Mini Project

Katherine Muller

2026-01-06

The purpose of this exercise is for you to implement the steps of a Bayesian data analysis using a relatively simple example.

1. Describe the study system.
2. Inspect the data.
3. Draw a causal diagram.
4. Define one or more research questions.
5. Define one or more specific variables that address a research question.
6. Design and implement a statistical model.
7. Obtain results.

Study system and data

For this exercise we will be using the ilri.sheep dataset from the agridat package.

First, install the package if you haven't already.

```
install.packages("agridat") # run this line
```

Next, load the package.

```
library(agridat)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggforce)
```

Read the documentation for the ilri.sheep dataset.

```
?ilri.sheep
```

Define one or more broad research questions or topics that could be addressed with this dataset.

You will refine these questions after you draw a conceptual model.

FILL IN: Examples: How does lamb genotype affect traits of interest? How do Dorpor and Red Maasi lambs differ in traits of interest? How does the parentage affect hybrid traits (i.e., Dorpor father and Red Maasi mother and vice versa)? How does sex affect traits of interest? How does environment (year) affect traits of interest? How do the effects of sex interact with the effects of genotype? What is the heritability of traits of interest?

Draw a causal diagram

Use your scientific reasoning to think though cause-and-effect relationships between variables in the ilri.sheep dataset. Denote variables as nodes and causal relationships as arrows.

