CHAPTER 3: NUMERICALLY SUMMARIZING DATA

[This chapter is based on Section 3.1 from the textbook]

In the previous chapter we learned how to summarize data, and we learned to build some graphs that help us visualize the data. In this chapter we will learn some numerical summaries of the data.

For any group of observations, we often want to know what is the shape of the data (centered, skewed right or skewed left), where is the center, and how 'disperse' the data is around the center. We will properly define these measures as we move on.

3.1 Measures of Central Tendency

Measures of central tendency try to summarize the data by providing the center. But what is the center? When we talk about center, we want to give one number that described the "typical" observation. We will see three different ways to give such a summary of the data.

The arithmetic mean

We start with the most popular measure.

Definition 3.1. *The arithmetic mean of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.*

For example, the arithmetic mean of the numbers 3, 8, 1, 4 is

$$\frac{3+8+1+4}{4} = \frac{16}{4} = 4$$

The definition above can be used for any set of quantitative variables. In particular, we can compute the population arithmetic mean, and a sample arithmetic mean as follows:

(i) **Population arithmetic mean:** We use the Greek letter μ (pronounced "mu") to denote it. If the population size is N, and $x_1, x_2, \ldots, x_N$ represent the value of the variable for each individual of the population, the population arithmetic mean is

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

(ii) **Sample arithmetic mean:** We use the symbol $\overline{x}$ (pronounced x bar) to denote it. If the sample size is n and the observations are represented by $x_1, \ldots, x_n$, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The symbol $\sum_{i=1}^{N}$ reads the sum of all the elements with index i taking values from 1 to N.

Let's see an example.

Example 3.1. *The following data represents the pulse rates (beats per minute) of nine students (the population).*

Student	*Pulse*
Perpectual Bempah	*76*
Megan Brooks	*60*
Jeff Honeycutt	*60*
Clarice Jefferson	*81*
Crystal Kurtenbach	*72*
Janette Lantka	*80*
Kevin McCarthy	*80*
Tammy Ohm	*68*
Kathy Wojdyla	*73*

Compute μ and $\overline{x}$ for a sample of 4 students.

Solution. To compute the population arithmetic mean μ we add up the pulse of every student and divide by the population size $N = 9$. We obtain:

$$\mu = \frac{76 + 60 + 60 + 81 + 72 + 80 + 80 + 68 + 73}{9} = 72.2$$

To compute $\overline{x}$ we first select a sample of 4 students. We can use any subset of 4 students, but in this case we use the students who's first name has 5 letters, that is, Megan, Kevin, Tammy and Kathy. If we use x_1, x_2, x_3, x_4 to denote their pulse (in that order), we have

$$x_1 = 60, \qquad x_2 = 80, \qquad x_3 = 68, \qquad x_4 = 73$$

and the sample arithmetic mean is

$$\overline{x} = \frac{60 + 80 + 68 + 73}{4} = 70.25$$

Observe that the sample arithmetic mean is smaller than the population arithmetic mean, that is, our sample underestimates the parameter μ. □

The median

We start with the definition.

Definition 3.2. *The median of a set of data is denoted by M and it is the value that lies in the middle of the data when arranged in ascending order. To compute it, we arrange the data, determine the number of observations n and:*

- *If n is an odd number, then the median is the observation in position $\dfrac{n+1}{2}$*
- *If n is an even number, then the median is the mean of the observations in positions $\dfrac{n}{2}$ and $\dfrac{n}{2} + 1$.*

Let's see some examples.

Example 3.2. *For the following sets of data, determine the median:*

(a) 76, 60, 60, 81, 72, 80, 80, 68, 73

(b) 39, 21, 9, 32, 30, 45, 11, 12, 39, 27

Solution.

(a) First observe that there are $n = 9$ observations. If we arrange them in ascending order, we obtain:

$$60 \quad 60 \quad 68 \quad 72 \quad 73 \quad 76 \quad 80 \quad 80 \quad 81$$

Since 9 is an odd number, the median is in the position $\dfrac{n+1}{2} = \dfrac{10}{2} = 5$, that is, the median is 73.

(b) Similarly to the previous case, we first observe that there are $n = 10$ observations. Next, we arrange the observations in ascending order and obtain

$$9 \quad 11 \quad 12 \quad 21 \quad 27 \quad 30 \quad 32 \quad 39 \quad 39 \quad 45$$

Since 10 is an even number, we compute the mean of the observations in positions $\frac{n}{2} = 5$ and $\frac{n}{2} + 1 = 6$, that is, the mean between 27 and 30. We obtain that the median is

$$\frac{27 + 30}{2} = 28.5$$

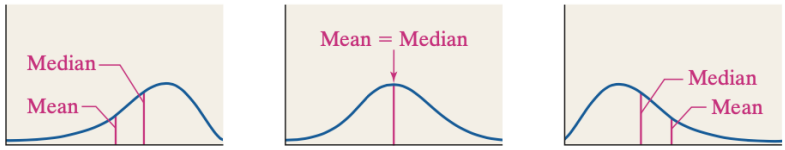
□

Comparing the mean and the median

Both, the mean and the median are indicating where is the center of the data. However, they use different ideas of 'center.' Both provide important information about the data, and if we study both at the same time, we can predict if the data is symmetric, skewed right or skewed left (recall these definitions from Chapter 2).

- Symmetric data: When the data is symmetric, the mean and median are roughly the same
- Skewed right: When the data is skewed right, the tail to the right of the peak in a histogram is longer. This means that most of the data has a small value, and a few data points have very large values. Then, the mean tends to be larger than the median because the mean uses the data values, and the median only uses their position.
- Skewed left: Similarly to the skewed right case, when the data is skewed left the mean is smaller than the median.

The following picture[1] summarizes this idea.



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

When the data is skewed right or left, the median is a better measure of the center than the mean. However, the mean is still very important and will help us in the process of statistics (as we will see in the following chapters).

The mode

The last measure of central tendency we will learn in this chapter is the mode, and we define it below.

