.
- Explain the impact that different prior distributions have on the posterior distribution.

In this activity, we will develop an intuition about Bayesian inference in the context of binary data using an online interactive resource. Binary data can only take two values, usually encoded as 0 (failure) or 1 (success).

For independent binary data $X_1, ..., X_n$, where each datum takes the value of 1 with probability p and the value of 0 with probability $1 - p$, recall that the frequentist estimator of p is given by the sample mean:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

An approximate $(1 - \alpha) \times 100\%$ confidence interval for p is given by

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}},$$

where $z_{1-\alpha/2}$ is the upper quantile of order $\alpha/2$ of a standard Normal distribution.

In the Bayesian paradigm, the parameter p is treated as a random variable. The initial uncertainty about p is reflected through a user-specified distribution called the *prior distribution.* A commonly used prior for p is a Beta distribution, $p \sim \text{Beta}(a, b)$. The parameters a and b, chosen by the practitioner to reflect their prior knowledge about p, are called *hyperparameters.* The mean of a Beta distribution with these parameters is given by

$$\mathsf{E}[p] = \frac{a}{a + b}.$$

Bayesian inference is based on the so-called *posterior distribution*, which is defined as the conditional distribution of p given the data: $p \,|\, X_1, ..., X_n$. Bayes' theorem is used to find the posterior distribution. In the case of binary data, and when the prior is a Beta distribution, the posterior is also a Beta distribution: $p \,|\, X_1, ..., X_n \sim \text{Beta}(a', b')$, where

$$a' = a + \sum_{i=1}^{n} X_i, \quad b' = b + n - \sum_{i=1}^{n} X_i.$$

A Bayesian estimator of p is given by the posterior mean, which for a $\text{Beta}(a', b')$ distribution is given by

$$\hat{p}_{\text{B}} = \mathsf{E}[p \,|\, X_1, ..., X_n] = \frac{a'}{a' + b'}.$$

As in frequentist statistics, we can construct intervals to quantify our (posterior) uncertainty about p. Specifically, we can find numbers c_* and c^* such that

$$\Pr(p \in [c_*, \, c^*] \,|\, X_1, ..., X_n) = 1 - \alpha.$$

The interval $[c_*, c^*]$ is called a $(1-\alpha) \times 100\%$ *credible interval* for p. Recall that you cannot say that the probability of p being in a confidence interval is $1-\alpha$. However, the probability of p being in the credible interval is indeed given by $1-\alpha$ because in the Bayesian paradigm p is considered random. Observe that c_* and c^* need not be unique—in this activity we will use what is known as the *highest posterior density* (HPD) interval, in which values outside the interval are less likely than values inside it; see this article.[1] Finally, descriptive statistics of the posterior distribution can inform further inference.

The resource can be accessed at

`https://shiny-apps.stat.ubc.ca/FlexibleLearning/FirstBayes/Beta-Binomial/`

The resource allows you to specify the hyperparameters a and b of the Beta prior distribution of p for the data to be used in the analysis. You can select from multiple available data sets, use your own data by providing the observed $\hat{p}$, or generate data by specifying the true value of p. Once you define the prior and the data you are going to work with, the resource calculates the posterior distribution of p and prints relevant descriptive statistics. You can also specify a credibility level α and the resource will calculate a $(1-\alpha) \times 100\%$ HPD credible interval.

Throughout Sections 1 to 3, we will study a (subset of) a data set which contains information about admissions to the University of California Berkeley.[2] You can view this as a random sample of admission results from previous years. Each observation encodes whether an applicant was accepted (encoded as $X = 1$) or rejected (encoded as $X = 0$). We are interested in estimating the value of $p = \Pr(X = 1)$, the probability of getting accepted. Before starting the activity, what do you think is the average admission rate to UC Berkeley?

[1] Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). *bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework*

[2] The data set is part of R's `datsets` package. See `http://unixlab.stat.ubc.ca/R/library/datasets/html/UCBAdmissions.html` for more information.

1 Defining the prior distribution

For this section, go to the *Prior distribution* tab of the resource. The purple curve corresponds to the density of a Beta(a, b) distribution, where a and b can be set on the left tab of the resource. To visualize how the parameters of the Beta distribution affect the shape of its density, visit this link: `https://upload.wikimedia.org/wikipedia/commons/7/78/PDF_of_the_Beta_distribution.gif`.

1) In the preamble to the activity, you were asked what you think is the average admission rate to UC Berkeley. Write down your answer. How certain are you of your answer?

