

Homework 1: Describing Data

Please hand in a hard copy of your homework in class and email your R code to Alex Goebel (AJGOEBEL@email.sc.edu).

1. Transit survey components [15 points]

The transportation office wants to estimate the proportion of all currently enrolled students who would use a late-night shuttle. It samples university email addresses and obtains 620 completed questionnaires.

- a. Identify the target population, sampled population, sampling frame, sampling unit, observation unit, and response.
- b. Explain why the target and sampled populations may differ.

Worked solution

- a. Target: all currently enrolled students. Sampled population: enrolled students reachable through the email list and willing/able to respond. Sampling unit and observation unit: a student. Response: whether the student would use the shuttle, possibly with supporting variables.
- b. Outdated addresses, inactive accounts, spam filtering, and nonresponse can exclude students systematically.

2. Commute-time distribution [20 points]

The following commute times in minutes were recorded for 24 students.

- a. Construct a frequency table using classes 10-14, 15-19, ..., 45-49.
- b. Draw a histogram and describe shape, center, spread, and unusual values.

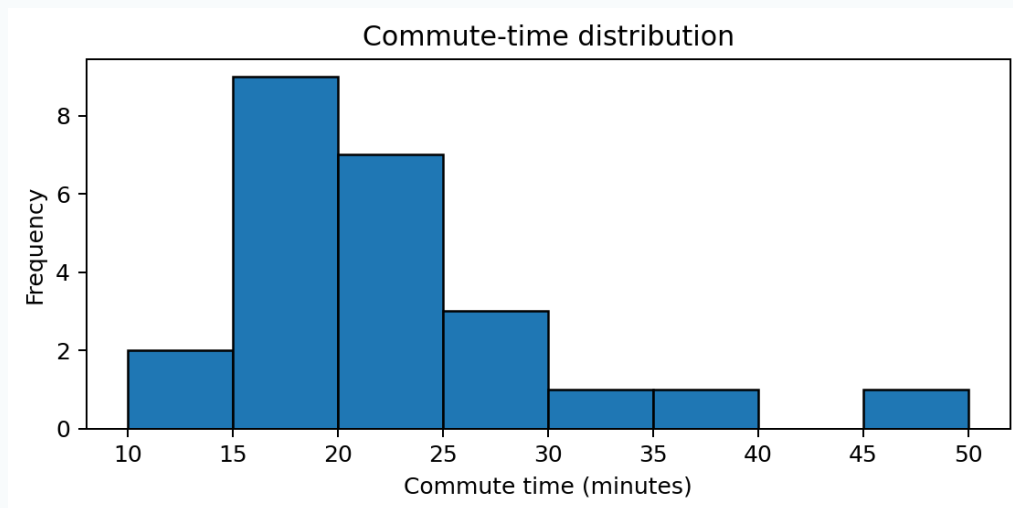
12	18	21	16	24	19	15	27
22	18	35	20	17	23	25	19
14	31	18	26	20	22	45	16

Worked solution

a. The class counts are shown below.

b. The distribution is unimodal with a right tail. Most observations lie between 15 and 24 minutes. The 45-minute observation is unusually high.

Class	Frequency
10-14	2
15-19	9
20-24	7
25-29	3
30-34	1
35-39	1
40-44	0
45-49	1



A five-minute-bin histogram reveals a right tail.

3. Center, spread, and boxplot [30 points]

Using the commute data, compute the mean, median, mode, range, sample variance, sample standard deviation, Q1, Q3, IQR, and the book's scaled MAD. Use the median-of-halves quartile convention. Apply the 1.5 IQR rule and sketch a boxplot.

Using the commute data, compute the mean, median, mode, range, sample variance, sample standard deviation, Q1, Q3, IQR, and the book's scaled MAD. Use the median-of-halves quartile convention. Apply the 1.5 IQR rule and sketch a boxplot.

Worked solution

Mean = 21.792; median = 20.0; mode = 18; range = 33.

Sample variance = 52.955; sample SD = 7.277. $Q1 = 17.5$, $Q3 = 24.5$, $IQR = 7.0$.

The unscaled median absolute deviation is 3.5, so scaled $MAD = 3.5/0.6745 = 5.189$. Fences are 7.0 and 35.0; 45 is an outlier. The upper whisker reaches 35.

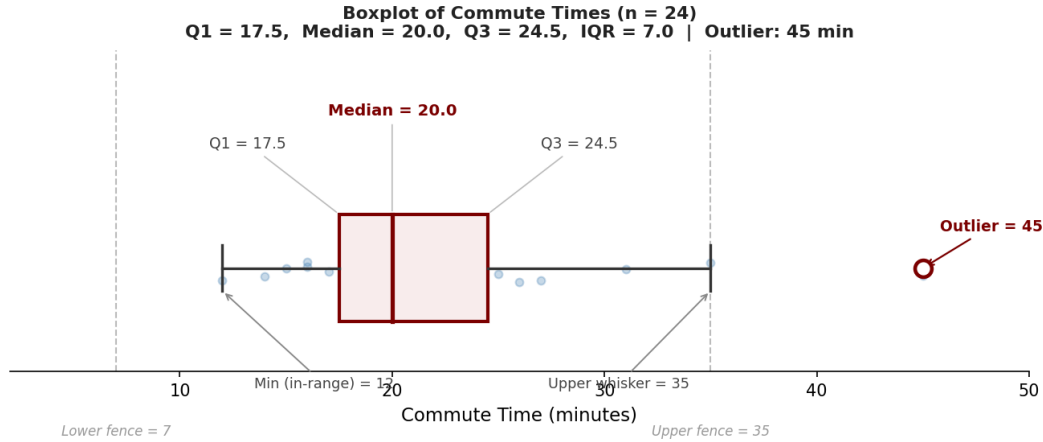


Figure: Boxplot of commute times. Box spans $Q1$ – $Q3$; median line at 20.0; upper whisker at 35; point at 45 is an outlier (beyond upper fence of 35).

4. Effect of the outlier [15 points]

Remove the 45-minute value and recompute the mean, median, sample SD. Which measures change most, and why?

Worked solution

Without 45: mean = 20.783, median = 20.0, sample SD = 5.460.

The mean and SD decrease noticeably because they use distances from the mean and squared deviations. The median is resistant and change much less.