**Defining a Bayesian prior distribution**

**Prerequisite knowledge**

To complete this activity, you will need to

* be familiar with discrete random variables and its support;
* know how to create a histogram from a table of the probabilities of a discrete random variable;
* be familiar with prior distributions in the context of Bayesian statistics.

**Learning outcomes**

After completing this activity, you should be able to

* specify a prior distribution for a parameter that takes on a discrete number of possible values based on your knowledge of the problem;
* explain how varying degrees of knowledge about a problem are expressed by different degrees of uncertainty in the prior;
* compare and contrast the degrees of uncertainty represented by two different prior distributions based on their histograms.

In this activity, we will guide you in creating your own Bayesian prior distribution for a discrete parameter.

**Problem setting**

We are interested in the average number of books read by Vancouverites (i.e., residents of Vancouver) in 2021. This number, which we will denote μV, is defined as

|  |  |  |
| --- | --- | --- |
| μV = | total number of books read by all Vancouverites in 2021 | . |
| population of Vancouver |

We will assume that μV ≤ 15 (i.e., that the average number of books read by Vancouverites in 2021 does not exceed 15). To make things simpler, we will also instead work with a *discretized* version of μV, denoted *M*V, that takes as possible values one of five different labels describing μV:

* “*Very low*”,that is, 0 ≤ μV < 3;
* “*Low*”, that is, 3 ≤ μV < 6;
* “*Medium*”, that is, 6 ≤ μV < 9;
* “*High*”, that is, 9 ≤ μV < 12;
* “*Very high*”, that is, 12 ≤ μV ≤ 15.

As we do not know μVin practice, we do not know *M*V (the discrete version of μV). We will specify a prior distribution for *M*V that captures our beliefs and uncertainty about *M*V.

**Specifying your prior distribution**

To help you specify your own prior distribution for *M*V, start by answering the following questions:

1. How many books do you read on average in any given year?
2. Do you think this is less than or above average in Vancouver? How certain are you of your answer?
3. Referring to the labels above (*Very low*, …, *Very high*), which label do you think most likely describes the true average number of books read by Vancouverites in 2021?
4. What about the second most likely label? And the least likely?

Now we are going to assign probabilities to each possible label. To make things easier, we will first give each label a “score” between 0 and 10, where 0 is “very unlikely” and 10 is “very likely.”

1. Using Table 1 at the end of this section, assign a score between 0 and 10 to each label.
2. Add up the scores of the five labels and write the total at the bottom of the second column of Table 1. You can use a calculator.
3. Finally, in the third column, divide the score of each label by the sum of all scores. Each label should be a number between 0 and 1, and the column should add up to 1.

|  |  |  |
| --- | --- | --- |
| Label | Score | Score / Sum of scores |
| *Very low*, i.e., [0,3) |  |  |
| *Low*, i.e., [3,6) |  |  |
| *Medium*, i.e., [6,9) |  |  |
| *High*, i.e., [9,12) |  |  |
| *Very high*, i.e., [12,15] |  |  |
|  | Sum of scores: | 1.0 |

*Table 1*

The result of question 7 (i.e., the third column) is your prior distribution for *M*V. Now let’s understand what this means. Recall that our parameter of interest is the label that describes the average number of books read by Vancouverites in 2021.

1. The prior is a distribution over the *support* of *M*V (i.e., the possible valuesof *M*V). Which of the following correctly describes the support of *M*V?
   1. {*Very low*, *Low*, *Medium*, *High*, *Very high*}*;*
   2. *R+*, the positive real numbers;
   3. *N*, the natural numbers (including 0);
   4. *{**0,1,…,14,15}*, the natural numbers up to 15;
   5. *[0,15]*, the interval of real numbers between 0 and 15.
2. Which column in your table represents your prior probability that *M*V = *m*, i.e., *P*(*M*V = *m*), where *m* is any possible value of *M*V?
3. Draw a histogram of your prior distribution. The *x*-axis should contain each label and the *y*-axis should contain *P*(*M*V = *m)*. You can use the accompanying Shiny web app available at:

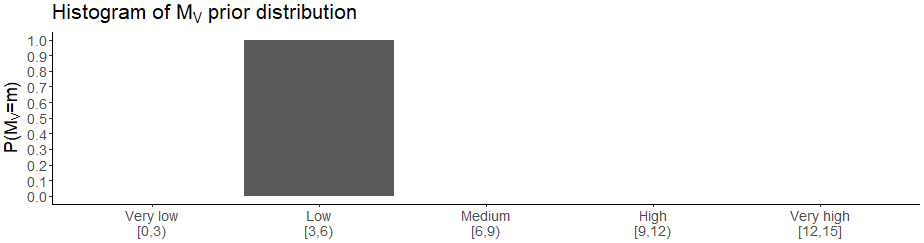
<https://shiny-apps.stat.ubc.ca/FlexibleLearning/FirstBayes/Histogram/>

1. Summarise your prior beliefs in words, using the table and the histogram as a reference.
2. Suppose that we select one Vancouverite at random from Vancouver’s population. Under our assumptions, can that Vancouverite have read over 15 books in 2021?