Definition 3.3. *The mode of a variable is the most frequent observation of the variable that occurs in the data set.*

(i) If no observation occurs more than once, we say that the data set has no mode

(ii) If only one observation is the most frequent, we say that the data set is unimodal

(iii) If a data set has two modes, we say that the data set is bimodal

(iv) If a data set has three or more modes, we say that it is multimodal

Observe that the mean and median only make sense for quantitative data, but the mode can be computed for both, quantitative and qualitative data.

Let's see an example.

Example 3.3. *Compute the mode(s) for the following data sets:*

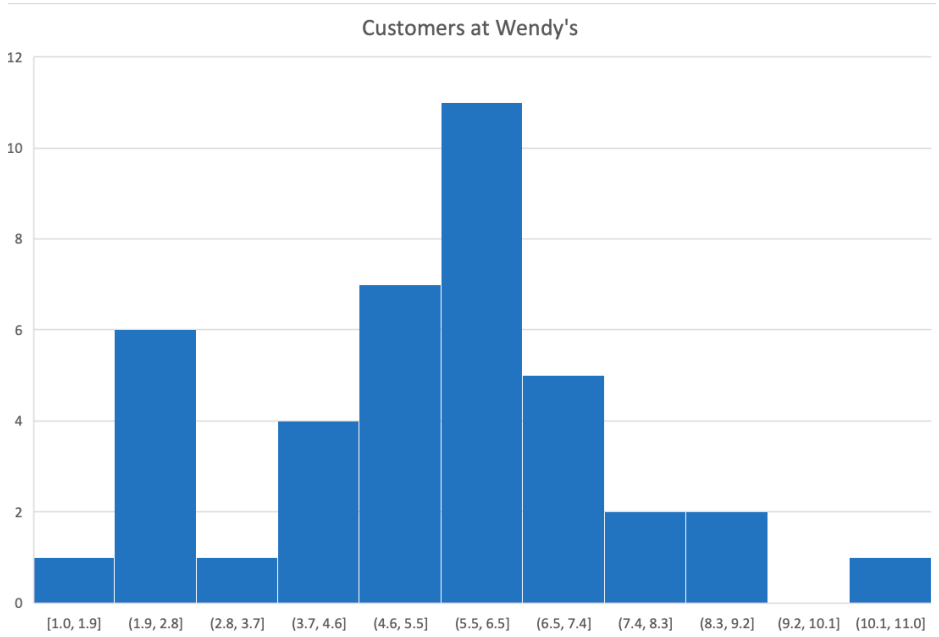
(a) The data set from Example 3.1

(b) The following data set, showing the location of injuries that requires rehabilitation by a physical therapist

[1] Figure 6 from Chapter 3 in the textbook

Category	*Frequency*
Back	*12*
Elbow	*1*
Groin	*1*
Hand	*2*
Hip	*2*
Knee	*5*
Neck	*1*
Shoulder	*4*
Wrist	*2*

(c) Consider the following histogram, where the horizontal axis represents the number of customers that arrive at a particular Wendy's in slots of 15 minutes at lunch time.



Solution.

(a) The sample is bimodal, and the modes are 60 and 80

(b) The sample is unimodal, and the mode is "Back"

(c) The sample is unimodal, and the mode is 6 customers

□

Let's end this section with one more example.

Example 3.4. *The following frequency table shows the number of children in a daycare separated out by their age.*

Age	*Frequency*
2	*3*
3	*7*
4	*6*
5	*1*

(a) How many children attend the daycare on this day?

(b) What is the mean age of the children at this daycare on this day?

(c) *What is the median of age of the children at this daycare on this day?*

(d) *What is the mode of age of the children at this daycare on this day?*

Solution.

(a) We add all the frequencies and obtain $3 + 7 + 6 + 1 = 17$

(b) We don't have the raw data, but we can easily obtain it using the frequency distribution. Observe that adding up all the raw data is equivalent to adding each observation multiplied by its frequency. Then, we have

$$\overline{x} = \frac{2 \cdot 3 + 3 \cdot 7 + 4 \cdot 6 + 5 \cdot 1}{17} = 3.294$$

(c) The number of data points is 17, which is an odd number. Then, we find the observation in position $\frac{17+1}{2} = 9$. We don't need to arrange the data in ascending order because we can use the frequency table. Observe that the first 3 data points equal 2, and the next 7 data points equal 3. Then the 9th data point must equal 3.

(d) We find the observation(s) with the highest frequency. In this case, the data set is unimodal and the mode is 3 years old.

□

3.2 Measures of Dispersion

In the last section we learned how to find the center of the data set, and we learned three ways to define the center (average, median and mode). In this chapter we will learn how to evaluate the dispersion of the data, that is, how the data is spread around the center. We will learn three measures, and we start with the range.

The range

We start with the definition.

Definition 3.4. *The range of a variable is denoted by R, and is the difference between the largest and the smallest value.*

Let's do a brief example.

Example 3.5. *Compute the range of the following data:* 39, 21, 9, 32, 30, 45, 11, 12, 39, 27

Solution. Observe that the smallest observation is 9 and the largest is 45. Therefore, the range is

$$R = 45 - 9 = 36$$

□

The range gives us an idea of how widely the data is spread. However, it is not the most powerful dispersion measure because it only considers two observations. Let's consider the following example.

Example 3.6. *Compute the range of the following data:* 39, 21, 9, 32, 30, 85, 11, 12, 39, 27

Solution. The data from is almost the same as in the previous example. The only difference is that the value 45 was replaced by 85. Hence, the new range is

$$R = 85 - 9 = 76$$

□

There is a huge difference in the range of both examples, even though we only changed one data point. In general, we will prefer using dispersion measures that use all the data points.

The standard deviation

Another way to measure dispersion around the mean is to compute the average deviation from the mean. Suppose that our sample is $x_1, x_2, \ldots, x_n$ and the sample arithmetic mean is $\overline{x}$. Then, the deviation of observation i from the mean is $x_i - \overline{x}$. Then, the average deviation from the mean is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$$

Similarly, we can define the average deviation of the population from the population mean as

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)$$

Let's see an example.

