Statistical Methods for Data Science 2025/2026

Homework 1 – Exercises

October 27, 2025

Deadline: November 10, 2025, 22:00 hrs.

Use the "Answer Sheet Homework 1_27-10-2025.docx" file to provide and upload your answers. Follow strictly the instructions provided therein. In opposite case, your answer(s) will not be graded. Work on your own. Please remember that it is academically dishonest and thus forbidden to copy from/cooperate with another classmate/student, discuss the exercises with a third person, or use AI tools such as ChatGPT. So, prepare the homework by yourself, referencing explicitly any resources (books, notes, papers, sites, etc.) you may use. Your upload will be checked for plagiarism using Turnitin.

NO DEADLINE EXTENSION WILL BE GIVEN.

Exercises

1. Consumer complaints are frequently reported to the *Better Business Bureau* (BBB). Some industries against whom the most complaints are reported to the BBB are banks; cable and satellite television companies; collection agencies; cellular phone providers; and new car dealerships (*USA Today*). The results for a sample of 200 complaints are contained in the file `BBB.csv`.
   a. Show the frequency and percent frequency of complaints by industry.
   b. Construct a bar chart and a pie chart of the percent frequency distribution.
   c. Which industry had the highest number of complaints?
   d. Comment on the percentage frequency distribution for complaints.
2. Consider the following data, contained in `data.txt`:
   8.9, 10.2, 11.5, 7.8, 10.0, 12.2, 13.5, 14.1, 10.0, 12.2, 6.8, 9.5, 11.5, 11.2, 14.9, 7.5, 10.0, 6.0, 15.8, 11.5
   a. Construct a dot plot.
   b. Construct a stem-and-leaf display, using a leaf unit of 0.1.
   c. Construct a stretched stem-and-leaf display, using a leaf unit of 0.1.
3. University endowments are financial assets that are donated by supporters to be used to provide income to universities. There is a large discrepancy in the size of university endowments. The file `Endowments.csv` provides a listing of many of the universities that have the largest endowments as reported by *the National Association of College and University Business Officers* in 2017. Summarize the data by constructing the following:
   a. A frequency distribution (classes 0–1.9, 2.0–3.9, 4.0–5.9, 6.0–7.9, and so on).
   b. A relative frequency distribution.
   c. A cumulative frequency distribution.
   d. A cumulative relative frequency distribution.
   e. What do these distributions tell you about the endowments of universities?
   f. Show a histogram. Comment on the shape of the distribution.
   g. What is the largest university endowment and which university holds it?
4. Each year *Forbes* ranks the world's most valuable brands. A portion of the data for 82 of the brands in the 2013 Forbes list is included in file `BrandValue.csv` (*Forbes website*). The data set includes the following variables:
   Brand: The name of the brand.

Industry: The type of industry associated with the brand, labeled Automotive & Luxury, Consumer Packaged Goods, Financial Services, Other, Technology.

Brand Value ($ billions): A measure of the brand's value in billions of dollars developed by Forbes based on a variety of financial information about the brand.

1-Yr Value Change (%): The percentage change in the value of the brand over the previous year.

Brand Revenue ($ billions): The total revenue in billions of dollars for the brand.

a. Prepare a crosstabulation of the data on Industry (rows) and Brand Value ($ billions). Use classes of [0,10), [10,20), [20,30), [30,40), [40,50), and [50,60] for Brand Value ($ billions).

b. Prepare a frequency distribution for the data on Industry.

c. Prepare a frequency distribution for the data on Brand Value ($ billions).

d. Compute the row and the column percentages for (b), (c).

e. How has the crosstabulation helped in preparing the frequency distributions in parts (b) and (c)?

f. What conclusions can you draw about the type of industry and the brand value?

5. The file `Snow.csv` contains temperature and snowfall data for 51 major U.S. cities over 30 years. For example, the average low temperature for Columbus, Ohio, is 44 degrees and the average annual snowfall is 27.5 inches.

   a. Construct a scatter diagram with the average annual low temperature on the horizontal axis and the average annual snowfall on the vertical axis.

   b. Does there appear to be any relationship between these two variables?

