Web Activity: Bayesian Analysis of Normally-distributed Data

Prerequisites:

- Be familiar with the probability concepts of a random variable and conditional probability.
- Know that, when making inference for Normally-distributed data with known variance σ^2, a random sample is used to estimate the value of the mean, μ.
- Be able to calculate and interpret confidence intervals for the mean of Normally-distributed data when the variance is known.
- 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 Normally-distributed data with mean μ and known variance σ^2:

- Given data and a Normal prior for μ, calculate by hand from a formula the parameters of the posterior distribution of μ.
- Given data and a Normal prior for μ, order, from largest to smallest, the prior mean of μ, the sample mean, and the posterior mean of μ.
- Given data and a Normal prior for μ, obtain the posterior mean and variance of μ using an online interactive resource.
- Compare and contrast the interpretation of a frequentist $(1-\alpha)\times 100\%$ confidence interval for μ with a Bayesian $(1-\alpha)\times 100\%$ credible interval for μ.
- 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 Normally-distributed data using an online interactive resource.

For independent data $X_1, ..., X_n$, where each datum follows a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is known, recall that the frequentist estimator of μ is given by the sample mean:

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

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\hat{\mu} \pm z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}},$$

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

In the Bayesian paradigm, the parameter μ is treated as a random variable. The initial uncertainty about μ is reflected through a user-specified distribution called the *prior distribution*. When the variance is known, a commonly used prior for μ is a Normal distribution, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. The parameters μ_0 and σ_0^2, chosen by the practitioner to reflect their prior knowledge about μ, are called *hyperparameters*.

Bayesian inference is based on the so-called *posterior distribution*, which is defined as the conditional distribution of μ given the data: $\mu \,|\, X_1, ..., X_n$. Bayes' theorem is used to find the posterior distribution. In the case of Normally-distributed data, and when the prior is a Normal distribution, the posterior is also a Normal distribution: $\mu \,|\, X_1, ..., X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$, where

$$\sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1},$$
$$\mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\hat{\mu}}{\sigma^2} \right).$$

The Bayesian inference process for Normal data with a Normal prior for the mean and known variance can be summarized as follows:

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2), \\ X_1, ..., X_n \,|\, \mu &\sim \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 \text{ known}, \\ \implies \mu \,|\, X_1, ..., X_n &\sim \mathcal{N}(\mu_1, \sigma_1^2). \end{aligned}$$

Observe that σ^2 refers to the variance of the population—which is known—whereas σ_0^2 and σ_1^2 refer to the prior and posterior variance of the mean μ, which in this context is a random variable.

A Bayesian estimator of μ is given by the posterior mean, $\hat{\mu}_{\mathrm{B}} = \mu_1$. As in frequentist statistics, we can construct intervals to quantify our (posterior) uncertainty about μ. Specifically, using the fact that $\mu \,|\, X_1, ..., X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$, we can find numbers c_* and c^* such that

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

The interval $[c_*,\, c^*]$ is called a $(1-\alpha) \times 100\%$ *credible interval* for μ. Recall that you cannot say that the probability of μ being in a confidence interval is $1 - \alpha$. However, the probability of μ being in the credible interval is indeed given by $1 - \alpha$ because in the Bayesian paradigm μ 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/Normal-Normal/`

