
Lecture 11: Describing a Distribution

For each question below, write the R code you would use and record the result when appropriate. I encourage you to complete this in an RMarkdown document so that you will have the code/explanations for your personal records. Answer in complete sentences when asked to explain.

Part 1: Describing the Distribution of `iris$Sepal.Length`

1. We will begin by looking at the `Sepal.Length` variable in the built-in `iris` dataset.

(a) Display the first 6 observations of the `iris` dataset.

(b) What does one row of the `iris` dataset represent?

(c) Is `Sepal.Length` quantitative or categorical? Explain briefly.

2. Step 1: Center

(a) Find the mean of `Sepal.Length`.

(b) Find the median of `Sepal.Length`.

(c) Find the 10% trimmed mean of `Sepal.Length`.

-
- (d) Compare the mean, median, and trimmed mean. What might this suggest about the shape of the distribution?

3. Step 2: Quartiles and Position

- (a) Display the five-number summary of `Sepal.Length`.
- (b) Find the quartiles of `Sepal.Length` using the `quantile()` function.
- (c) Compute the range of the 2nd Quartile ($\text{Median} - Q1$).
- (d) Compute the range of the 3rd Quartile ($Q3 - \text{Median}$).
- (e) Compute the range of the 1st Quartile ($Q1 - \text{Minimum}$).
- (f) Compute the range of the 4th Quartile ($\text{Maximum} - Q3$).
- (g) Based on your quartile analysis, does the distribution appear roughly symmetric, left-skewed, or right-skewed? Explain.

4. Step 3: Spread

- (a) Find the variance of `Sepal.Length`.
- (b) Find the standard deviation of `Sepal.Length`.
- (c) Find the IQR of `Sepal.Length`. What does this tell us about the spread?

5. Step 4: Visualizing the Distribution

- (a) Create a histogram of `Sepal.Length`.
- (b) Add a density curve or create a density plot of `Sepal.Length`.
- (c) Does the visualization agree with your conclusions from the center and quartile analysis? Explain.

6. Step 5: Describing the Distribution

Write 2–3 complete sentences describing the distribution of `Sepal.Length`. Your description should mention the center, spread, and shape.

Part 2: Describing the Distribution of `airquality$Ozone`

7. Now we will look at the `Ozone` variable in the built-in `airquality` dataset.

(a) Display the first 6 observations of the `airquality` dataset.

(b) What does one row of the `airquality` dataset represent?

(c) Is `Ozone` quantitative or categorical? Explain briefly.

8. Step 1: Center

(a) Find the mean of `Ozone`. You should get a value, not `NA`.

(b) Find the median of `Ozone`.

(c) Find the 10% trimmed mean of `Ozone`.

(d) Compare the mean, median, and trimmed mean. What does this suggest about skew or outliers?

9. Step 2: Quartiles and Position

- (a) Display the five-number summary of `Ozone`.
- (b) Find the quartiles of `Ozone` using the `quantile()` function.
- (c) Compute the range of the 2nd Quartile ($\text{Median} - Q1$).
- (d) Compute the range of the 3rd Quartile ($Q3 - \text{Median}$).
- (e) Compute the range of the 1st Quartile ($Q1 - \text{Minimum}$).
- (f) Compute the range of the 4th Quartile ($\text{Maximum} - Q3$).
- (g) Based on your quartile analysis, does the distribution appear roughly symmetric, left-skewed, or right-skewed? Explain.

10. Step 3: Spread

- (a) Find the variance of `Ozone`.

(b) Find the standard deviation of `Ozone`.

(c) Find the IQR of `Ozone`.

(d) Use the $1.5 \times \text{IQR}$ rule to determine the lower and upper outlier cutoffs.

(e) Are there any potential outliers based on these cutoffs?

11. Step 5: Visualizing the Distribution

(a) Create a histogram of `Ozone` with a density curve added to the plot.

(b) Does the visualization agree with your conclusions from the center and quartile analysis? Explain.

12. Step 6: Describing the Distribution

Write 2–3 complete sentences describing the distribution of `Ozone`. Your description should mention the center, spread, and shape.

Part 3: Comparing the Two Distributions

13. Compare `iris$Sepal.Length` and `airquality$Ozone`.

(a) Which variable appears more symmetric?

(b) Which variable appears more skewed?

(c) Which variable appears more affected by extreme values?

(d) For which variable would the mean and standard deviation be more appropriate? Why?

(e) For which variable would the median and IQR be more appropriate? Why?

14. **Reflection:** In one or two sentences, explain why we should use both numerical summaries and visualizations when describing a distribution.