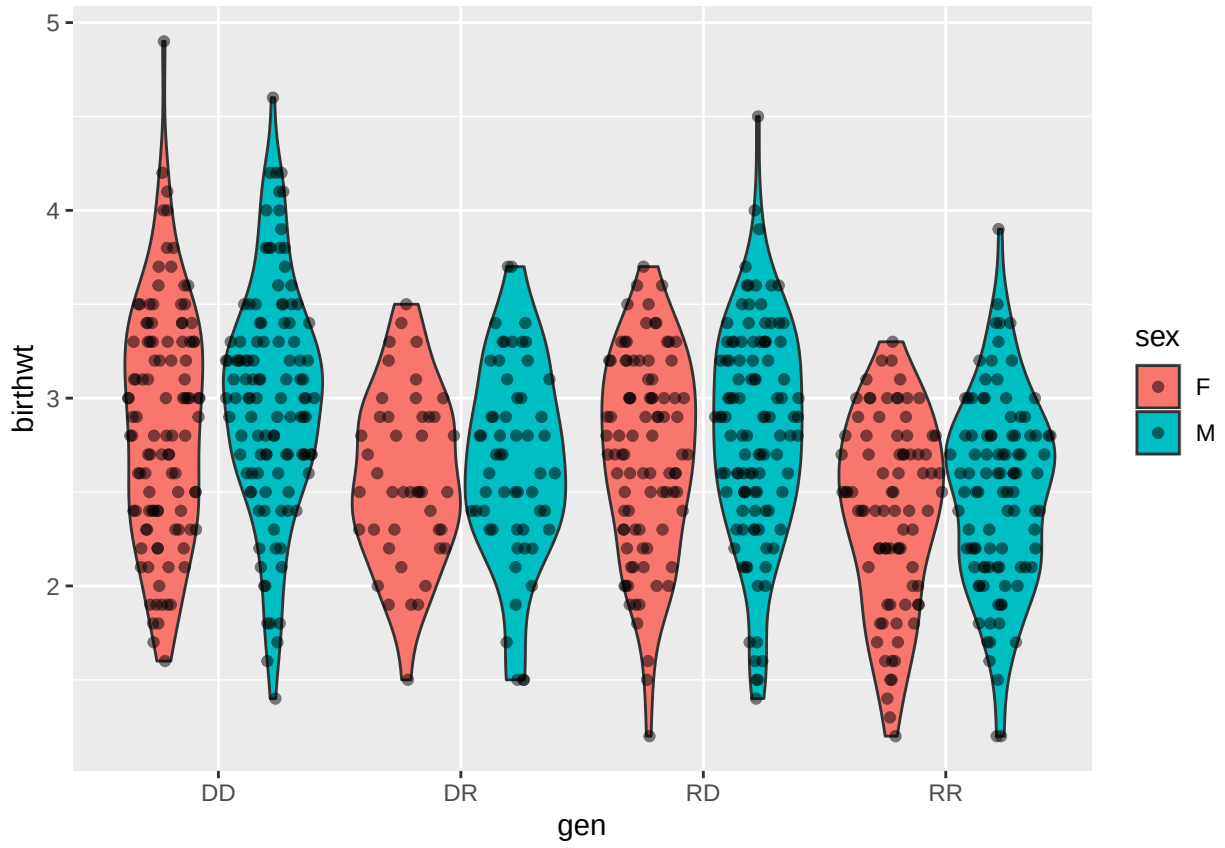
Exploratory data visualization

Before building any statistical models, visualize your data. Use your exploratory questions as a guide.

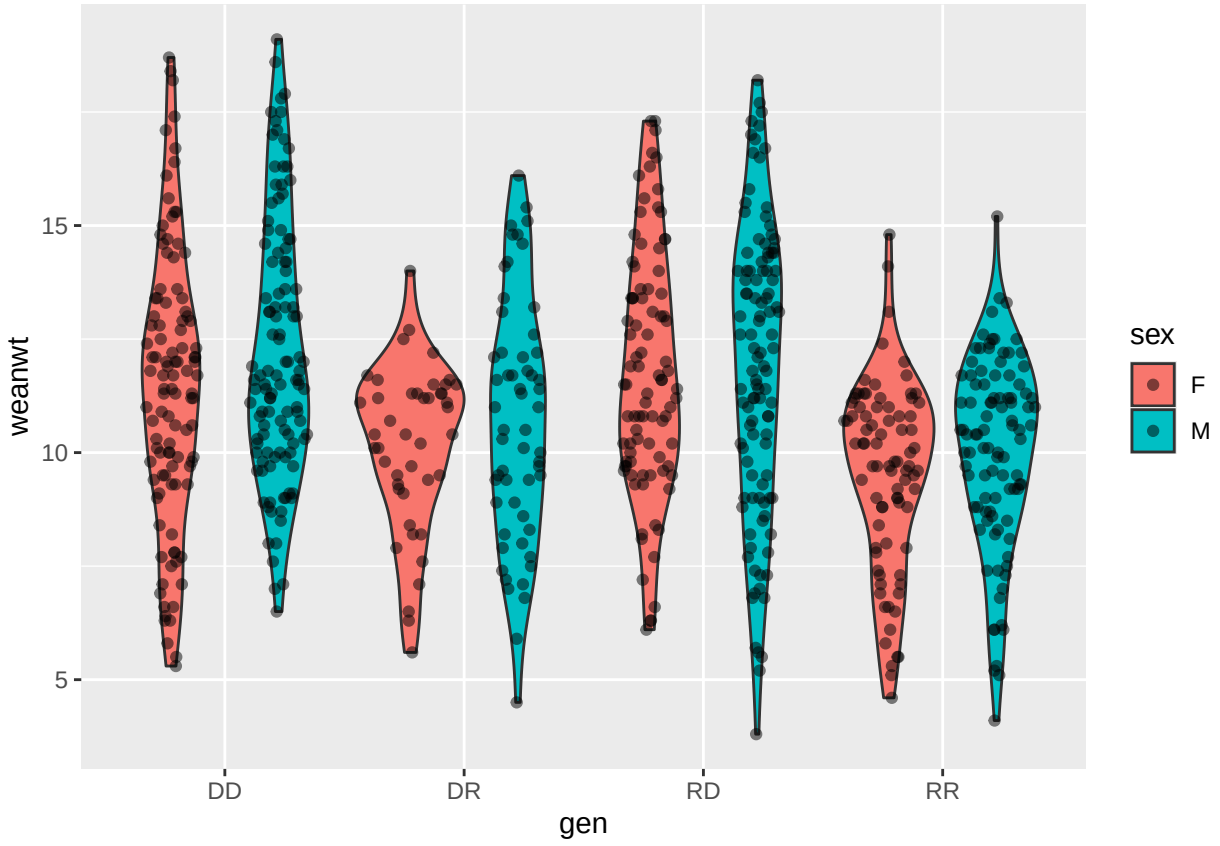
What do you want to see? Try sketching some plots on a piece of paper.

