1. **Generate scatterplots of petal length *vs.* petal width, and of sepal length *vs.* sepal** **width**. They should look something like this:

A graph of different species

Description automatically generated

We’re going to focus on the relationship between petal length and petal width (the first sub-plot above). We’d like to answer questions like:

* Does the mean petal length differ across species?
* Does the mean petal width differ across species?
* If we were to observe a new flower of a given species, in what length-width region might we expect it to fall?
* How correlated are petal lengths and widths across species?

1. Let’s start with estimating differences in petal lengths across species. Open iris\_bayes\_blanks.R. The first section (“What are the mean petal lengths by species?”) gives us the R code needed to answer this question. **Read through the R code in this block**, and make sure you have a rough understanding of what it’s doing.

Then, **open iris\_petallengths\_blanks.stan and fill in the missing pieces.** Note that there are three code blocks: the data{} block, the parameters{} block, and the model{} block. Everything is given except for the parameters of the model (i.e., the data-generating process). What should these be?

Once you’ve gotten the Stan code written correctly, you should be able to run the rest of the code in this first block of iris\_bayes\_blanks.R. **Look at fig\_iris\_petallengths\_chains. What are they telling you? Do they look “healthy”? Why/why not?**

Last, **look at fig\_postfit\_iris\_petallengths.** It should look like the figure below. **What’s plotted here? What does it tell us about the mean petal lengths across species?**

**A diagram of a graph

Description automatically generated with medium confidence**

1. Now, let’s do the same thing, but for petal widths. Look at the second block of code in iris\_bayes\_blanks.R, called “What are the mean petal widths by species?”. It has the same structure as the previous block, but now there are a lot more blanks for you to fill in. **Fill them in!**

In the end, fig\_postfit\_iris\_petalwidths should look like the following:

A diagram of a graph

Description automatically generated with medium confidence

**What is this figure telling us?** How does it differ from the corresponding figure for petal lengths?

1. We’ve now independently generated descriptions of the petal lengths and petal widths of the three iris species. Let’s examine them together, in the “Some plotting” block of iris\_bayes\_blank.R.

The combined\_summaries data frame combines the posterior summary statistics (mean and bounds of the 95% credible interval) for the parameters we’ve estimated. **Examine this table. Do the entries make sense? What are they telling us?**

Let’s compare the entries of this table with the data. This is done in fig\_pointswithmeans, which should look like this:

A graph with lines and dots

Description automatically generated

**What is this figure showing us? Does it seem like the fits are behaving well?**

The places where the colored lines intersect represent the posterior “means of means”; that is, they mark the posterior mean for the mean petal length and petal width. The shaded bands represent confidence intervals around those means.

However – we’ve plotted the 95% credible intervals for the mean length and the mean width separately. The shaded squares, where the shaded bands intersect, represent something like the 95% credible region for the mean petal length and width – but it’s a bit too big! That’s because the probability of getting a mean length and width in the corner of one of these squares isn’t 5%, but rather 5% \* 5% = 0.25%! How can we identify the 95% credible *region* for where the mean length and width lie?

We can combine our posterior mean length and width estimates and generate a 95% credible *contour* containing these values – it’ll look like an ellipse. The stat\_ellipse() function from ggplot helps us with this.

The data frame iris\_draws\_clean combines the posterior draws for the mean length and width for each species into a data frame. **Check the R code to make sure you understand how the data frame is generated.**

We can use these draws to generate 95% credible ellipses for the mean petal length and petal width. This is done in fig\_pointswithmeans\_ellipse, which should look like this:

A diagram of a graph

Description automatically generated

**How do the ellipses compare to the shaded boxes, where the shaded bands intersect? Why might this make sense? What do the ellipses represent?**

Note that these ellipses don’t contain much of the data – and that’s ok! The ellipses are telling us where the *means* of the petal lengths and widths might sit, but not what the length and width of an individual flower might be. Let’s take a look at this second question: if we were given a new flower of a given species, where, according to the model, might we expect its petal length and width to fall?

This requires us to construct a *prediction interval*. While a *credible interval* tells us about a model parameter, a *prediction interval* combines model parameters to tell us where we might expect new data to fall.

