Stats 60: Learning stats via simulation

Stats 60 TAs

2017-01-01

Contents

Chapter 1

Prerequisites

Chapter 2

Simulation

Chapter 3

Descriptive statistics and data vizualization

As an analyst, often our first goal is to **describe** a data set. What we mean by “describe” here is something like: provide an efficient represenation of the data that is easy for another person to understand. Two effective tools for this task are *descriptive statistics* and *data visualization.* It is important to emphasize that the task here is to summarize and communicate with other humans, which means that you need to do more than just figure out how to do the computations. In his statistics book, ognitive scientist, Dan Navarro, has a really nice paragraph on the philosophy of descriptive statistics, so I thought I would include it here:

> Thus it is no small thing to say that the first task of the statistician and the scientist is to summarise the data, to find some collection of numbers that can convey to an audience a sense of what has happened. This is the job of descriptive statistics, but it’s not a job that can be told solely using the numbers. You are a data analyst, not a statistical software package. Part of your job is to take these statistics and turn them into a description. When you analyse data, it is not sufficient to list off a collection of numbers. Always remember that what you’re really trying to do is communicate with a human audience. The numbers are important, but they need to be put together into a meaningful story that your audience can interpret. That means you need to think about framing. You need to think about context. And you need to think about the individual events that your statistics are summarising

With that framing in mind, let’s dive into some common statistical techniques that we can use to efficiently describe our data.

3.1 Measures of central tendency

Answers the question: Where are the data? What is the long-run average value of repetitions of the same experiment or data-generating process?

3.1.1 Mean

For a data set, the mean provides information about the “central tendency” or the “center of mass” of the data, and is typically denoted using the symbol $\bar{x}$. Note that if our data set consists of random samples from a larger population, we need to be careful to limit the use of the mean to describe our *sample* since the population mean (typically denoted μ) is different.

To calculate the mean, we take the sum of each value in our data set and then divide by the number of data points. In formal notation this looks like:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

In R, we can quickly compute the mean using the built-in function `mean()`. Here we are using the Iris data set, which comes with your R installation. The data include a set of measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

```
m_plength <- mean(d$Petal.Length)
m_plength
```

```
## [1] 3.758
```

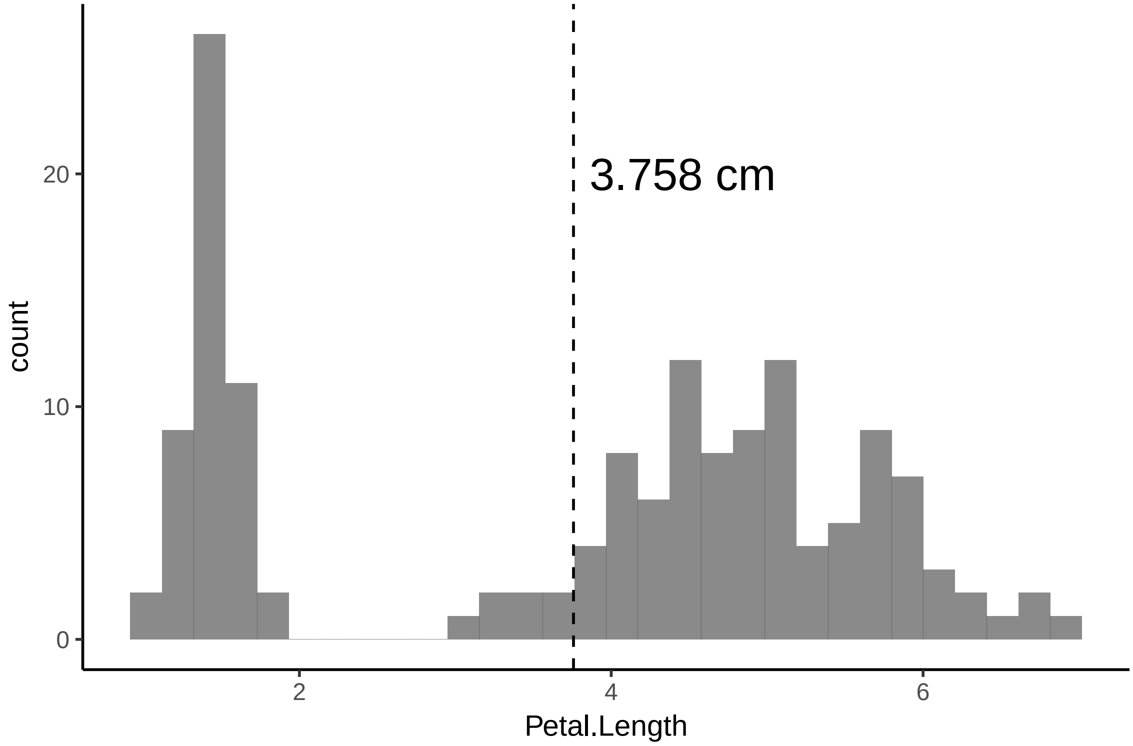
Note that we could also use our dplyr skills to compute the mean like this.

```
d %>% summarise(m = mean(Petal.Length))
```

```
##       m
## 1 3.758
```

Let's practice our data viz skills and plot the distribution of petal lengths and include the mean as a vertical dashed line.