At this point, we will work together as a group to code some plots. The

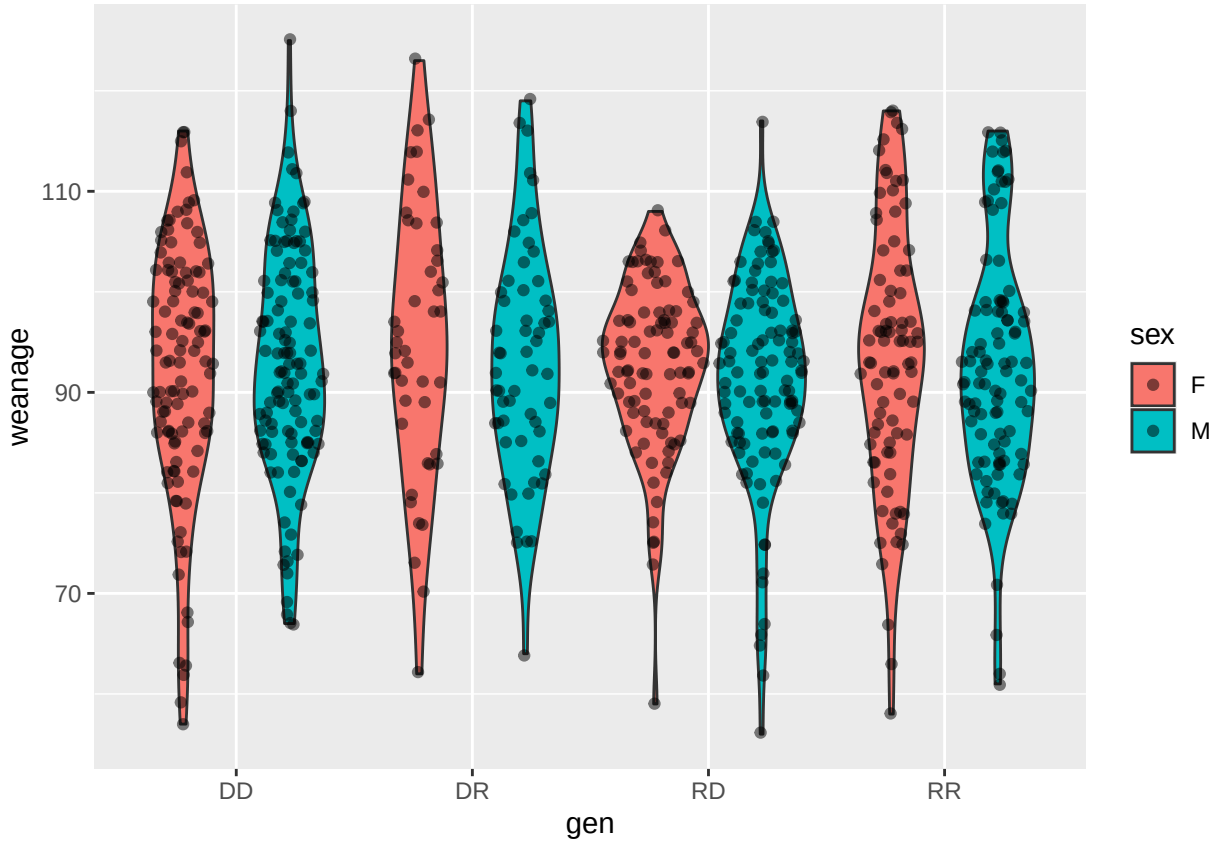
```
data(ilri.sheep) # load data
sheep <- ilri.sheep

# sheep %>%
#   drop_na() %>%
#   ggplot(aes(x = gen, y = birthwt, fill = sex)) +
#   geom_violin()+
#   geom_point(position = position_jitterdodge())

sheep %>%
  drop_na() %>%
  ggplot(aes(x = gen, y = birthwt, fill = sex)) +
  geom_violin()+
  ggforce::geom_sina(alpha= 0.5)
```



