DATA 505: Lab 3

Descriptive Statistics and Summarizing Data

Solution

2025-09-09

Table of contents

1 Setup & Goals **1**

2 The data **2**

3 Getting acquainted with the data **2**

4 Summarizing distributions **3**

5 Challenges **7**

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

```
str(Boston)
```

```
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

- 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)

```
?Boston
```

4 Summarizing distributions

4.1 Crime rate

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

```
mean(Boston$crim)
```

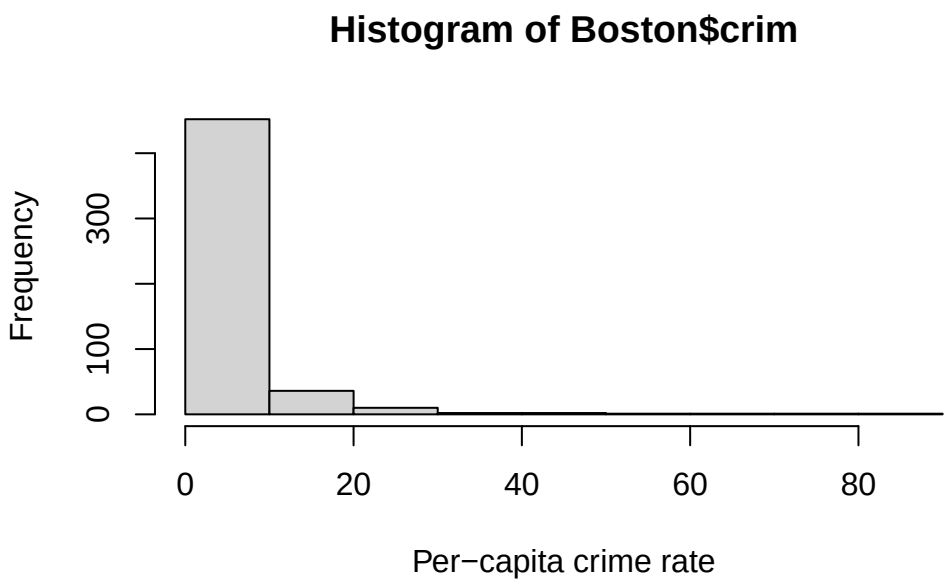
```
[1] 3.613524
```

```
median(Boston$crim)
```

```
[1] 0.25651
```

- Which of these descriptive statistics would be a better measure of the crime rate in a typical Boston town? Why?
  - Median is a better measure of a "typical" / middle value.
- Create a histogram of the per-capita crime rate

```
hist(Boston$crim, xlab = "Per-capita crime rate")
```



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

```
sd(Boston$crim)
```

```
[1] 8.601545
```

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

```
quantile(Boston$crim, c(0.25, 0.75))
```

```
     25%      75%
0.082045 3.677083
```

- Provide a description of the distribution of this variable. Address the shape of the distribution, its center, and its spread.
  - The distribution is unimodal and strongly right skewed. The mode is near zero, with the median crime rate (0.25651) being notably lower than the mean (3.6). There is a high degree of variation: the standard deviation is 8.6. The middle 50% of tracts have crime rates between 0.082045 and 3.6770825.

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()`)

```
table(Boston$charles_river)
```

```
Not bounding river Bounding Charles river
               471                      35
```

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

```
prop.table(table(Boston$charles_river))
```

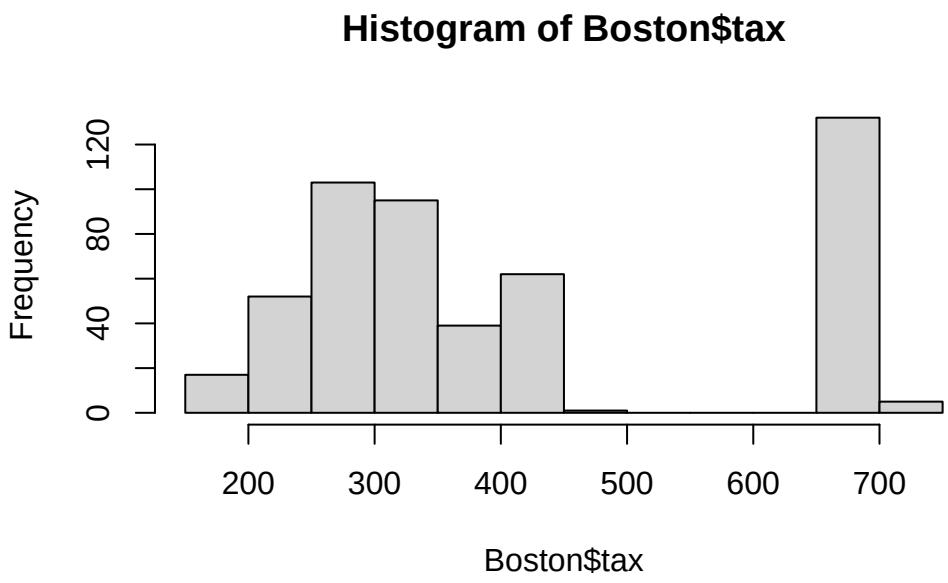
```
Not bounding river Bounding Charles river
        0.93083004             0.06916996
```

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.

