DATA 505: Lab 3

Descriptive Statistics and Summarizing Data

YOUR NAME HERE

2025-09-09

Table of contents

1 Setup & Goals

Just like in Labs 1 and 2, you will edit a `.qmd` Quarto script, render the result to a PDF, and submit both on Moodle (if you do not succeed at rendering the PDF, still submit the `.qmd` script: you may still get full "satisfactory" credit if you have made a reasonable effort).

Goals for Lab 3:

- Continue to explore the Boston housing dataset
- Use descriptive statistics and simple visual displays to describe the distributions of variables in this dataset

2 The data

- The data for this lab come from the `MASS` package
- This is one of a short list of official recommended R packages on CRAN; as a result it was most likely installed automatically when you installed R. In case it was not, we will do:

```
# install the MASS package if needed
if(!"MASS" %in% installed.packages()){
  install.packages("MASS")
}

# load the MASS package and verify it is now ready for use
stopifnot(require(MASS))
```

```
Loading required package: MASS
```

- To load the data from `MASS` into your environment, the command is

```
data("Boston", package = "MASS")
```

3 Getting acquainted with the data

- Use the `str()` function to print a concise summary of the `Boston` object

```
# YOUR CODE HERE
```

- Look up the help page for the dataset to understand more about the columns (recall: to access help in R, you can write in the console `?` followed by the name of the function or object for which you want documentation)

```
# YOUR CODE HERE
```

4 Summarizing distributions

4.1 Crime rate

- What is the mean per-capita crime rate? What is the median?

```
# YOUR CODE HERE
```

- Which of these descriptive statistics would be a better measure of the crime rate in a typical Boston town? Why?
  - ***YOUR ANSWER HERE***
- Create a histogram of the per-capita crime rate

```
# YOUR CODE HERE
```

- What is the standard deviation of the per-capita crime rate?

```
# YOUR CODE HERE
```

- What are the 25th and 75th percentiles (a.k.a. first quartile (Q1) and third quartile (Q3))? is the interquartile range?

```
# YOUR CODE HERE
```

- Provide a description of the distribution of this variable. Address the shape of the distribution, its center, and its spread.
  - ***YOUR DESCRIPTION HERE***

4.2 Charles-river adjacency

Next we will explore a categorical (binary, in this case) variable. First we will recode the variable as a factor.

```
Boston$charles_river <- factor(
  Boston$chas,
  levels = c(0L, 1L),
  labels = c("Not bounding river", "Bounding Charles river")
)
```

- How many of the tracts are bounding the Charles river? (Hint: use `table()`)

```
# YOUR CODE HERE
```

- What proportion of the tracts are bounding the Charles river? (Hint: use `prop.table()`)

```
# YOUR CODE HERE
```

4.3 Property tax rate (completed example)

The `"tax"` column contains the tract-level average property-tax rate per 10,000 USD of assessed home value.

- Provide a description of the distribution of this variable. Address the shape of the distribution, its center, and its spread. Provide at least one visual display.
  - Tax rates across these tracts have an asymmetric, bimodal distribution, as we can see in the histogram below. The median tract has a tax rate of 330, while the mean tax rate is 408.2. The distribution has large spread: the standard deviation is 168.5, fairly high relative to the man. The middle 50% of tracts have tax rates between 279 and 666.

4.4 Home value (your turn)

The `"medv"` column contains the median home value (in 1000 USD) for homes within each tract.

- Provide a description of the distribution of this variable. Address the shape of the distribution, its center, and its spread. Provide at least one visual display.
  - ***YOUR DESCRIPTION HERE***

5 Challenges

- Compare the home values in Charles-river-adjacent tracts against other tracts which do not bound the river: how do the means compare? Can you create side-by-side box plots to visually compare the distributions of home values in these two groups? Describe, in text, any differences that you see (Hint: refer to `?boxplot`. There are several varieties (a.k.a. S3 methods) of this function, depending on the arguments that are supplied: use the S3 method for class `formula`. Look at the very first example in the examples section for an example of the syntax.)
  - ***YOUR INTERPRETATION***

```
# YOUR CODE HERE
```

- In many situations with asymmetric or otherwise irregular continuous distributions, a **normalizing transformation** may be advantageous as a first step for data analysis. For example, for right-skewed distributions, applying a log transform is often a good step.
  - Take the logarithm of the median home values. Save the result in a new variable. Draw a histogram of this transformed variable, and describe its distribution.
    - ***YOUR DESCRIPTION***

```
# YOUR CODE HERE
```