16

Exploratory data analysis: numerical summaries

The classical way to describe important features of a dataset is to give several numerical summaries. We discuss numerical summaries for the center of a dataset and for the amount of variability among the elements of a dataset, and then we introduce the notion of quantiles for a dataset. To distinguish these quantities from corresponding notions for probability distributions of random variables, we will often add the word *sample* or *empirical*; for instance, we will speak of the sample mean and empirical quantiles. We end this chapter with the *boxplot*, which combines some of the numerical summaries in a graphical display.

16.1 The center of a dataset

The best-known method to identify the *center* of a dataset is to compute the *sample mean*

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{16.1}$$

For the sake of notational convenience we will sometimes drop the subscript n and write $\bar{x}$ instead of $\bar{x}_n$. The following dataset consists of hourly temperatures in degrees Fahrenheit (rounded to the nearest integer), recorded at Wick in northern Scotland from 5 p.m. December 31, 1960, to 3 a.m. January 1, 1961. The sample mean of the 11 measurements is equal to 44.7.

43 43 41 41 41 42 43 58 58 41 41

Source: V. Barnett and T. Lewis. *Outliers in statistical data.* Third edition, 1994.

Another way to identify the center of a dataset is by means of the *sample median*, which we will denote by $\mathrm{Med}(x_1, x_2, \ldots, x_n)$ or briefly Med_n. The sample median is defined as the middle element of the dataset when it is put in ascending order. When n is odd, it is clear what this means. When n is even,

we take the average of the two middle elements. For the Wick temperature data the sample median is equal to 42.

Quick exercise 16.1 Compute the sample mean and sample median of the dataset

$$4.6 \quad 3.0 \quad 3.2 \quad 4.2 \quad 5.0.$$

Both methods have pros and cons. The sample mean is the natural analogue for a dataset of what the expectation is for a probability distribution. However, it is very sensitive to *outliers*, by which we mean observations in the dataset that deviate a lot from the bulk of the data.

To illustrate the sensitivity of the sample mean, consider the Wick temperature data displayed in Figure 16.1. The values 58 and 58 recorded at midnight and 1 a.m. are clearly far from the bulk of the data and give grounds for concern whether they are genuine (58 degrees Fahrenheit seems very warm at midnight for New Year's in northern Scotland). To investigate their effect on the sample mean we compute the average of the data, leaving out these measurements, which gives 41.8 (instead of 44.7). The sample median of the data is equal to 41 (instead of 42) when leaving out the measurements with value 58. The median is more robust in the sense that it is hardly affected by a few outliers.

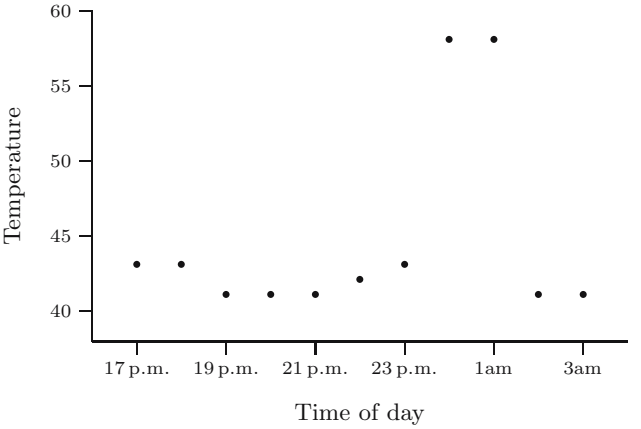


Fig. 16.1. The Wick temperature data.

It should be emphasized that this discussion is only meant to illustrate the sensitivity of the sample mean and by no means is intended to suggest we leave out measurements that deviate a lot from the bulk of the data! It is important to be *aware* of the presence of an outlier. In that case, one could try to find out whether there is perhaps something suspicious about this measurement. This might lead to assigning a smaller weight to such a measurement or even to

removing it from the dataset. However, sometimes it is possible to reconstruct the exact circumstances and correct the measurement. For instance, after further inquiry in the temperature example it turned out that at midnight the meteorological office changed its recording unit from degrees Fahrenheit to 1/10th degree Celsius (so 58 and 41 should read 5.8°C and 4.1°C). The corrected values in degrees Fahrenheit (to the nearest integer) are

$$43 \quad 43 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad \textit{42} \quad \textit{42} \quad \textit{39} \quad \textit{39}.$$

For the corrected data the sample mean is 41.5 and the sample median is 42.