**Others’ prior distribution**

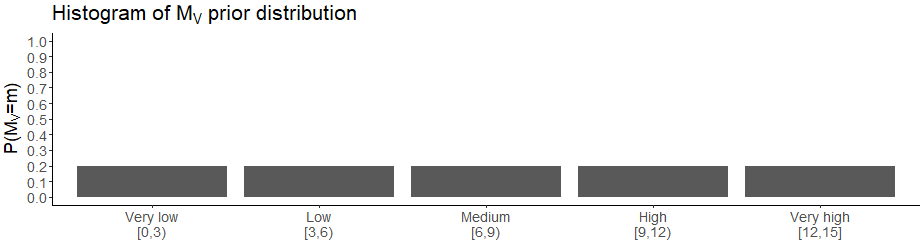
Let’s try to understand our prior distribution better by considering what other histograms we could have obtained through the exercise in the previous section.

Suppose that Lily, a librarian at the Vancouver Public Library, has been observing the reading habits of Vancouverites since 2010. When asked for her scores of the labels, Lily gives a score of 10 to *Low* and a score of 0 to all other labels. Her resulting histogram is shown in Figure 1.

*Figure 1*

1. Summarise Lily’s prior beliefs by referencing her histogram and using words like certainty.

Now suppose that we ask the same questions to Thiago, a tourist from Mexico visiting Vancouver for the first time. When asked for his scores, Thiago gives a score of 4 to all labels. His resulting histogram is shown in Figure 2.

*Figure 2*

1. Summarise Thiago’s prior beliefs by referencing his histogram and using words like uncertainty.
2. Describe how Lily and Thiago’s histograms differ in terms of their shape. Can you relate the shape of their histograms to their prior beliefs?
3. Refer back to the histogram you created in question 10. Does your histogram more closely resemble Lily’s histogram, Thiago’s histogram, or some histogram in between? What does your histogram say about your own level of uncertainty?
4. Refer back to your answer for question 11. Do you think your summary of your prior beliefs still reflects the histogram you created in question 10? If not, adjust your answer for question 11.

**Defining a prior for a different city**

Now suppose we are interested in the label that describes the average number of books read in 2021 by residents of Guanajuato, a city north of Mexico City in Mexico. We will denote this parameter by *MG*, and we will repeat most of the questions from the previous section.

1. Recall how many books you read on average each year from question 1. Do you think you read, on average, more or less than residents of Guanajuato? Are you more or less certain than you were in question 2?

*(Hint: how much do you know about reading habits of the residents of Guanajuato?)*

1. Referring to the labels defined in the first section, which label do you think most likely describes the true average number of books read by residents of Guanajuato in 2021?
2. What about the second most likely label? And the least likely?
3. How hard was picking these labels compared to picking them for Vancouver?

Just as with Vancouver, now we are again going to assign probabilities to each possible label. To make things easier, we will first give each label a “score” between 0 and 10, where 0 is “very unlikely” and 10 is “very likely.”

1. Using Table 2 at the end of this section, assign a score to each label.

*(Hint: you may assign the same score to multiple labels if you think the labels equally likely describe the true value of MG.)*

1. Add up all the scores of the five labels and write the total at the bottom of the second column of Table 2. You can use a calculator.
2. Finally, divide the score of each label by the sum of all scores and write the result for each label in the third column.

The result of question 24 (i.e., the third column), is your prior distribution for the label describing the average number of books read by residents of Guanajuato in 2021, *MG*.

1. Draw a histogram of your prior distribution for *MG*. The *x*-axis should contain each label and the *y*-axis should contain *P*(*MG* = *m*). You can use the accompanying Shiny web app available at:

<https://shiny-apps.stat.ubc.ca/FlexibleLearning/FirstBayes/Histogram/>

1. Summarise your prior beliefs for *MG* in words, using the table and the histogram as a reference.

|  |  |  |
| --- | --- | --- |
| Label | Score | Score / Sum of scores |
| *Very low*, i.e., [0,3) |  |  |
| *Low*, i.e., [3,6) |  |  |
| *Medium*, i.e., [6,9) |  |  |
| *High*, i.e., [9,12) |  |  |
| *Very high*, i.e., [12,15] |  |  |
|  | Sum of scores: | 1.0 |

*Table 2*

**Comparing Vancouver and Guanajuato**

We are now going to compare the prior distributions that you specified for *MV* and *MG*. Refer to the histograms and the tables you created in the previous sections. Feel free to use words like spread, symmetry, and location.

1. Would you say that you have less, about the same, or more uncertainty about *MG* (the label describing the average number of books read by residents of Guanajuato in 2021) than about *MV* (the same thing, but for Vancouverites)?
2. Compare the shape of the histogram of *MV* that you created in question 10 with the shape of the histogram of *MG* from question 25. Is your answer for question 27 reflected in the shapes of each histogram? If so, how?
3. Are the prior distributions that you specified for *MV* and for *MG* very different? If so, what were the differences between the two problem contexts that led to this difference? If not, why were the differences between the two problem contexts irrelevant for specifying the prior distributions?