2) What are the current values of the hyperparameters of the prior distribution? What are the prior mean and variance?

3) How is the prior mean represented in the plot?

4) Modify the a and b settings in the left tab of the resource until you find a combination with the same mean as what you answered in question 1).

5) For these values of a and b, what is the variance of the prior distribution? Do you think it is too large, adequate, or too small in terms of reflecting your uncertainty in question 1)? Use the curve on the right plot to inform your thoughts.

6) If you answered too large or too small in question 5), modify the hyperparameters until you find a density with the same (or similar) mean as in question 1), but with a variance that appropriately reflects your uncertainty.

7) Suppose that instead of UC Berkeley, we were to study admission rates of your university. Would the average admission rate be different than the one you selected in question 1)? Would you feel more certain about this value?

2 Selecting a data set

For this section, go to the *Data and likelihood* tab of the resource. The upper plot displays the number (and percentage) of successes and failures in the data set. The curve in the lower plot corresponds to the likelihood function.

1) Recall that we are going to analyze the UC Berkeley admissions data set. Make the proper selection in the left tab of the resource.

2) For this data set, what is the value of $\hat{p}$? Is it similar to the prior mean you selected in Section 1?

3) Based on the value of $\hat{p}$ from question 2), construct a 90% approximate confidence interval for p. Do you need any assumptions for this approximate confidence interval to be valid?
Hint: $z_{0.95} = 1.645$.

4) Interpret, using as few technical terms as possible, the confidence interval you obtained in question 3).

3 Analyzing the posterior distribution

Go to the *Posterior analysis* tab of the resource. The curve in the plot corresponds to the posterior distribution of p.

A1) Calculate the hyperparameters of the posterior distribution, a' and b'.
Hint: use the equation in the preamble to the activity and the information in the Data and likelihood *tab of the resource.*

A2) Based on question A1), identify the posterior distribution of p.

A3) What is the posterior mean of p? How is it represented in the plot on the right?

A4) The posterior mean is ____________the prior mean.
(*Fill in the blank with one of <u>smaller than</u>, <u>equal to</u>, or <u>greater than</u>.*)

A5) The posterior mean is ____________$\hat{p}$.
(*Fill in the blank with one of <u>smaller than</u>, <u>equal to</u>, or <u>greater than</u>.*)

A6) Based on questions A4) and A5), the posterior mean is ____________the prior mean and $\hat{p}$.
(*Fill in the blank with one of <u>smaller than both</u>, <u>greater than both</u>, <u>between</u>, or <u>not comparable to</u>.*)

A7) Modifying the α setting on the left tab of the resource, obtain a 90% credible interval for p.

A8) How is the credible interval you obtained in question A7) represented in the plot on the right-hand side?

A9) Considering that p is a random variable, interpret the credible interval you obtained in question A7).
Hint: refer to the preamble to the activity.

A10) Compare the width and overall location of the credible interval of question A7) with the confidence interval you obtained in Section 2, question 3).

A11) Contrast the interpretation of both intervals. (Refer to Section 2, question 4).) Which interpretation do you find more intuitive?

A12) What is the posterior variance of p?

A13) Compared to the prior variance of p in Section 1, the posterior variance is ___________. Briefly explain without calculations why you think this is so. (*Fill in the blank with one of smaller than, equal to, or greater than.*)

Now go to the *Summary* tab of the resource. The display on the right is called a *triplot* because it shows the prior density, likelihood, and posterior density in a single plot.

B1) The first table in the left tab shows the prior, sample, and posterior means. Do the values in this table reflect what you answered in question A6)?

B2) How is the relationship between the prior, sample, and posterior means represented in the plot?

B3) The second table in the left tab shows the hyperparameters you chose and the updated ones. Do the new hyperparameters correspond to the ones you calculated in question A1)?

Optional questions. The questions in this box are not necessary for completing the activity. They are more theoretical than the questions in the rest of the activity.

C1) Using the formula for the hyperparameter update given in the preamble of the activity, prove that the posterior mean is a weighted average of the prior and sample means:

$$\hat{p}_{\mathrm{B}} = \frac{a+b}{a+b+n}\hat{p}_{\mathrm{prior}} + \frac{n}{a+b+n}\hat{p}. \tag{1}$$

C2) How is Eqn. (1) related to your answer in question A6)?