Quick exercise 16.2 Consider the same dataset as in Quick exercise 16.1. Suppose that someone misreads the dataset as

$$4.6 \quad 30 \quad 3.2 \quad 4.2 \quad 50.$$

Compute the sample mean and sample median and compare these values with the ones you found in Quick exercise 16.1.

16.2 The amount of variability of a dataset

To quantify the amount of variability among the elements of a dataset, one often uses the *sample variance* defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

Up to a scaling factor this is equal to the average squared deviation from $\bar{x}_n$. At first sight, it seems more natural to define the sample variance by

$$\tilde{s}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

Why we choose the factor $1/(n-1)$ instead of $1/n$ will be explained later (see Chapter 19). Because s_n^2 is in different units from the elements of the dataset, one often prefers the *sample standard deviation*

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2},$$

which is measured in the same units as the elements of the dataset itself.

Just as the sample mean, the sample standard deviation is very sensitive to outliers. For the (uncorrected) Wick temperature data the sample standard deviation is 6.62, or 0.97 if we leave out the two measurements with value 58.

For the corrected data the standard deviation is 1.44. A more robust measure of variability is the *median of absolute deviations* or MAD, which is defined as follows. Consider the absolute deviation of every element x_i with respect to the sample median:

$$|x_i - \text{Med}(x_1, x_2, \ldots, x_n)|$$

or briefly

$$|x_i - \text{Med}_n|.$$

The MAD is obtained by taking the median of all these absolute deviations

$$\text{MAD}(x_1, x_2, \ldots, x_n) = \text{Med}(|x_1 - \text{Med}_n|, \ldots, |x_n - \text{Med}_n|). \tag{16.2}$$

Quick exercise 16.3 Compute the sample standard deviation for the dataset of Quick exercise 16.1 for which it is given that the values of $x_i - \bar{x}_n$ are:

$$-1.0,\ 0.6,\ -0.8,\ 0.2,\ 1.0.$$

Also compute the MAD for this dataset.

Just as the sample median, the MAD is hardly affected by outliers. For the (uncorrected) Wick temperature data the MAD is 1 and equal to 0 if we leave out the two measurements with value 58 (the value 0 seems a bit strange, but is a consequence of the fact that the observations are given in degrees Fahrenheit rounded to the nearest integer). For the corrected data the MAD is 1.

Quick exercise 16.4 Compute the sample standard deviation for the misread dataset of Quick exercise 16.2 for which it is given that the values of $x_i - \bar{x}_n$ are:

$$11.6,\ -13.8,\ -15.2,\ -14.2,\ 31.6.$$

Also compute the MAD for this dataset and compare both values with the ones you found in Quick exercise 16.3.

16.3 Empirical quantiles, quartiles, and the IQR

The sample median divides the dataset in two more or less equal parts: about half of the elements are less than the median and about half of the elements are greater than the median. More generally, we can divide the dataset in two parts in such a way that a proportion p is less than a certain number and a proportion $1 - p$ is greater than this number. Such a number is called the $100p$ *empirical percentile* or the *pth empirical quantile* and is denoted by $q_n(p)$. For a suitable introduction of empirical quantiles we need the notion of order statistics.

The *order statistics* consist of the same elements as in the original dataset $x_1, x_2, \ldots, x_n$, but in ascending order. Denote by $x_{(k)}$ the kth element in the ordered list. Then

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

are called the order statistics of $x_1, x_2, \ldots, x_n$. The order statistics of the Wick temperature data are

$$41 \quad 41 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad 43 \quad 43 \quad 58 \quad 58.$$

Note that by putting the elements in order, it is possible that successive order statistics are the same, for instance, $x_{(1)} = \cdots = x_{(5)} = 41$. Another example is Table 15.2, which lists the order statistics of the Old Faithful dataset.