6. Electric plug-in vehicle sales have been increasing worldwide. The file `ElectricVehicles.csv` displays data collected by the *U.S. Department of Energy* on electric plug-in vehicle sales in the words top markets in 2013 and 2015. (Data compiled by *Argonne National Laboratory, U.S. Department of Energy website*, https://www.energy.gov/eere/vehicles/fact-918-march-28-2016-global-plug-light-vehicle-sales-increased-about-80-2015)

   a. Construct a side-by-side bar chart with year as the variable on the horizontal axis. Comment on any trend in the display.

   b. Convert the above table to percentage allocation for each year. Construct a stacked bar chart with year as the variable on the horizontal axis.

   c. Is the display in part (a) or part (b) more insightful? Explain.

7. The creator of a new online multiplayer survival game has been tracking the monthly downloads of the newest game. File `OnlineGame.csv` contains the monthly downloads (in thousands) for each month of the current and previous year.

   a. Compute the mean, median, and mode for number of downloads in the previous year.

   b. Compute the mean, median, and mode for number of downloads in the current year.

   c. Compute the first and third quartiles for downloads in the previous year.

   d. Compute the first and third quartiles for downloads in the current year.

   e. Compare the values calculated in parts (a)-(d) for the previous and current years. What does this tell you about the downloads of the game in the current year compared to the previous year?

8. *Martinez Auto Supplies* has retail stores located in eight cities in California. The price they charge for a particular product in each city varies because of differing competitive conditions. For instance, the price they charge for a case of a popular brand of motor oil in each city and the number of cases that they sold last quarter in each city are shown in `AutoSupplies.csv`. Compute the average sales price per case for this product during the last quarter.

9. *The New York Times* reported that Apple has unveiled a new iPad marketed specifically to school districts for use by students (*The New York Times website*). The 9.7-inch iPads will have faster processors and a cheaper price point in an effort to take market share away from Google Chromebooks in public school districts. Suppose that the data contained in `iPad.txt` file represent the percentages of students currently using Apple iPads for a sample of 18 U.S. public school districts.
   a. Compute the mean and median percentage of students currently using Apple iPads.
   b. Compare the first and third quartiles for these data.
   c. Compute the range and interquartile range for these data.
   d. Compute the variance and standard deviation for these data.
   e. Are there any outliers in this data?
   f. Based on your calculated values, what can we say about the percentage of students using iPads in public school districts?
10. The following times were recorded by the quarter mile and mile runners of a university track team (times are in minutes).

    Quarter-Mile times: .92, .98, 1.04, .90, .99
    Mile times: 4.52, 4.35, 4.60, 4.70, 4.50

    After viewing this sample of running times, one of the coaches commented that the quarter milers turned in the more consistent times. Use the standard deviation and the coefficient of variation to summarize the variability in the data. Does the use of the coefficient of variation indicate that the coach’s statement should be qualified?
11. Consider a sample with a mean of 30 and a standard deviation of 5.
    a. Use Chebyshev’s theorem to determine the percentage of the data within each of the following ranges: 20 to 40, 15 to 45.
    b. Suppose the data has a bell-shaped distribution. Use the empirical rule to determine the percentage of data within each of the above ranges.
12. *Consumer Reports* provides overall customer satisfaction scores for AT&T, Sprint, T-Mobile, and Verizon cell-phone services in major metropolitan areas throughout the United States. The rating for each service reflects the overall customer satisfaction considering a variety of factors such as cost, connectivity problems, dropped calls, static interference, and customer support. A satisfaction scale from 0 to 100 is used, with 0 indicating completely dissatisfied and 100 indicating completely satisfied. Suppose that the ratings for the four cell-phone services in 20 metropolitan areas are as recorded in `CellService.csv`.
    a. Consider T-Mobile first. What is the median rating?
    b. Develop a five-number summary for the T-Mobile service.
    c. Are there any outliers for T-Mobile? Explain.
    d. Repeat parts (b) and (c) for the other three cell-phone services.
    e. Show the boxplots for the four cell-phone services on one graph. Discuss what a comparison of the boxplots tells about the four services. Which service does Consumer Reports recommend as being best in terms of overall customer satisfaction?
13. The file `StockComparison.csv` contains monthly adjusted stock prices for technology company Apple, Inc., and consumer-goods company Procter & Gamble (P&G) from 2013–2018.
    a. Develop a scatter diagram with Apple stock price on the horizontal axis and P&G stock price on the vertical axis.
    b. What appears to be the relationship between these two stock prices?
    c. Compute and interpret the sample covariance.
    d. Compute the sample correlation coefficient. What does this value indicate about the relationship between the stock price of Apple and the stock price of P&G?