Example 3.7. *Suppose our dataset is* $2, 6, 7, 9, 11$. *Compute the average deviation from the arithmetic mean. Assume that the dataset represents the population.*

Solution. We first compute the arithmetic mean, and obtain:

$$\mu = \frac{2+6+7+9+11}{5} = 7$$

Now we compute the average deviation from the mean using the following table:

Observation notation	Observation value	Deviation $x_i - \mu$
x_1	2	$x_1 - \mu = 2 - 7 = -5$
x_2	6	$x_2 - \mu = 6 - 7 = -1$
x_3	7	$x_3 - \mu = 7 - 7 = 0$
x_4	9	$x_4 - \mu = 9 - 7 = 2$
x_5	11	$x_5 - \mu = 11 - 7 = 4$

Then, the total deviation is:

$$\sum_{i=1}^{5}(x_i - \mu) = (-5) + (-1) + 0 + 2 + 7 = 0$$

and, therefore, the average deviation is also 0. □

Obtaining an average deviation equal to 0 is not an accident. It is an immediate consequence of the definition of the arithmetic mean. We won't do a mathematical proof, but if you're interested, let me know and I can show it to you. An alternative to compute a measure of dispersion is making sure that all the individual deviations from the arithmetic mean are nonnegative. There are many ways to do that, but in this case we will square the individual deviations.

Definition 3.5.

(i) The population variance is denoted by σ^2 *(pronounced sigma squared). For a population with* N *individuals and observations denoted by* $x_1, \ldots, x_N$, *the variance is computed as follows:*

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

(ii) The sample variance is denoted by s. *For a sample with* n *individuals and observations denoted by* x_i *with* $i = 1, \ldots, n$, *the variance is computed as follows:*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Observe that in both cases we are computing the average of the squared deviations from the mean. In the case of population variance, we use the population mean μ, and for the sample variance, we use the sample mean $\overline{x}$.

Let's see an example.

Example 3.8. *Consider tha dataset from Example 3.7*

(a) Compute the population variance

(b) Consider a sample of prime numbers and compute the sample variance

Solution.

(a) Recall that the population mean is $\mu = 7$. We use the following table to compute the population variance:

Observation notation	Observation value	Squared deviation: $(x_i - \mu)^2$
x_1	2	$(x_1 - \mu)^2 = (-5)^2 = 25$
x_2	6	$(x_2 - \mu)^2 = (-1)^2 = 1$
x_3	7	$(x_3 - \mu)^2 = 0^2 = 0$
x_4	9	$(x_4 - \mu)^2 = 2^2 = 4$
x_5	11	$(x_5 - \mu)^2 = 4^2 = 16$

Then, the population variance is:

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \mu)^2 = \frac{25 + 1 + 0 + 4 + 16}{5} = 9.2$$

(b) Our sample is now $2, 7, 11$ and the sample size is 3. The sample arithmetic mean is

$$\overline{x} = \frac{2 + 7 + 11}{3} = 6.67$$

We use the following table to compute the sample variance:

Observation notation	Observation value	Squared deviation: $(x_i - \overline{x})^2$
x_1	2	$(x_1 - \overline{x})^2 = (2 - 6.67)^2 = (-4.47)^2 = 21.78$
x_2	7	$(x_2 - \overline{x})^2 = (7 - 6.67)^2 = 0.33^2 = 0.11$
x_3	11	$(x_3 - \overline{x})^2 = (11 - 6.67)^2 = 4.33^2 = 18.78$

Then, the sample variance is

$$s^2 = \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \overline{x})^2 = \frac{21.78 + 0.11 + 18.78}{2} = 20.33$$

□

Comparing the variance and arithmetic mean is tricky because they are measured in different units. For example, if the data is in dollars, the arithmetic mean is also in dollars but the variance is in squared dollars. Hence, we use the following dispersion measure instead.

Definition 3.6.

(i) The population standard deviation is denoted by σ (pronounced sigma) and computed as $\sigma = \sqrt{\sigma^2}$, that is, as the square root of the population variance. Writing the whole expression yields

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

(ii) The sample standard deviation is denoted by s and computed as $s = \sqrt{s^2}$, that is, the square root of the sample variance. The whole expression is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Example 3.9. *Following the computations from Example 3.7, we obtain the following standard deviations:*

$$\sigma = \sqrt{9.2} = 3.03$$
$$s = \sqrt{20.33} = 4.5$$

As a general rule, the standard deviation shows how disperse the data are around the mean. The higher the standard deviation, the more spread apart the data points are. A smaller standard deviation is related to smaller deviations from the arithmetic mean and, hence, more consistent outcomes. Let's see an example to illustrate this idea.

Example 3.10 (Exercise 21 from section 3.2 of the textbook)**.** *Ethan and Drew went on a 10-day fishing trip. The number of smallmouth bass caught and released by the two boys each day was as follows:*

Ethan	*9*	*24*	*8*	*9*	*5*	*8*	*9*	*10*	*8*	*10*
Drew	*15*	*2*	*3*	*18*	*20*	*1*	*17*	*2*	*19*	*3*

These 10 days of fishing represent the population.

(a) Compute the population arithmetic mean of each fisherman

(b) Compute the range of each fisherman

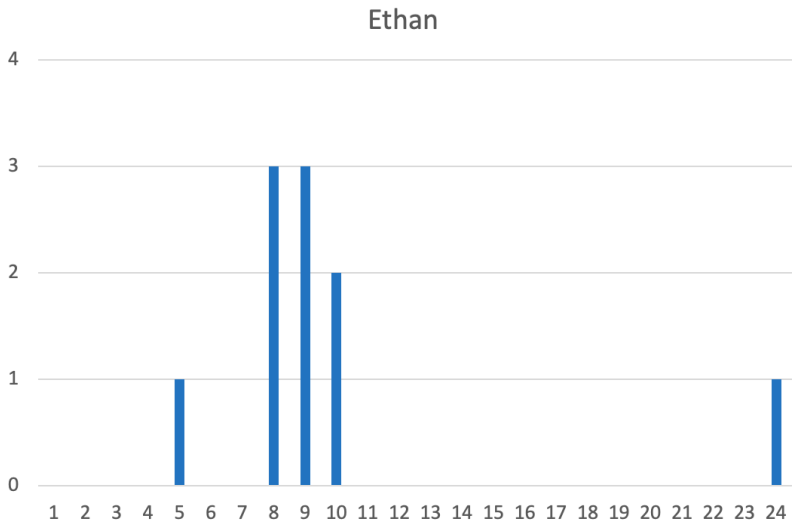
(c) Compute the population standard deviation of each fisherman

(d) Use a graph to show each fisherman's performance using the number of fish on the horizontal axis. Which fisherman is more consistent? Relate your conclusion to the standard deviation value.