To compute empirical quantiles one linearly interpolates between order statistics of the dataset. Let $0 < p < 1$, and suppose we want to compute the pth empirical quantile for a dataset $x_1, x_2, \ldots, x_n$. The following computation is based on requiring that the ith order statistic is the $i/(n+1)$ quantile. If we denote the integer part of a by $\lfloor a \rfloor$, then the computation of $q_n(p)$ runs as follows:

$$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$

with $k = \lfloor p(n+1) \rfloor$ and $\alpha = p(n+1) - k$. On the left in Figure 16.2 the relation between the pth quantile and the empirical distribution function is illustrated for the Old Faithful data.

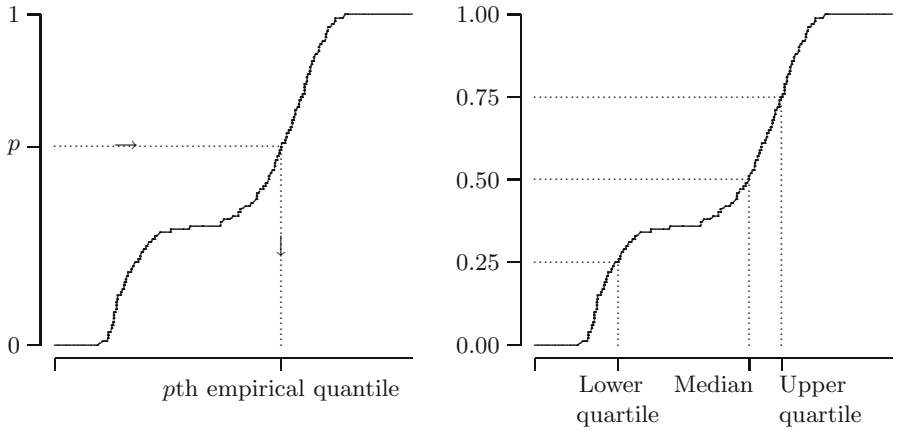


Fig. 16.2. Empirical quantile and quartiles for the Old Faithful data.

QUICK EXERCISE 16.5 Compute the 55th empirical percentile for the Wick temperature data.

Lower and upper quartiles

Instead of identifying only the center of the dataset, Tukey [35] suggested to give a five-number summary of the dataset: the minimum, the maximum, the sample median, and the 25th and 75th empirical percentiles. The 25th empirical percentile $q_n(0.25)$ is called the *lower quartile* and the 75th empirical percentile $q_n(0.75)$ is called the *upper quartile*. Together with the median, the lower and upper quartiles divide the dataset in four more or less equal parts consisting of about one quarter of the number of elements. The relation of the two quartiles and the median with the empirical distribution function is illustrated for the Old Faithful data on the right of Figure 16.2. The distance between the lower quartile and the median, relative to the distance between the upper quartile and the median, gives some indication on the skewness of the dataset. The distance between the upper and lower quartiles is called the *interquartile range*, or IQR:

$$\text{IQR} = q_n(0.75) - q_n(0.25).$$

The IQR specifies the range of the middle half of the dataset. It could also serve as a robust measure of the amount of variability among the elements of the dataset. For the Old Faithful data the five-number summary is

Minimum	Lower quartile	Median	Upper quartile	Maximum
96	129.25	240	267.75	306

and the IQR is 138.5.

Quick exercise 16.6 Compute the five-number summary for the (uncorrected) Wick temperature data.

16.4 The box-and-whisker plot

Tukey [35] also proposed visualizing the five-number summary discussed in the previous section by a so-called box-and-whisker plot, briefly *boxplot*. Figure 16.3 displays a boxplot. The data are now on the vertical axis, where we left out the numbers on the axis in order to explain the construction of the figure. The horizontal width of the box is irrelevant. In the vertical direction the box extends from the lower to the upper quartile, so that the height of the box is precisely the IQR. The horizontal line inside the box corresponds to the sample median. Up from the upper quartile we measure out a distance of 1.5 times the IQR and draw a so-called *whisker* up to the largest observation that lies within this distance, where we put a horizontal line. Similarly, down from the lower quartile we measure out a distance of 1.5 times the IQR and draw a whisker to the smallest observation that lies within this distance, where we also put a horizontal line. *All* other observations beyond the whiskers are marked by ∘. Such an observation is called an *outlier*.