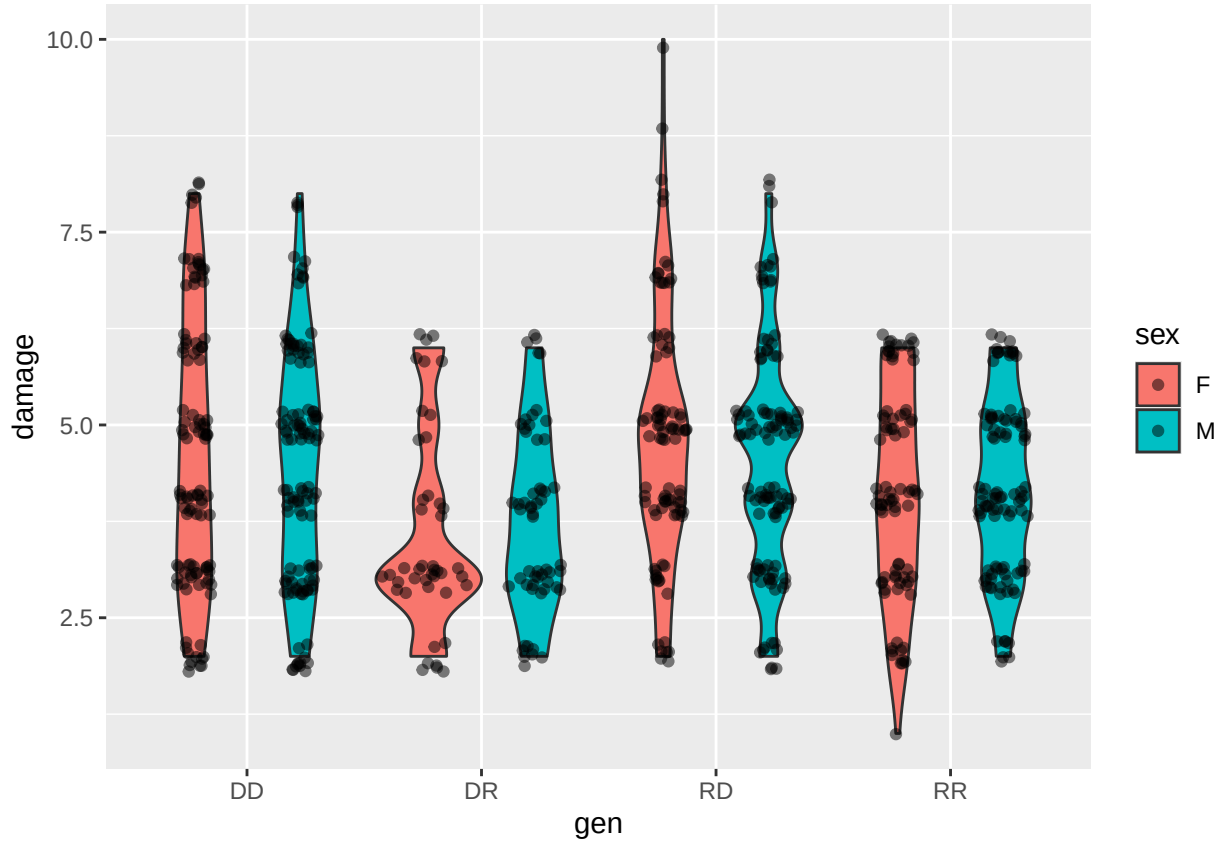

```
sheep %>%
  drop_na() %>%
  ggplot(aes(x = gen, y = weanwt, fill = sex)) +
  geom_violin()+
  ggforce::geom_sina(alpha= 0.5)
```




```
sheep %>%
  drop_na() %>%
  ggplot(aes(x = gen, y = weanage, fill = sex)) +
  geom_violin()+
  ggforce::geom_sina(alpha= 0.5)
```




```
sheep %>%
  drop_na() %>%
  ggplot(aes(x = gen, y = damage, fill = sex)) +
  geom_violin()+
  ggforce::geom_sina(alpha= 0.5)
```



Describe some patterns you can see from your exploratory visualization

Develop a set of hypotheses and/or predictions that we can evaluate with a statistical model

Prepare the data

Before you do any kind of analysis, you will need to make sure that the data are in a form you can work with. When importing a data table into R, R guesses what type of data is in each column based on its contents. The two main data types are character and numeric. Continuous variables should be numbers (type num, int, or dbl). If the column contains any non-numeric string, R will read it as a character. For example, if I write "missing" for missing values in a numeric column, R will interpret the whole column as characters.