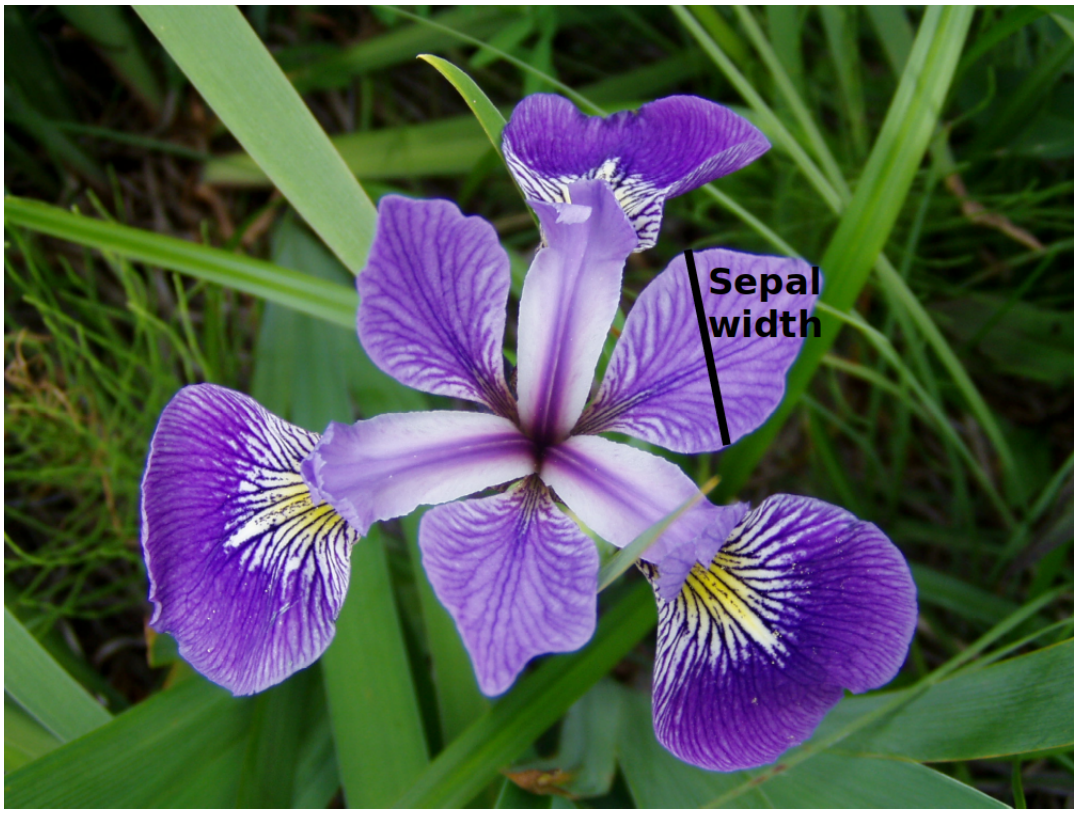
The resource allows you to specify the hyperparameters μ_0 and σ_0^2 of the Normal prior distribution of μ for the data to be used in the analysis. You can select from multiple available data sets, use your own data by providing the observed $\hat{\mu}$, or generate data by specifying the true value of μ. Once you define the prior and the data you are going to work with, the resource calculates the posterior distribution of μ 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.

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

Problem setting

Throughout Sections 1 to 3, we will study a data set which contains information about three species of Iris flowers.[2] Specifically, the data set contains measurements of the sepal widths of a random sample of Irises found in the Gaspé Peninsula in Quebec, Canada. The sepal width is measured at the widest part of the widest petal. See below for a picture of one of these flowers with the sepal width highlighted.

Note that the data are strictly positive, whereas a Normal distribution can also take negative values. However, as you will see in Section 2, the data are sufficiently away from zero and relatively symmetric. In such cases, it is common to use a Normal distribution to model the data, as we will do. We are interested in estimating the average sepal width in centimeters, μ. Before starting the activity, what do you think is the average sepal width in cm of the Irises?



Iris versicolor. Original photo taken by Danielle Langlois in July 2005 at the Forillon National Park of Canada, Quebec, Canada. Downloaded from Wikipedia under a CC BY-SA 3.0 License here. Image modified to show sepal width measurement.

[2]The data set, which has since become a classical data set to analyze, was originally studied in Edgar Anderson (1935). "The irises of the Gaspé Peninsula". *Bulletin of the American Iris Society. 59*: 2–5. See here 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 $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, where μ_0 and σ_0 can be set on the left tab of the resource.

1) In the preamble to the activity, you were asked what you think is the average sepal width in cm of the Iris flower. 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 μ_0 setting in the left tab of the resource until you find a combination with the same mean as what you answered in question 1).

5) Modify the σ_0 parameter to a value that appropriately reflects your current uncertainty.

6) Suppose that instead of Iris flowers, we were to study the sepal width of roses. Would the average sepal width 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 a histogram of the data set. The curve in the lower plot corresponds to the likelihood function.

1) Recall that we are going to analyze the sepal width of Iris flowers. Make the proper selection in the left tab of the resource.

2) Describe the shape of the histogram in the upper plot.

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

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

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

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 μ.

Part A.

A1) Calculate the hyperparameters of the posterior distribution, μ_1 and σ_1^2.
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 μ.