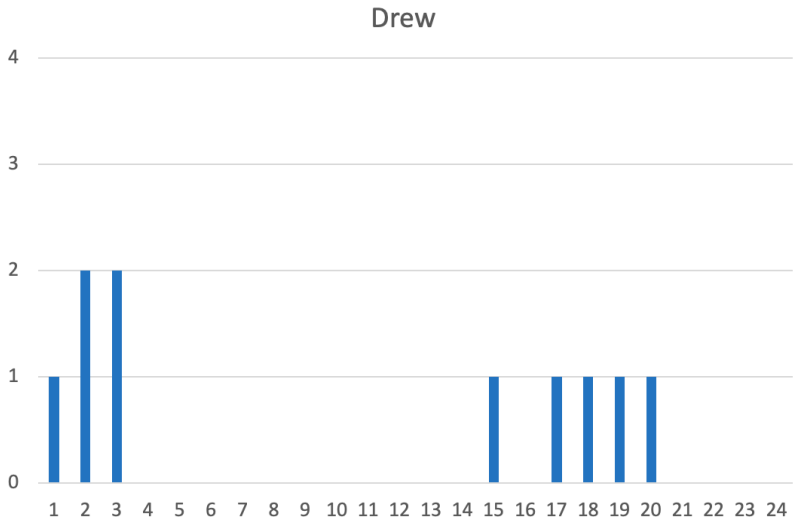
Solution. We obtain the following parameters:

Parameter	Ethan	Drew
Mean μ	10	10
Range R	19	19
Standard deviation σ	4.9	7.9

We use frequency bar charts to evaluate their performance. We obtain the following graph for Ethan



and the following for Drew



Both fishermen have the same average and range. However, from the graphs we see that Ethan he gets between 8 and 10 fish most of the time, and Drew either gets 1-3 or over 15 fish. Hence, Ethan is more consistent or less disperse. This is exactly what the standard deviation tells us, since Ethan's is smaller than Drew's. □

In the last example of the section we compute the standard deviation using a frequency distribution instead of the raw data.

Example 3.11. *Consider the sample of daycare from Example 3.4 and assume it is a population. Compute the standard deviation.*

Solution. Recall that the mean is $\mu = 3.294$ and the number of observations is 17.

Similarly to the computation of the arithmetic mean, we multiply each observation by its frequency when we sum. Specifically, we use the following table:

Age	Frequency	Squared deviation: $(x_i - \mu)^2$
2	3	$(2 - 3.294)^2 = 1.674$
3	7	$(3 - 3.294)^2 = 0.038$
4	6	$(4 - 3.294)^2 = 0.498$
5	1	$(5 - 3.294)^2 = 2.910$

Then, we compute the population standard deviation summing the product between the frequency and the squared deviation of each category (age). The variance is

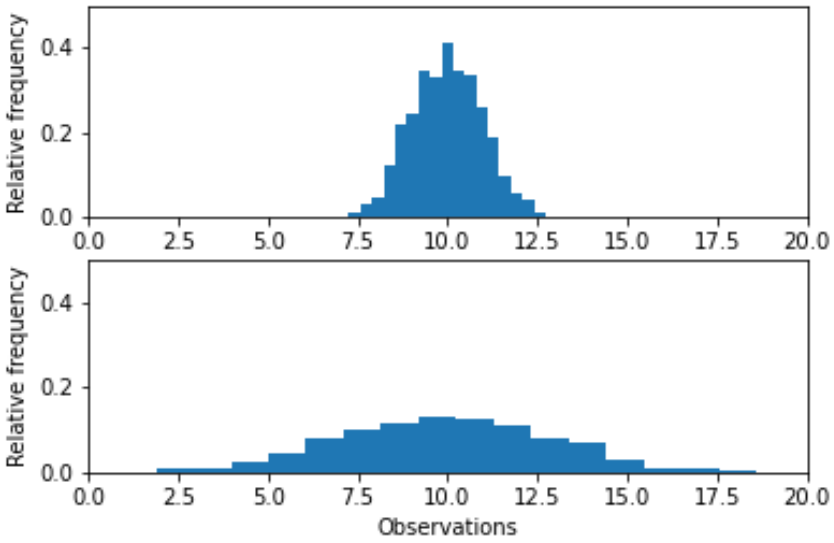
$$\sigma^2 = \frac{3 \cdot 1.674 + 7 \cdot 0.038 + 6 \cdot 0.498 + 1 \cdot 2.910}{17} = 0.658$$

and the standard deviation

$$\sigma = \sqrt{0.658} = 0.811$$

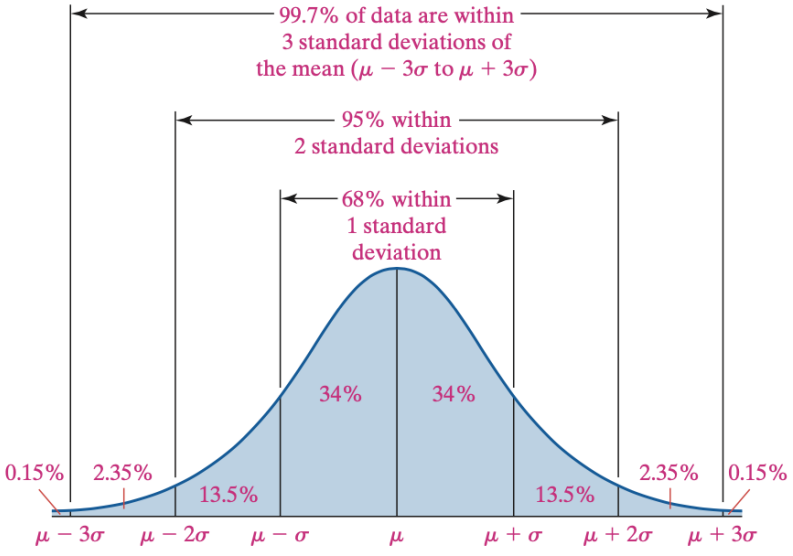
Having a small standard deviation makes sense because most of the kids that attend daycare are between 3-4 years. □

Let's study a visualization of two datasets with the same arithmetic mean and different standard deviation.