C3) Referring only to the numerators of the weights in Eqn. (1), the sample mean is assigned a weight proportional to n. To what is the weight assigned to the prior mean proportional to?

C4) One way to interpret Eqn. (1) is that the sample mean is given a weight proportional to the number of observations. Likewise, you can think of a fictional "prior sample" related to the weight of the prior mean. How does the size of this prior sample relate to a and b?
(*Note: this sample size might not be an integer.*)

C5) For the values of the hyperparameters that you chose in Section 1, what is the sample size of this fictional prior sample? Do you think this is too large, regarding how sure you felt at that time?

C6) The uniform distribution, a special case of the Beta distribution when $a = b = 1$, is called *non-informative* because it is considered to provide very little prior information. What prior sample size would this distribution have? Do you think this is small enough to justify calling it non-informative?

C7) What (non-negative) values of a and b result in a prior sample of size 0? What would the posterior mean be equal to in this case?

C8) What is the issue with using the distribution characterized by the values of a and b in question C7)?
(*Hint: recall that the density of a Beta distribution for p is given by* $f(p;\, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$.)

4 Your turn: analyzing FedEx on-time deliveries

In this part of the activity, you will pretend to work at a bookstore that receives their book orders via FedEx. You will be studying the proportion of packages that are delivered on time. Note that the true on-time delivery rate might be known by FedEx, but they do not make this information public. You will pretend to have placed multiple book orders, some of which were delivered on time and some were not. You will use this information to report to the bookstore manager how likely are future orders to arrive on time.

A1) What values can each datum take? To what event would each of these values correspond to?

A2) What is the distribution of each datum? (Assume each book order was placed independently.)

A3) What is the parameter of interest?

A4) Briefly explain the relationship between the parameter of interest and the data.

A5) Without looking at any plot in the resource, what do you think is the average on-time delivery rate of FedEx? Are you more certain about this guess than you were about the admissions rate of UC Berkeley?

A6) Go to the *Prior distribution* panel of the resource. Set the values of the hyperparameters so that the prior distribution reflects your belief about the on-time delivery rate of FedEx.
Hint: change the settings and observe how the curve in the plot changes. Use this curve, along with the displayed prior mean and variance, to inform your choice.

A7) Suppose that you placed 15 book orders, of which 13 arrived on time. In your sample, what is the value of $\hat{p}$? Calculate an approximate 80% confidence interval for p.

A8) Go to the *Data and likelihood* panel of the resource. Modify the controls on the left tab of the resource to reflect the data that you observed.
Hint: make sure to select Custom data.

A9) Using the *Posterior analysis* and *Summary* panels of the resource, report the following:

6.1) The posterior mean and variance of p.

6.2) An 80% credible interval for p.

6.3) The posterior distribution.

A10) Compare the width of the confidence and credible intervals. Which one is longer?

A11) Write a brief summary of your analysis. It should include information on the sample you observed, your prior beliefs about p, and a description of the properties of the posterior distribution. Imagine this is the report that you will hand in to the bookstore manager—they should be able to understand it. You can assume they know basic statistics, but not Bayesian inference.

A12) Do you think your results would be the same for other bookstores in the city? What about during Christmas time? Briefly explain.

A13) According to an article by Supply Chain Dive[3]—a popular business journal—the on-time delivery rate of FedEx in May 2020 was 91%. Did the credible interval capture this value? What about the approximate confidence interval?

[3] Leonard, Matt. "FedEx, UPS on-time performance falls in May." *Supply Chain Dive*, June 15, 2020. `https://www.supplychaindive.com/news/fedex-ups-on-time-performance-delivery-may/579785/`. Accessed October 30, 2020.

Finally, we will study the impact that different prior distributions can have on the data analysis.

B1) What would have happened if your prior and your data were very different? Answer this question by first going to the *Prior distribution* panel of the resource and setting $a = 2$ and $b = 10$. Write down the prior mean and variance for this choice of prior distribution.

B2) Do you think this prior distribution reflects too much confidence about the value of p?

B3) Calculate the same quantities as in question A9) using the *Posterior analysis* and *Summary* panels of the resource. What is the difference between the two posterior means? Which prior distribution yields a longer credible interval? Which posterior distribution has a larger variance?

B4) Go to the *Summary* panel of the resource. Referring to the plot on the right side, how likely are the values of p from the credible interval under the prior distribution? Under the likelihood?

B5) Based on questions B1) to B4), briefly explain the impact that the prior distribution can have on the posterior distribution.