```
hist(Boston$tax)
```

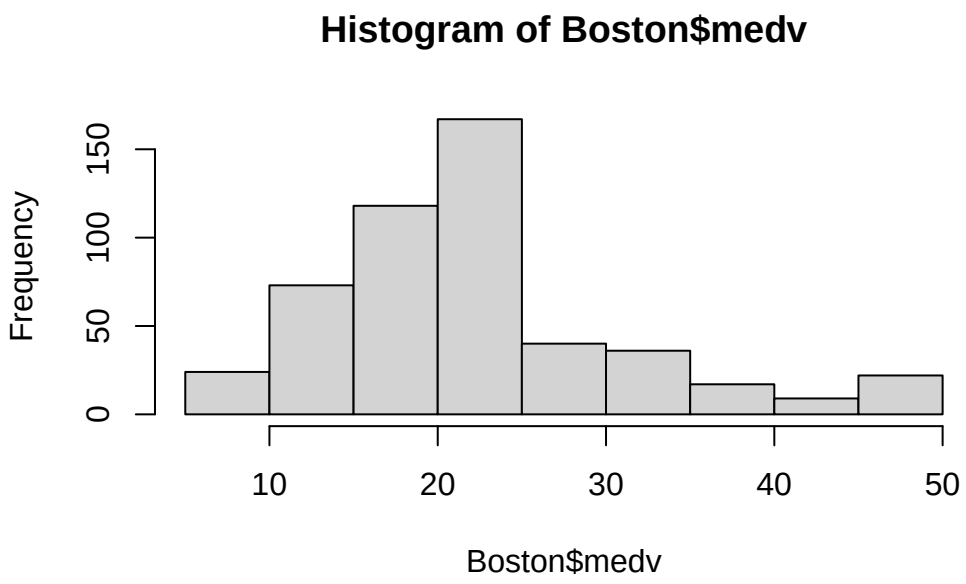


4.4 Home value (your turn)

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.
  - Median home values across these tracts have a unimodal, slightly right-skewed distribution, as we can see in the histogram below. The median tract has a home values of 21.2, while mean across tracts is 22.5. The standard deviation 9.2 indicates a moderate degree of spread relative to the mean. The middle 50% of tracts have home values between 17.025 and 25.

```
hist(Boston$medv)
```



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.)
  - Census tracts that bound the river have higher median home values on average than census tracts which do not bound the river.

```
# comparing means ------------------------------------------------------------

# option 1: tapply()
tapply(Boston$medv, Boston$charles_river, mean)
```

```
    Not bounding river Bounding Charles river
              22.09384                28.44000
```

```r
# option 2: split() and sapply()
sapply(split(Boston$medv, Boston$charles_river), mean)
```