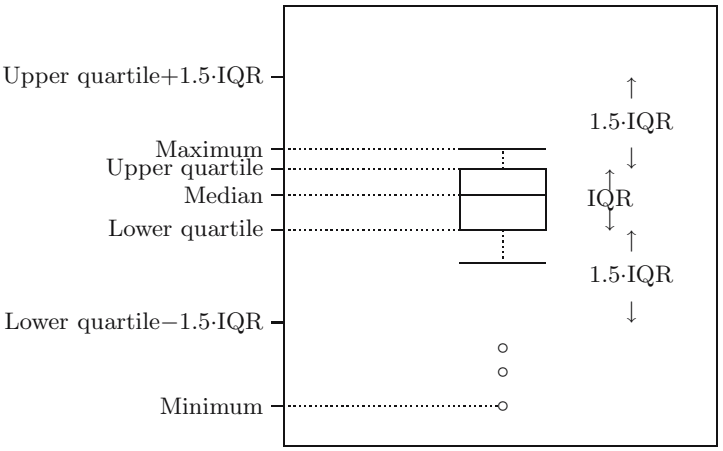


Fig. 16.3. A boxplot.

In Figure 16.4 the boxplots of the Old Faithful data and of the software reliability data (see also Chapter 15) are displayed. The skewness of the software reliability data produces a boxplot with whiskers of very different length and with several observations beyond the upper quartile plus 1.5 times the IQR. The boxplot of the Old Faithful data illustrates one of the shortcomings of the boxplot; it does not capture the fact that the data show two separate peaks. However, the position of the sample median inside the box does suggest that the dataset is skewed.

Quick exercise 16.7 Suppose we want to construct a boxplot of the (uncorrected) Wick temperature data. What is the height of the box, the length of both whiskers, and which measurements fall outside the box and whiskers? Would you consider the two values 58 extreme outliers?

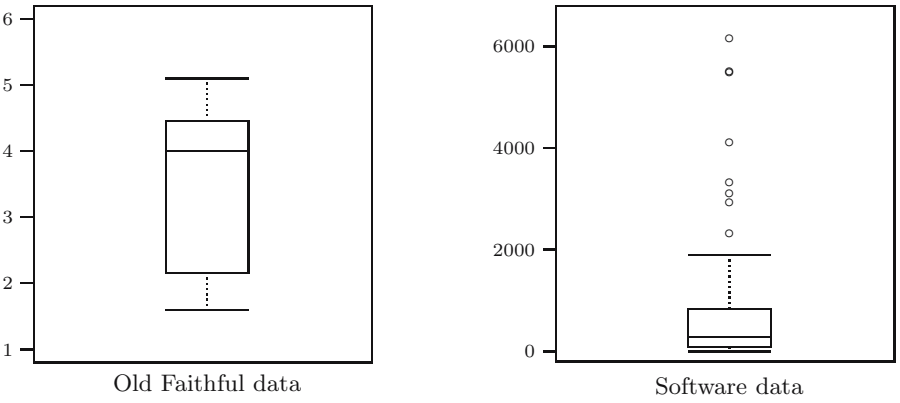


Fig. 16.4. Boxplot of the Old Faithful data and the software data.

Using boxplots to compare several datasets

Although the boxplot provides some information about the structure of the data, such as center, range, skewness or symmetry, it is a poor graphical display of the dataset. Graphical summaries such as the histogram and kernel density estimate are more informative displays of a single dataset. Boxplots become useful if we want to compare several sets of data in a simple graphical display. In Figure 16.5 boxplots are displayed of the average drill time for dry and wet drilling up to a depth of 250 feet for the drill data discussed in Section 15.5 (see also Table 15.4). It is clear that the boxplot corresponding to dry drilling differs from that corresponding to wet drilling. However, the question is whether this difference can still be attributed to chance or is caused by the drilling technique used. We will return to this type of question in Chapter 25.