In the top panel, the standard deviation is 1 and in the bottom panel, 3. Observe how both histograms are centered around the same value (10), but the bottom panel shows more spread data.

The histograms above are bell shaped, as many data sets. In these cases, we have an empirical rule to know which % of the dataset lies within k standard deviations of the mean. We show the rule in the following picture[2]



[2]Figure 13 in Chapter 3 of the textbook

Let's do an example.

Example 3.12 (Exercise 31 from Section 3.2 of the textbook)**.** *The weight in grams of the pair of kidneys in adult males between ages of 40 and 49 has bell-shaped distribution with a mean of 325 grams and a standard deviation of 30 grams.*

(a) About 95% of kidney pairs will be between what weights?

(b) What percentage of kidney pairs weights between 235 grams and 415 grams?

(c) What percentage of kidney pairs weighs less than 235 grams or more than 415 grams?

(d) What percentage of kidney pairs weighs more than 385 grams?

Solution. We use the empirical rule. Observe that $\mu = 325$ grams and $\sigma = 30$ grams.

(a) 95% of the data is between

$$\mu - 2\sigma = 325 - 2 \cdot 30 = 265$$

and

$$\mu + 2\sigma = 325 + 2 \cdot 30 = 385$$

Hence, 95% of kidney pairs weigh between 265 grams and 385 grams.

(b) We first need to find how many standard deviations apart from the mean 235 and 415 are. We have:

$$235 = 325 - k \cdot 30 \qquad \Longrightarrow \qquad k = \frac{325 - 235}{30} = 3$$

and

$$415 = 325 + k \cdot 30 \qquad \Longrightarrow \qquad k = \frac{415 - 325}{30} = 3$$

Hence, 99.7% of the kidney pairs weigh between 235 grams and 415 grams

(c) We use our computation from the previous part, noticing that now we need to compute the remaining percentage. Then, the percentage of kidneys that weigh less than 235 grams or more than 415 grams is

$$100\% - 99.7\% = 0.3\%$$

Another way to approach this problem is looking at the empirical rule picture. Observe that 0.15% of the observations weigh less than 235 grams ($\mu - 3\sigma$) and 0.15% more than 415 grams ($\mu + 3\sigma$). Then, the percentage of kidney pairs that weigh less than 235 grams or more than 415 grams is the sum of both percentages, that is, $0.15\% + 0.15\% = 0.3\%$

(d) From part (a), recall that $385 = \mu + 2\sigma$. Then, the empirical rule says that the percentage of kidney pairs that weigh more than 385 grams is 2.35%.

□

3.3 Measures of Position and Outliers

When we are studying a dataset and we focus on a single observation, we would like to know if it is a "typical" observation or not. In other words, we would like to know how close it is to the mean. As we saw in the previous section, this measure of "how close to the mean" should be studied using the standard deviation. Specifically, we would like to know how many standard deviations away from the mean our observation is.

We formally introduce this notion in the following definition.

Definition 3.7. *The z-score represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it as follows:*

- *Population z-score:* $z = \dfrac{x - \mu}{\sigma}$
- *Sample z-score:* $z = \dfrac{x - \overline{x}}{s}$

The z-score is unitless, has mean 0 and standard deviation 1.

Observe that the z-score is positive if the observation is larger than the mean, it is negative if the observation is smaller than the mean, and is zero if the observation equals the mean.

Since the z-score is unitless and always has the same mean and standard deviation, it is also useful to compare samples that may seem different. For example, scores in an exam of the students from two different classes. Let's see some examples.

Example 3.13. *Babies born after a gestation period of 32-35 weeks have a mean weight of 2600 grams, and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and standard deviation of 470 grams.*

Suppose a 34-week gestation period baby weights 3000 grams and a 40-week gestation period baby weighs 3900 grams. Which baby weighs less relative to the gestation period?

Solution. Observe that both babies were born with 400 grams over their respective means. However, the mean and standard deviation of each type of baby are different. Then, we use the z-scores.

The z-score of the 34-gestation period baby is

$$z_{34} = \frac{3000 - 2600}{660} = 0.61$$

and the z-score of the 40-week gestation period baby is

$$z_{40} = \frac{3900 - 3500}{470} = 0.85$$

□

Example 3.14. *Bob and Mary run a marathon. Bob finished in 213 minutes, and the finishing time among all men had mean 242 minutes and standard deviation 57 minutes. Mary finished in 241 minutes, and the finishing time among all women had mean 273 minutes and standard deviation of 52 minutes. Who did better in the race?*

Solution. We compute each of their z-scores:

$$z_{Bob} = \frac{213 - 242}{57} = -0.51$$
$$z_{Mary} = \frac{241 - 273}{52} = -0.62$$

Observe that both, Bob and Mary, ran faster than the mean. Then, their z-score is negative.

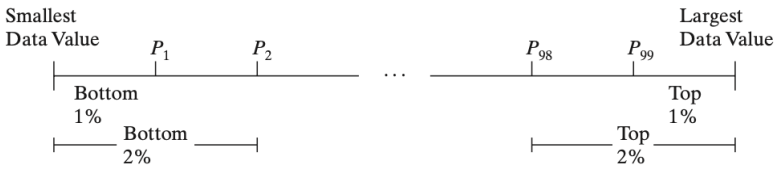
Bob is 0.52 standard deviations below the mean, and Mary is 0.62 standard deviations below the mean. Since Mary is further away from the mean than Bob, then Mary did better. □

If you recall the empirical rule, the z-scores are telling us where in the horizontal axis we are, that is, how many standard deviations apart from the data. In the case of bell-shaped distributions, this number is immediately related to the percentage of data that lies above and below a certain observation. We now learn this concept for more general datasets, that is, not necessarily bell shaped.