We’ll use a Monte Carlo approach to generate approximate prediction intervals. This is done in the definition of the data frame predbound\_df. To generate this data frame, we first extract the posterior means for (1) the mean petal length/width and the (2) standard deviation of the petal lengths/widths. Then, using these posterior means, we generate 1,000 draws from the *multivariate normal distribution* defined by (1) and (2). Effectively, we have generated simulated data from our model! **Look at predbound\_df and interpret the values in each column.**

Using this “simulated data”, we can draw an ellipse around the most likely 95% of these simulated data points. That’s what stat\_ellipse() does for us. Using stat\_ellipse() lets us generate fig\_predint, which should look like this:

A diagram of a graph

Description automatically generated

Whoa! Those intervals are really big, and there’s a lot of space inside those ellipses that don’t really contain any data! **Why is this? What haven’t we accounted for?**

1. If our goal is to identify regions of the petal length/width space where a new flower of a given species is likely to sit, our job is clearly not done yet. We need to account for the *correlation* between petal lengths and widths. How can we do this?

We can accomplish this by using a *multivariate normal* model for our data. Previously, we assumed that the petal length and petal width were drawn from two independent normal distributions, but this doesn’t account for the fact that flowers with long petals are also likely to have wide petals, and *vice versa*. A multivariate normal distribution helps us account for this correlation.

A multivariate normal distribution is like a normal distribution but in 2 or more dimensions. A 1-dimensional normal distribution looks like this:

A diagram of a function

Description automatically generated

A 2-dimensional (multivariate) normal distribution looks like this:

A green and red grid on a square

Description automatically generated

We can visualize a bivariate normal distribution using contours like this (it’s like we’re looking at the above figure top-down, and the contours represent different elevations on the curve):

A graph of a concentric circle

Description automatically generated

A multivariate normal distribution takes a *vector* as its mean and a *matrix* as its “standard deviation” term (now called a “covariance matrix”):

The covariance matrix must be *symmetric,* meaning that . The elements of the matrix also have interpretations: and give the variance for the first and second element of the outcome, respectively. is the *covariance* between the first and second elements. The *correlation* between the first and second elements is

For the standard multivariate normal distribution (analogous to the standard Normal(0,1) distribution), we have

A draw from a Multivariate Normal distribution is also a vector with the same length as the mean vector (in this case, length 2).

We can introduce *correlation* between the outputs of a multivariate normal distribution by changing the off-diagonal elements of the covariance matrix. For example, a

distribution would give us contours that look like:

A blue line drawing of a planet

Description automatically generated

This looks like it might be a better framework for our data-generating process of combined petal lengths and widths. Let’s generate some Stan code to estimate a mean vector and covariance matrix for a multivariate normal distribution!

Let’s move to the final code block in iris\_bayes\_blanks.R, titled “Try a full multivariate normal fit.” **Look at the code used to generate the list iris\_ex3, and examine the contents of iris\_ex3. What do the different elements of this list represent?**

A template for the Stan code needed to fit a multivariate normal model to the data is given in iris\_petalLW\_blanks.stan. **Open this stan file and fill in the blanks, like you did for the petal lengths file in Step 2. How is the code here similar to what you’ve done in the previous fits? How is it different?**

As in Step 4, we can generate posterior credible regions for the means. This is done in fig\_postmean\_mvn which looks like this:

A graph of different colored dots

Description automatically generated

**How does this figure differ from fig\_pointswithmeans\_ellipse? Why? Do we like the differences?**

Similarly, as in Step 4, we can generate posterior prediction regions. This is done in fig\_postpred\_mvn, which looks like this:

A diagram of a diagram

Description automatically generated

**How does this figure differ from fig\_predint? Why? Do we like the differences?**

The final thing we’ll do in this exercise is ask: is there evidence for correlation between the petal lengths and widths for each of the three species?

The posterior correlation coefficients are calculated in pos\_corr\_df; **you’ll need to fill in the blank on the final line, which computes the correlation using the columns of the data frame generated up to that point. To do this, use the mathematical expression for the correlation () from a few pages back.**

These correlation coefficients are plotted in fig\_post\_corr, which looks like this:

A diagram of a normal distribution

Description automatically generated

**What does this figure tell us about the correlation between petal length and petal width for each species?**