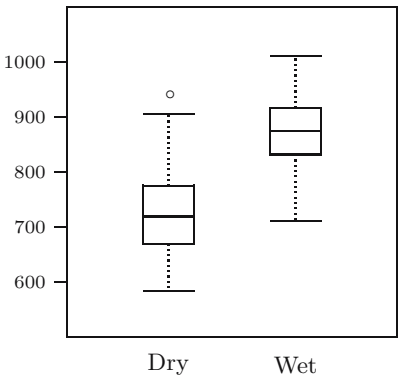


Fig. 16.5. Boxplot of average drill times.

16.5 Solutions to the quick exercises

16.1 The average is

$$\bar{x}_n = \frac{4.6 + 3.0 + 3.2 + 4.2 + 5.0}{5} = \frac{20}{5} = 4.$$

The median is the middle element of 3.0, 3.2, 4.2, 4.6, and 5.0, which gives $\text{Med}_n = 4.2$.

16.2 The average is

$$\bar{x}_n = \frac{4.6 + 30 + 3.2 + 4.2 + 50}{5} = \frac{90}{5} = 18,$$

which differs 14.4 from the average we found in Quick exercise 16.1. The median is the middle element of 3.2, 4.2, 4.6, 30, and 50. This gives $\text{Med}_n = 4.6$, which only differs 0.4 from the median we found in Quick exercise 16.1. As one can see, the median is hardly affected by the two outliers.

16.3 The sample variance is

$$s_n^2 = \frac{(-1)^2 + (0.6)^2 + (-0.8)^2 + (0.2)^2 + (1.0)^2}{5-1} = \frac{3.04}{4} = 0.76$$

so that the sample standard deviation is $s_n = \sqrt{0.76} = 0.872$. The median is 4.2, so that the absolute deviations from the median are given by

$$0.4 \quad 1.2 \quad 1.0 \quad 0.0 \quad 0.8.$$

The MAD is the median of these numbers, which is 0.8.

16.4 The sample variance is

$$s_n^2 = \frac{(11.6)^2 + (-13.8)^2 + (-15.2)^2 + (-14.2)^2 + (31.6)^2}{5-1} = \frac{1756.24}{4} = 439.06$$

so that the sample standard deviation is $s_n = \sqrt{439.06} = 20.95$, which is a difference of 20.19 from the value we found in Quick exercise 16.3. The median is 4.6, so that the absolute deviations from the median are given by

$$0.0 \quad 25.4 \quad 1.4 \quad 0.4 \quad 45.4.$$

The MAD is the median of these numbers, which is 1.4. Just as the median, the MAD is hardly affected by the two outliers.

16.5 We have $k = \lfloor 0.55 \cdot 12 \rfloor = \lfloor 6.6 \rfloor = 6$, so that $\alpha = 0.6$. This gives

$$q_n(0.55) = x_{(6)} + 0.6 \cdot (x_{(7)} - x_{(6)}) = 42 + 0.6 \cdot (43 - 42) = 42.6.$$

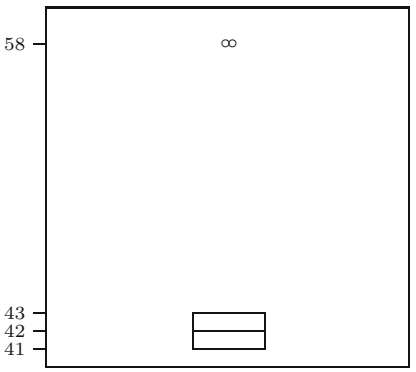
16.6 From the order statistics of the Wick temperature data

$$41 \quad 41 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad 43 \quad 43 \quad 58 \quad 58$$

it can be seen immediately that minimum, maximum, and median are given by 41, 58, and 42. For the lower quartile we have $k = \lfloor 0.25 \cdot 12 \rfloor = 3$, so that $\alpha = 0$ and $q_n(0.25) = x_{(3)} = 41$. For the upper quartile we have $k = \lfloor 0.75 \cdot 12 \rfloor = 9$, so that again $\alpha = 0$ and $q_n(0.75) = x_{(9)} = 43$. Hence for the Wick temperature data the five-number summary is

Minimum	Lower quartile	Median	Upper quartile	Maximum
41	41	42	43	58

16.7 From the five-number summary for the Wick temperature data (see Quick exercise 16.6), it follows immediately that the height of the box is the IQR: $43 - 41 = 2$. If we measure out a distance of 1.5 times 2 down from the lower quartile 41, we see that the smallest observation within this range is 41, which means that the lower whisker has length zero. Similarly, the upper whisker has length zero. The two measurements with value 58 are outside the box and whiskers. The two values 58 are clearly far away from the bulk of the data and should be considered extreme outliers.