Definition 3.8. *The k^{th} percentile of a dataset is denoted by P_k and is a value such that $k\%$ of the observations are less than or equal to the value.*

Hence, percentiles are useful to give the relative standing of an observation with respect to the rest of the dataset.

The simplest example is the median, which corresponds to the 50^{th} percentile because 50% of the data is below the median. If we arrange the data in ascending order, then each data point represents a percentile, as shown in the figure below.



Example 3.15. *Mary just received the results of her SAT exam. Her SAT Mathematics score of 600 is in the 74^{th} percentile. What does this mean?*

Solution. It means that 74% of the SAT Math scores are below 600 and 26% are above. Then, 74% of the students who took the exam scored worse than Mary and 26% of them scored better. □

A special type of percentiles are quartiles, which we define below.

Definition 3.9. *Quartiles divide the dataset in four equal parts:*

- *The first quartile, denoted by Q_1, divides the bottom 25% of the data from the top 75%.*
- *The second quartile, denoted by Q_2, divides the bottom 50% of the data from the top 50% and is equivalent to the median.*
- *The third quartile, denoted by Q_3, divides the bottom 74% of the data from the top 25%.*

To compute the quartiles, we use a similar procedure than for the median. Specifically, we follow these 3 steps:

(1) Arrange data in ascending order

(2) Determine the median M (or second quartile Q_2)

(3) Divide the data set into halves: the observations below the median and the observations above.

- The first quartile Q_1 is the median of the bottom half of the data
- The third quartile Q_2 is the median of the top half of the data

Let's see an example.

Example 3.16. *The Highway Loss Data Institute routinely collects data on collisions coverage claims. Collisions coverage insures against physical damage to an insured individual's vehicle. The data in the table below represents a random sample of 18 collision coverage claims based on data obtained from Highway Loss Data Institute. Find and interpret the first, second, and third quartiles for collision coverage claims.*

$ 6751	*$ 9908*	*$ 3461*	*$ 2336*	*$21,147*	*$ 2332*
$ 189	*$ 1185*	*$ 370*	*$ 1414*	*$ 4668*	*$ 1953*
$ 10,034	*$ 735*	*$ 802*	*$ 618*	*$ 180*	*$ 1657*

Solution. We follow the steps above. We first order the data in ascending order and obtain

$ 180	$ 189	$ 370	$ 618	$ 735	$ 802	$ 1185	$ 1414	$ 1657
$ 1953	$ 2332	$ 2336	$ 3461	$ 4668	$ 6751	$ 9908	$ 10,034	$ 21,147

The number of data points $n = 18$ is even, so the median is computed as the average of the observations in positions $\frac{18}{2} = 9$ and $\frac{18}{2} + 1 = 10$. We obtain:

$$M = Q_2 = \frac{1657 + 1953}{2} = 1805$$

To compute the first quartile we compute the median of the first row of the data after rearranging, that is, we compute the median of

$$\$\ 180 \quad \$\ 189 \quad \$\ 370 \quad \$\ 618 \quad \$\ 735 \quad \$\ 802 \quad \$\ 1185 \quad \$\ 1414 \quad \$\ 1657$$

There are 9 data points, so the median is in the position $\frac{9+1}{2} = 5$. We obtain

$$Q_1 = 735$$

To compute the third quartile we compute the median of the second row, that is, the median of

$$\$\ 1953 \quad \$\ 2332 \quad \$\ 2336 \quad \$\ 3461 \quad \$\ 4668 \quad \$\ 6751 \quad \$\ 9908 \quad \$\ 10{,}034 \quad \$\ 21{,}147$$

We obtain:

$$Q_3 = 4668$$

Based on the quartiles, we conclude that:

- 25% of the collision claims are less than \$ 735 and 75% of the collision claims are above \$ 735
- 50% of the collision claims are less than \$ 1805
- 75% of the collision claims are less than \$ 4668, and 25% of the collision claims are above \$ 4668

□

When we started discussing measures of dispersion we talked about the range, and we said it may not be representative because it is highly affected by the extreme values of a dataset. An alternative is the range of values considering only the 50% of the central data. We properly define this measure below.

Definition 3.10. *The interquartile range is denoted by IQR is the range of the middle 50% of the observations. That is,*

$$IQR = Q_3 - Q_1$$

Let's see an example.

Example 3.17. *Compute the interquartile range of the collision claim data from Example 3.16.*

Solution. We obtain

$$IQR = Q_3 - Q_1 = 4668 - 735 = 3993$$

That is, the range of the middle 50% data is \$ 3993. □

As stated above, the interquartile range is not affected by the extreme values. Then, it is great measure of dispersion when we have skewed (left or right) data; even better than the standard deviation.

We have the following table to summarize which measure we should use for central tendency and spread.

Shape	Center	Dispersion
Symmetric	Mean	Standard deviation
Skewed (left or right)	Median	Interquartile range

While studying a dataset, we should always keep an eye for observations that are atypical. They may help us to identify mistakes or certain observations that are not well represented by the rest of the data. For example, if we are studying the price of the cars we see in a specific corner, and suddenly we see a Rolls-Royce. Should we ignore it? Should we add it to the sample? If we ignore it, we would not be truthful and we will be biasing the data. However, if we add it and treat it exactly as all the other prices, our measures of central tendency and dispersion will be considerably affected.

The answer is that we should register its price, and keep in mind that it is an extreme value.

Definition 3.11. *An outlier is an extreme observation. To check if an observation is an outlier, we use the following steps:*

(1) Determine the first and third quartile

(2) Compute the interquartile range

(3) Determine the fences, that is, cutoff points for determining outliers:

$$\text{Lower fence} = Q1 - 1.5IQR$$
$$\text{Upper fence} = Q3 + 1.5IQR$$

(4) If an observation is smaller than the lower fence, or greater than the upper fence, it is considered an outlier.

Let's close this section with an example.