A3) What is the posterior mean of μ? 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 smaller than, equal to, or greater than.*)

A5) The posterior mean is _____$\hat{\mu}$.
(*Fill in the blank with one of smaller than, equal to, or greater than.*)

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

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

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

A9) Considering that μ 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 4).

A11) Contrast the interpretation of both intervals. (Refer to Section 2, question 5).) Which interpretation do you find more intuitive?

A12) What is the posterior variance of μ?

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

A14) Finally, if you had to summarise the posterior distribution with only one number, which one would it be? Briefly explain.

Part B.
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)?

Part C (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.

In the Bayesian analysis of Normally-distributed data, it is common to work with the *precision*, denoted by τ and defined as the multiplicative inverse of the variance: $\tau = \frac{1}{\sigma^2}$.

C1) Using the formula for the hyperparameter update given in the preamble of the activity, prove that the posterior precision $\tau_1 = \frac{1}{\sigma_1^2}$ is a linear combination of the prior precision $\tau_0 = \frac{1}{\sigma_0^2}$ and the population precision $\tau = \frac{1}{\sigma^2}$:

$$\tau_1 = \tau_0 + n\tau. \tag{1}$$

C2) In Equation (1) above, to what value would τ_1 be close to if the sample size n is large?

C3) How is Eqn. (1) related to your answer in question A13)?

C4) Use Eqn. (1) to determine which is larger: σ_1^2 or σ^2/n.

C5) 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:

$$\mu_1 = \frac{\mu_0\tau_0 + n\hat{\mu}\tau}{\tau_1}. \quad (2)$$

C6) How is Eqn. (2) related to your answer in question A6)?

C7) In Equation (2) above, to what value would μ_1 be close to if the sample size n is large?
(*Hint: use your answer to question C2).*)

4 Your turn: analyzing average cycling speeds

In this part of the activity, you will pretend to help the Youth Cycling Committee of the (fictional) City of Aldera to determine if the average speed in km/h of their cyclists has increased. For this purpose, 15 cyclists on the committee will complete a lap around a pre-defined circuit and record their speed. A census in which all cyclists of the committee participated 5 years ago concluded that the average speed in that circuit was 26.5 km/h, with a standard deviation of 1.2 km/h. You will assume that the standard deviation is still 1.2 km/h, and that the speed of each cyclist follows a Normal distribution, to determine whether the average speed is now larger than 26.5 km/h.

A1) Let $X_1, ..., X_{15}$ be the data corresponding to your analysis. What distribution does each datum follow? Define any parameters that you use.

A2) What is the parameter of interest in words?

A3) Without looking at any plot in the resource, what do you think is the average speed of the cyclists? Are you more certain about this guess than you were about the average sepal width of the Iris flowers?

A4) 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 average speed of the cyclists.
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.

A5) Suppose these are the speeds you record:

28.0, 26.5, 28.5, 27.3, 27.6, 26.2, 27.8, 26.0, 25.4, 29.8, 26.9, 27.1, 29.0, 26.8, 25.6.

In your sample, what is $\hat{\mu}$? Calculate a 90% confidence interval for μ.

A6) 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.

A7) Using the *Posterior analysis* and *Summary* panels of the resource, report the following:

A7.1) The posterior mean and variance of μ.

A7.2) A 90% credible interval for μ.

A7.3) The posterior distribution.

A8) Compare the width of the confidence and credible intervals. Which one is longer?

A9) Do you think your results indicate that the average speed has changed since the census 5 years ago?

A10) Write a brief summary of your analysis. It should include information on the sample you observed, your prior beliefs about μ, and a description of the properties of the posterior distribution. Imagine you will hand in the report to the Youth Cycling Committee manager. You can assume they know basic statistics, but not Bayesian inference.You should write it such that the manager will be able to understand it.

A11) Do you think your results would be the same if the cyclists had done a considerably longer circuit? Briefly explain.

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 $\mu_0 = 20$ and $\sigma_0 = 0.3$. Write down the prior distribution for this choice of prior distribution.

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

B3) Calculate the same quantities as in question A7) using the *Posterior analysis* and *Summary* panels of the resource. Comparing the quantities in A7) and B3):

B3.1) What is the difference between the two posterior means?

B3.2) Which prior distribution yields a longer credible interval?

B3.3) 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 μ 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.