16.6 Exercises

16.1 ⊡ Use the order statistics of the software data as given in Exercise 15.4 to answer the following questions.

- **a.** Compute the sample median.
- **b.** Compute the lower and upper quartiles and the IQR.
- **c.** Compute the 37th empirical percentile.

16.2 Compute for the Old Faithful data the distance of the lower and upper quartiles to the median and explain the difference.

16.3 ⊞ Recall the example about the space shuttle *Challenger* in Section 1.4. The following table lists the order statistics of launch temperatures during take-offs in degrees Fahrenheit, including the launch temperature on January 28, 1986.

31	53	57	58	63	66	67	67	67	68	69	70
70	70	70	72	73	75	75	76	76	78	79	81

- **a.** Find the sample median and the lower and upper quartiles.
- **b.** Sketch the boxplot of this dataset.

c. On January 28, 1986, the launch temperature was 31 degrees Fahrenheit. Comment on the value 31 with respect to the other data points.

16.4 ⊡ The sample mean and sample median of the uncorrected Wick temperature data (in degrees Fahrenheit) are 44.7 and 42. We transform the data from degrees Fahrenheit (x_i) to degrees Celsius (y_i) by means of the formula

$$y_i = \frac{5}{9}(x_i - 32),$$

which gives the following dataset

$$\frac{55}{9} \quad \frac{55}{9} \quad 5 \quad 5 \quad 5 \quad \frac{50}{9} \quad \frac{55}{9} \quad \frac{130}{9} \quad \frac{130}{9} \quad 5 \quad 5.$$

a. Check that $\bar{y}_n = \frac{5}{9}(\bar{x}_n - 32)$.

b. Is it also true that $\text{Med}(y_1, \ldots, y_n) = \frac{5}{9}(\text{Med}(x_1, \ldots, x_n) - 32)$?

c. Suppose we have a dataset $x_1, x_2, \ldots, x_n$ and construct $y_1, y_2, \ldots, y_n$ where $y_i = ax_i + b$ with a and b being real numbers. Do similar relations hold for the sample mean and sample median? If so, state them.

16.5 Consider the uncorrected Wick temperature data in degrees Fahrenheit (x_i) and the corresponding temperatures in degrees Celsius (y_i) as given in Exercise 16.4. The sample standard deviation and the MAD for the Wick data are 6.62 and 1.

a. Let s_F and s_C denote the sample standard deviations of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respectively. Check that $s_C = \frac{5}{9}s_F$.

b. Let MAD_F and MAD_C denote the MAD of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respectively. Is it also true that $\text{MAD}_C = \frac{5}{9}\text{MAD}_F$?

c. Suppose we have a dataset $x_1, x_2, \ldots, x_n$ and construct $y_1, y_2, \ldots, y_n$ where $y_i = ax_i + b$ with a and b being real numbers. Do similar relations hold for the sample standard deviation and the MAD? If so, state them.

16.6 ⊞ Consider two datasets: $1, 5, 9$ and $2, 4, 6, 8$.

a. Denote the sample means of the two datasets by $\bar{x}$ and $\bar{y}$. Is it true that the average $(\bar{x} + \bar{y})/2$ of $\bar{x}$ and $\bar{y}$ is equal to the sample mean of the combined dataset with 7 elements?

b. Suppose we have two other datasets: one of size n with sample mean $\bar{x}_n$ and another dataset of size m with sample mean $\bar{y}_m$. Is it always true that the average $(\bar{x}_n + \bar{y}_m)/2$ of $\bar{x}_n$ and $\bar{y}_m$ is equal to the sample mean of the combined dataset with $n + m$ elements? If no, then provide a counterexample. If yes, then explain this.

c. If $m = n$, is $(\bar{x}_n + \bar{y}_m)/2$ equal to the sample mean of the combined dataset with $n + m$ elements?