Example 3.18. *Check the collision coverage claim data from Example 3.16 for outliers.*

Solution. We already have the quartiles, so we compute the fences. We obtain:

$$\text{Lower fence} = 735 - 1.5 \cdot 3993 = -5164.5$$
$$\text{Upper fence} = 4668 + 1.5 \cdot 3993 = 10,567$$

None of the prices is below the lower fence which, indeed, is negative. The value \$ 21,147 is greater than the upper fence and, hence, it is considered an outlier. □

3.4 The Five-Number Summary and Boxplots

In this section we have learned how to compute several measures that can numerically summarize the data. As we have discussed, if the data is:

- Symmetric, the best way to measure the center and dispersion is using the mean and standard deviation
- Skewed (left or right), the best way to measure center and dispersion is using the median and interquartile range

Additionally, in symmetric data the median equals the mean and the interquartile range is still very useful. Hence, if we wanted a few numbers that will give us a great summary of the data, we use the rule defined below.

Definition 3.12. *The five-number summary of a dataset corresponds to:*

(i) Smallest data value (minimum)

(ii) First quartile Q_1

(iii) Median M (or Q_2)

(iv) Third quartile Q_3

(v) Largest data value (maximum)

We have seen along the semester that, usually, graphs provide a better representation of the data because they help us visualize the observations and immediate draw some conclusions. We close this chapter with a graphic representation of the five-number summary.

Definition 3.13. *To construct a boxplot, follow these steps:*

(1) The horizontal axis represents the values in the dataset

(2) Draw a vertical line at the value of Q_1, M and Q_3 and enclose the three lines in a box.

(3) Determine the lower and upper fences, and mark them with an open squared bracket in your plot

(4) Draw a horizontal line from Q_1 to the smallest data value that is largest than the lower fence. Similarly, draw a horizontal line from Q_3 to the maximum data value that is smallest than the upper fence.

(5) Any data values that are outside the fences should be marked with a $*$ symbol

(6) Remove the temporary brackets to mark the lower and upper fence

Let's do some examples. We will use the graphical calculator to obtain the five-number summary and draw the boxplot.

Example 3.19. *The following data represent the weight in grams of a random sample of 25 Tylenol tablets.*

0.608	*0.601*	*0.606*	*0.602*	*0.611*
0.608	*0.610*	*0.610*	*0.607*	*0.600*
0.608	*0.608*	*0.605*	*0.609*	*0.605*
0.610	*0.607*	*0.611*	*0.608*	*0.610*
0.612	*0.598*	*0.600*	*0.605*	*0.603*

Source: Kelly Roe, student at Joliet Junior College

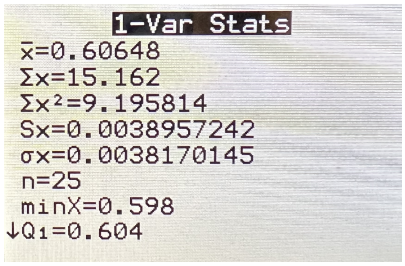
Build a boxplot of the data. Are there any outliers?

Before solving the example, let's learn how to compute the five-number summary using a TI83/84.

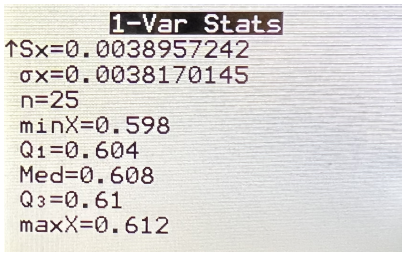
Algorithm 3.1 (Five-number summary using TI83/84)**.**

1. *Enter the raw data.*
   - 1.1. *Press the* `stat` *button*
   - 1.2. *Press* `1` *to tner the* `1:Edit...` *menu*
   - 1.3. *Enter the first observation under* `L1` *and press* `Enter`*. Repeat this step until you enter all the observations.*
   - 1.4. *Press* `2nd` *and* `mode` *buttons to quit.*
2. *Press the* `stat` *button and use the arrows to navigate until the* `CALC` *menu*
3. *Press* `1` *to enter the menu* `1-Var Stats`
4. *Make sure that the first line says* `List: L1`*, use the arrows to navigate until* `Calculate` *and press* `Enter`

Solution. Following the steps from Algorithm 3.1 we obtain the following output from the calculator:



and after pressing the down arrow, we obtain the following:



Hence, the five-number summary is:

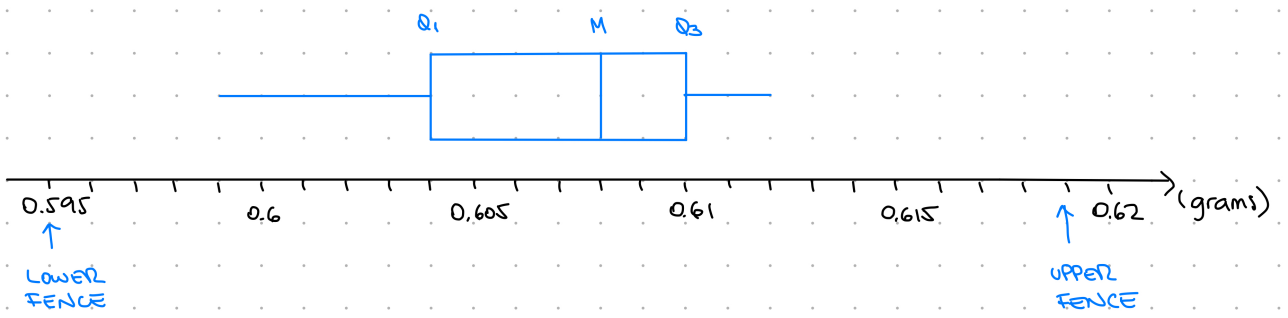
$$\text{Min} = 0.598 \qquad Q_1 = 0.604 \qquad M = 0.608 \qquad Q_3 = 0.610 \qquad \text{Max} = 0.612$$