```
    Not bounding river Bounding Charles river
              22.09384                28.44000
```

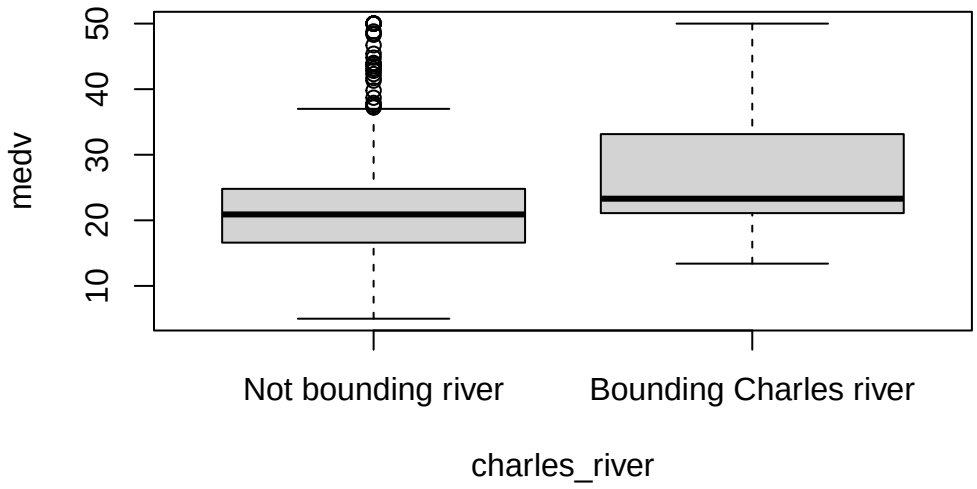
```r
# option 3: subsetting
river_values <- Boston$medv[Boston$charles_river == "Bounding Charles river"]
non_river_values <- Boston$medv[Boston$charles_river != "Bounding Charles river"]

# option 4: dplyr
if(!"dplyr" %in% installed.packages()) install.packages("dplyr")
library(dplyr)
```

```
Attaching package: 'dplyr'

The following object is masked from 'package:MASS':

    select

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
Boston |>
  group_by(charles_river) |>
  summarize(mean_value = mean(medv))
```

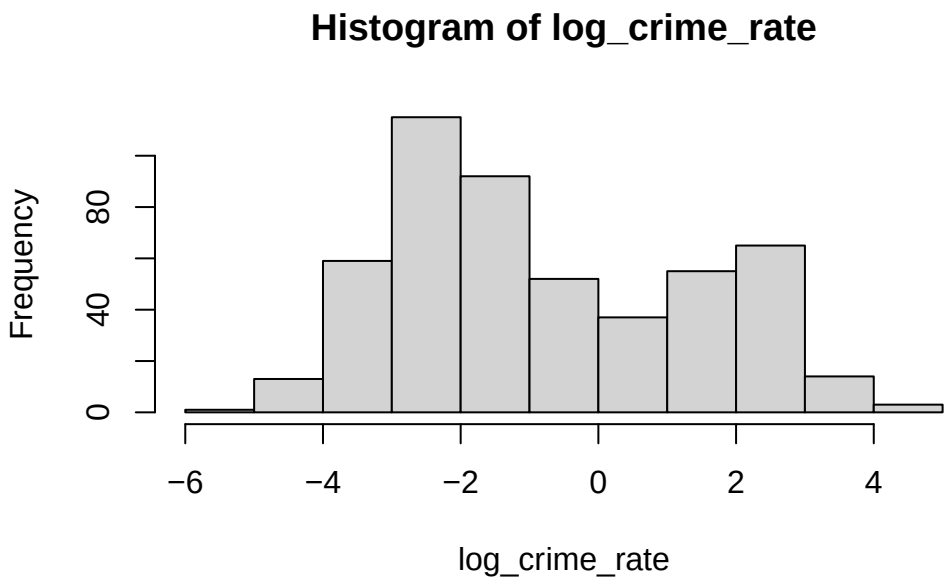
```
# A tibble: 2 x 2
  charles_river          mean_value
  <fct>                       <dbl>
1 Not bounding river           22.1
2 Bounding Charles river       28.4
```

```r
# boxplot ------------------------------------------------------------------

boxplot(medv ~ charles_river, data = Boston)
```



- 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 per-capita crime rate. Save the result in a new variable. Draw a histogram of this transformed variable, and describe its distribution.
    - The distribution of crime rates appears bimodal on the logarithmic scale. On this scale, the distribution is asymmetric with a hint of right-skew. The median suggests a typical Boston census tract has a crime rate of 0.25651 on the original scale. The middle 50% of census tracts have crime rates ranging from 0.082045 to 3.6770825.

```r
log_crime_rate <- log(Boston$crim)
hist(log_crime_rate)
```


Histogram of log_crime_rate
Frequency
0
40
80
−6
−4
−2
0
2
4
log_crime_rate