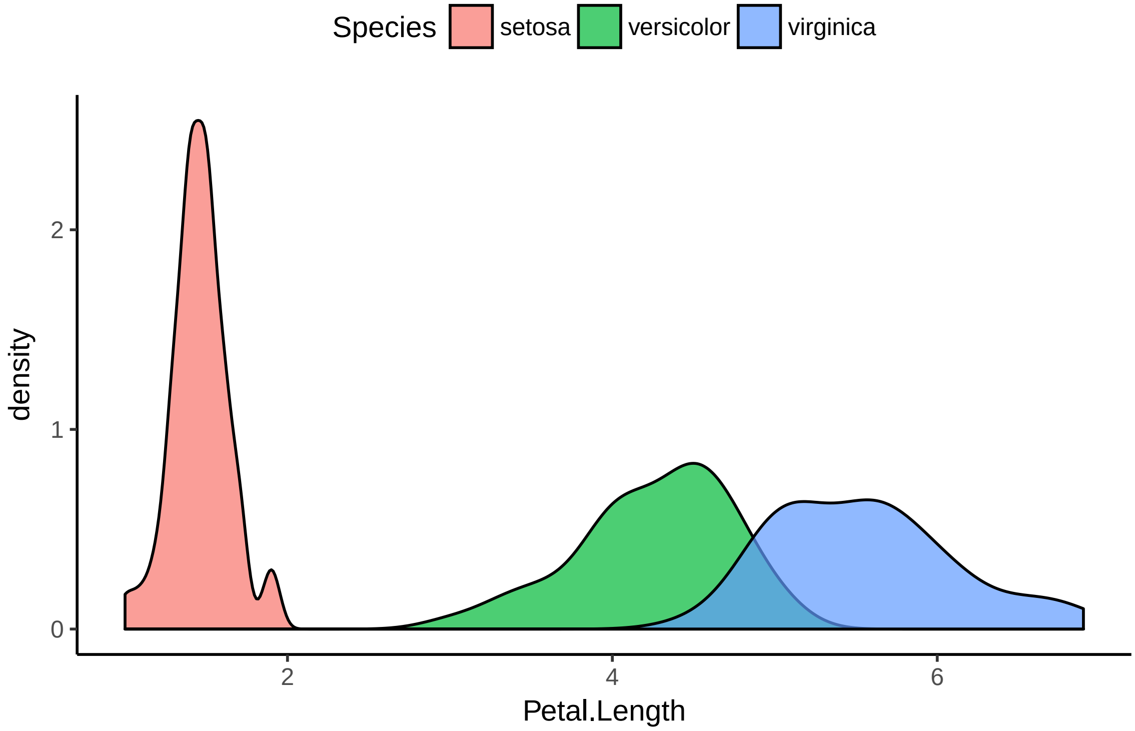
```
d %>%
  ggplot(aes(x = Petal.Length)) +
  geom_histogram(alpha = 0.7) +
  geom_vline(xintercept = m_plength, linetype = "dashed") +
  annotate(geom = "text", x = m_plength + .7, y = 20,
           label = paste(m_plength, "cm"), size = 6)
```



What do we see? Well, it looks like the mean does tell us about the center location of these data. One way to think about this is is that the mean serves as a "balancig point" of the data distribution, with the number to the left of the mean being balanced by the numbers to the right of the mean.

But, let's return to our original goal of providing a useful description of these data. Is `m_plength` telling us anything useful about these data? Not really. And if we color our plot based on the species of iris, we can see what's going on here.

```
d %>%
  ggplot(aes(x = Petal.Length, fill = Species)) +
  geom_density(alpha = 0.7) +
  theme(legend.position = "top")
```