The interquartile range, lower and upper fences are:

$$IQR = Q_3 - Q_1 = 0.610 - 0.604 = 0.006$$

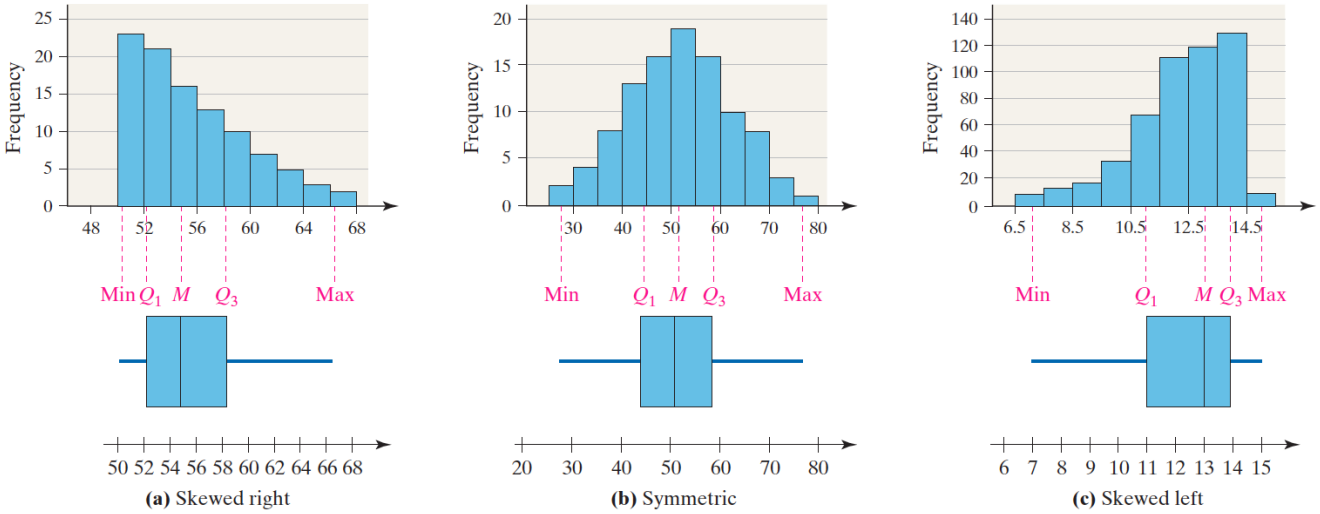
$$\text{Lower fence} = Q_1 - 1.5IQR = 0.604 - 1.5 \cdot 0.06 = 0.595$$
$$\text{Upper fence} = Q_3 + 1.5IQR = 0.610 + 1.5 \cdot 0.06 = 0.619$$

We obtain the following boxplot:



All the data points are withing the lower and upper fence, so there are no outliers. □

Boxplots are great for summarizing data and, indeed, we can also conclude whether the data is symmetric, skewed right or skewed left. The following picture[3] shows a guideline:



Observe that the location of the box with respect to the horizontal lines, and the location of the median with respect to quartiles Q_1 and Q_3 tell us whether the data is symmetric or skewed. Based on the diagrams above, we see that Example 3.19 shows a skewed left dataset because the boxplot is similar to panel (c).

Similarly to bar charts, boxplots are great to compare samples. Let's see an example.

[3] Figure 22 from the textbook

Example 3.20. *Do store-brand chocolate chip cookies have fewer chips per cookie than Keebler's Chips Deluxe Chocolate Chip Cookies? To find out, a student randomly selected 21 cookies for each brand and counted the number of chips in the cookies. The results are shown next.*

Keebler			*Store Brand*		
32	*23*	*28*	*21*	*23*	*24*
28	*28*	*29*	*24*	*25*	*27*
25	*20*	*25*	*26*	*26*	*21*
22	*21*	*24*	*18*	*16*	*24*
21	*24*	*21*	*21*	*30*	*17*
26	*28*	*24*	*23*	*28*	*31*
33	*20*	*31*	*27*	*33*	*29*

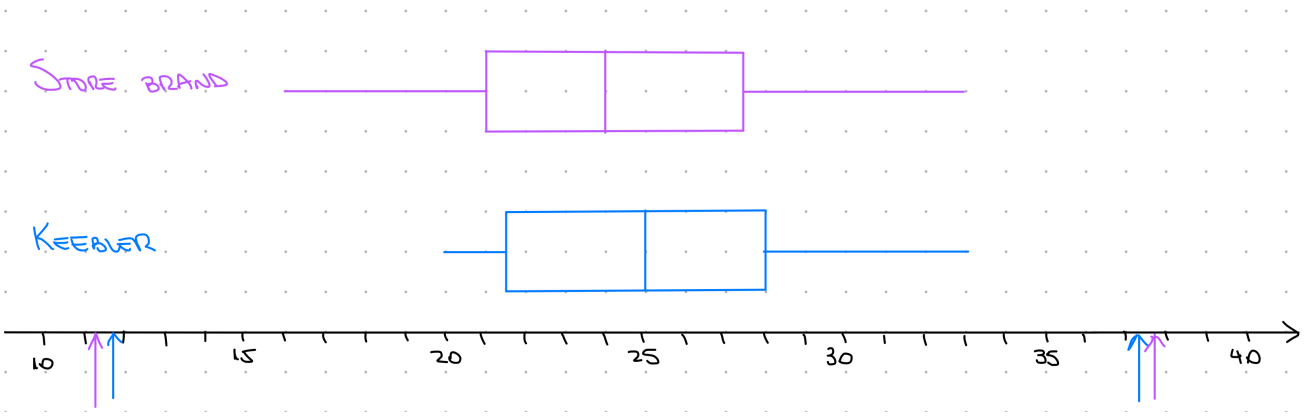
Source: Trina McNamara, student at Joliet Junior College.

Solution. We draw both boxplots to answer this question, that is, one for the Keebler's cookies and one for the store brand.

We start computing the five-number summary and the fences for each dataset. We obtain

Statistic	Keebler	Store Brand
Minimum	20	16
Q_1	21.5	21
M	25	24
Q_3	28	27.5
Maximum	33	33
IQR	6.5	6.5
Lower fence	11.75	11.25
Upper fence	37.75	37.25

We obtain the following boxplots:



The arrows under the horizontal axis indicate the lower and upper fences.

We observe that there is no substantial difference between the Keebler's cookies and the store brand. Indeed, the value of the quartiles are very similar.

If we look at the spread of all values, we see that the store brand cookies' minimum is smaller than the Keebler's minimum. The maximum of both samples are equal. Hence, we conclude that the Keebler's cookies are more consistent. □