Are the numbers numbers?

Based on your knowledge of the dataset, identify the columns that should be **continuous**. Verify that R has read them as numeric (num, int, or dbl).

```
str(sheep)
```

```
## 'data.frame':    882 obs. of  12 variables:
##  $ year   : int  91 91 91 91 91 91 91 91 91 91 ...
##  $ lamb   : int  627 629 635 636 638 639 640 642 643 644 ...
##  $ sex    : Factor w/ 2 levels "F","M": 2 1 2 2 1 1 2 2 2 1 ...
```

```
##  $ gen    : Factor w/ 4 levels "DD","DR","RD",..: 1 1 1 1 1 1 3 1 1 1 ...
##  $ birthwt: num  2.7 2.9 2.5 2.7 3 2.4 3.4 2.5 3.8 2.5 ...
##  $ weanwt : num  16.3 18.4 14.7 15.6 NA 10.8 15.5 NA 19.1 11.4 ...
##  $ weanage: int  125 112 109 108 NA 107 107 NA 107 107 ...
##  $ ewe    : int  1682 1082 1520 1450 5183 1471 1116 5138 1169 1595 ...
##  $ ewegen : Factor w/ 2 levels "D","R": 1 1 1 1 1 1 1 1 1 1 ...
##  $ damage : int  2 5 2 5 3 2 4 3 5 2 ...
##  $ ram    : int  1980 4908 1974 4911 4909 1973 1981 4909 1973 4910 ...
##  $ ramgen : Factor w/ 2 levels "D","R": 1 1 1 1 1 1 2 1 1 1 ...
```

Are the categories factors?

For analysis, it can be helpful to convert categorical variables into factors. For example, in the sheep dataset, the variable sex is a factor with two levels, F and M. This means that R interprets this variable as an integer with F = 1 and M = 2. You can specify how to order the factor levels, but alphabetical is the default. Because R interprets factors as integers, you have to be careful with them if you have lots of data processing steps. Here we will use factors for categorical variables to help with analysis.

Based on your knowledge of the dataset, identify the columns that should be **categorical**. In the sheep data, several categorical variables have already been converted to factors.

```
str(sheep)
```

```
## 'data.frame':    882 obs. of  12 variables:
##  $ year   : int  91 91 91 91 91 91 91 91 91 91 ...
##  $ lamb   : int  627 629 635 636 638 639 640 642 643 644 ...
##  $ sex    : Factor w/ 2 levels "F","M": 2 1 2 2 1 1 2 2 2 1 ...
##  $ gen    : Factor w/ 4 levels "DD","DR","RD",..: 1 1 1 1 1 1 3 1 1 1 ...
##  $ birthwt: num  2.7 2.9 2.5 2.7 3 2.4 3.4 2.5 3.8 2.5 ...
##  $ weanwt : num  16.3 18.4 14.7 15.6 NA 10.8 15.5 NA 19.1 11.4 ...
##  $ weanage: int  125 112 109 108 NA 107 107 NA 107 107 ...
##  $ ewe    : int  1682 1082 1520 1450 5183 1471 1116 5138 1169 1595 ...
##  $ ewegen : Factor w/ 2 levels "D","R": 1 1 1 1 1 1 1 1 1 1 ...
##  $ damage : int  2 5 2 5 3 2 4 3 5 2 ...
##  $ ram    : int  1980 4908 1974 4911 4909 1973 1981 4909 1973 4910 ...
##  $ ramgen : Factor w/ 2 levels "D","R": 1 1 1 1 1 1 2 1 1 1 ...
```

There are several categorical variables that are coded using integers. Convert these to factors.

First, transform data so it is easier to model.

Standardize continuous metrics (subtract the mean and divide by standard deviation).

```
## the scale function standardizes data. These options are the defaults.
sheep <- sheep %>% mutate(std_birthwt = scale(birthwt,center = TRUE, scale=TRUE)[,1],
                          std_weanwt = scale(weanwt, center = TRUE, scale = TRUE)[,1],
                          std_weanage = scale(weanage, center = TRUE, scale = TRUE)[,1],
                          std_damage = scale(damage, center = TRUE, scale = TRUE)[,1])
```

We're going to start with a frequentist linear modeling tool...

Design a model