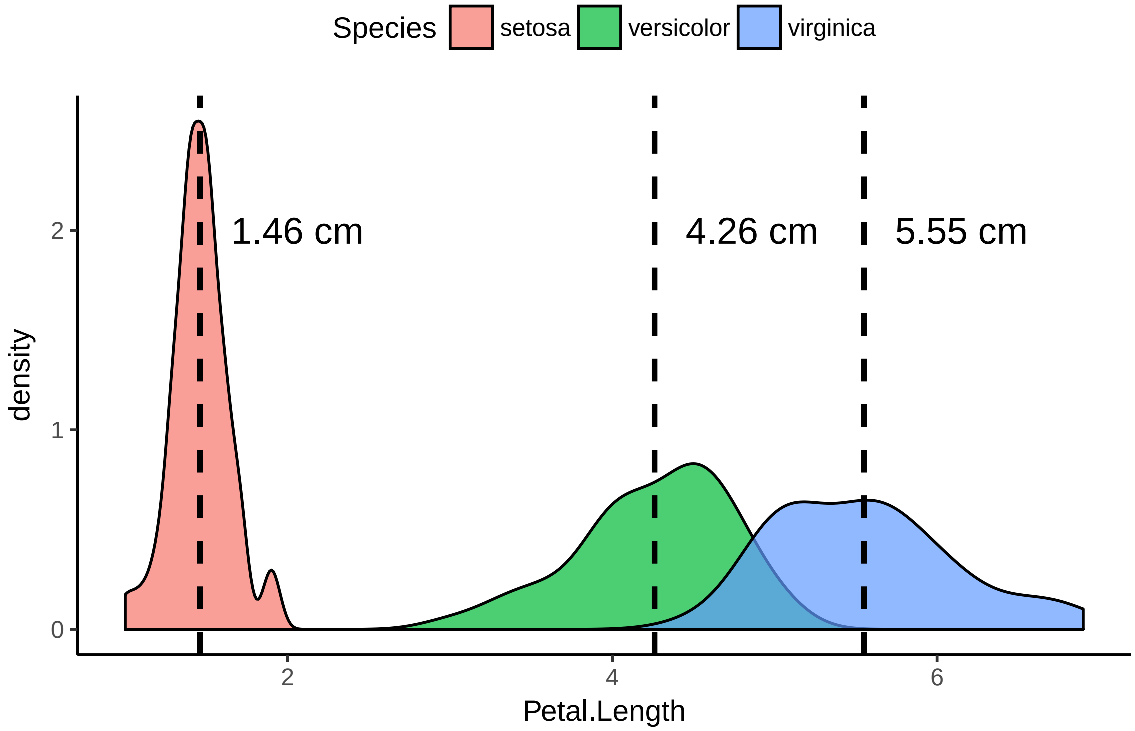


It looks like there are actually three different distributions of petal length in our data set. We can use our dplyr skills to compute the mean for each distribution separately:

```
ms_petal_length <- d %>%
  group_by(Species) %>%
  summarise(m = mean(Petal.Length)) %>%
  mutate(m = round(m, digits = 2))
```

And add that information to our plot:

```
d %>%
  ggplot(aes(x = Petal.Length, fill = Species)) +
  geom_density(alpha = 0.7) +
  geom_vline(aes(xintercept = m), data = ms_petal_length, linetype = "dashed", size = 1) +
  geom_text(aes(label = paste(m, "cm"), x = m + 0.6, y = 2), data = ms_petal_length, size = 5) +
  theme(legend.position = "top")
```



I would argue that these three numbers – the mean for each type of iris species in the data set – give us a more useful description of the central tendency. And you can say something like, “The average petal length for the setosa species is 1.46.”

3.1.2 Median

3.1.3 Mode

3.2 Measures of variability

Answers the question: How spread out are the data?

3.2.1 Range

3.2.2 Interquartile range

The interquartile range (IQR) refers to the middle 50% of a data distribution. It provides a useful way to describe how spread out the data are. To compute the IQR, you divide your data into quartiles or four equal parts. You then subtract the first quartile value from the third quartil value: $IQR = Q3Q1$

Things get a little trickier when we are trying to compute the IQR for continuous-valued distributions. We need to use calculus to integrate the probability density function to get the cumulative distribution function (CDF). The CDF provides the following pieces of information:

- The lower quartile, Q1, is a number such that integral of the PDF from $-\infty$ to Q1 equals 0.25
- The upper quartile, Q3, is such a number that the integral from $-\infty$to Q3 equals 0.75.

3.2.3 Mean Absolute Deviation

The mean absolute deviation (MAD) is another measure of the amount of dispersion in your data. Intuitively, it provides a measure of how far your data are from the mean. Here is the formal definition:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Really, this formula is just a compact way to describe the following algorithm:

1. Compute the mean of the sample
2. Compute the difference between each data point and the mean
3. Add all of those differences
4. Divide by the number of data points in the sample

3.2.4 Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

3.2.5 Standard deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

3.3 Standard scores

Chapter 4

T-test

Bibliography