16.7 Consider the two datasets from Exercise 16.6.

a. Denote the sample medians of the two datasets by Med_x and Med_y. Is it true that the sample median $(\text{Med}_x + \text{Med}_y)/2$ of the two sample medians is equal to the sample median of the combined dataset with 7 elements?

b. Suppose we have two other datasets: one of size n with sample median Med_x and another dataset of size m with sample median Med_y. Is it always true that the sample median $(\text{Med}_x + \text{Med}_y)/2$ of the two sample medians is equal to the sample median of the combined dataset with $n+m$ elements? If no, then provide a counterexample. If yes, then explain this.

c. What if $m = n$?

16.8 ⊞ Compute the MAD for the combined dataset of 7 elements from Exercise 16.6.

16.9 Consider a dataset $x_1, x_2, \ldots, x_n$ with $x_i \neq 0$. We construct a second dataset $y_1, y_2, \ldots, y_n$, where

$$y_i = \frac{1}{x_i}.$$

a. Suppose dataset $x_1, x_2, \ldots, x_n$ consists of $-6, 1, 15$. Is it true that $\bar{y}_3 = 1/\bar{x}_3$?

b. Suppose that n is odd. Is it true that $\bar{y}_n = 1/\bar{x}_n$?

c. Suppose that n is odd and each $x_i > 0$. Is it true that $\text{Med}(y_1, \ldots, y_n) = 1/\text{Med}(x_1, \ldots, x_n)$? What about when n is even?

16.10 ⊡ A method to investigate the sensitivity of the sample mean and the sample median to extreme outliers is to replace one or more elements in a given dataset by a number y and investigate the effect when y goes to infinity. To illustrate this, consider the dataset from Quick Exercise 16.1:

$$4.6 \quad 3.0 \quad 3.2 \quad 4.2 \quad 5.0$$

with sample mean 4 and sample median 4.2.

a. We replace the element 3.2 by some real number y. What happens with the sample mean and the sample median of this new dataset as $y \to \infty$?

b. We replace a number of elements by some real number y. How many elements do we need to replace so that the sample median of the new dataset goes to infinity as $y \to \infty$?

c. Suppose we have another dataset of size n. How many elements do we need to replace by some real number y, so that the sample mean of the new dataset goes to infinity as $y \to \infty$? And how many elements do we need to replace, so that the sample median of the new dataset goes to infinity?

16.11 Just as in Exercise 16.10 we investigate the sensitivity of the sample standard deviation and the MAD to extreme outliers, by considering the same dataset with sample standard deviation 0.872 and MAD equal to 0.8. Answer the same three questions for the sample standard deviation and the MAD instead of the sample mean and sample median.

16.12 ⊡ Compute the sample mean and sample median for the dataset

$$1, 2, \ldots, N$$

in case N is odd and in case N is even. You may use the fact that

$$1 + 2 + \cdots + N = \frac{N(N+1)}{2}.$$

16.13 Compute the sample standard deviation and MAD for the dataset

$$-N, \ldots, -1, 0, 1, \ldots, N.$$

You may use the fact that

$$1^2 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}.$$

16.14 Check that the 50th empirical percentile is the sample median.

16.15 ⊞ The following rule is useful for the computation of the sample variance (and standard deviation). Show that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x}_n)^2$$

where $\bar{x}_n = (\sum_{i=1}^{n} x_i)/n$.

16.16 Recall Exercise 15.12, where we computed the mean and second moment corresponding to a density estimate $f_{n,h}$. Show that the variance corresponding to $f_{n,h}$ satisfies:

$$\int_{-\infty}^{\infty} t^2 f_{n,h}(t)\,\mathrm{d}t - \left(\int_{-\infty}^{\infty} t f_{n,h}(t)\,\mathrm{d}t\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + h^2 \int_{-\infty}^{\infty} u^2 K(u)\,\mathrm{d}u.$$

16.17 Suppose we have a dataset $x_1, x_2, \ldots, x_n$. Check that if $p = i/(n+1)$ the pth empirical quantile is the ith order statistic.