ACCESS TO SCIENCE, ENGINEERING AND AGRICULTURE: MATHEMATICS 1

MATH00030

TRIMESTER 1 2023/2024

DR. ANTHONY BROWN

Contents

8. Statistics 1
8.1. Measures of Centre: Mean, Median and Mode 1
8.2. Measures of Spread: Standard Deviation and Interquartile Range 3
8.3. Line of Best Fit: Least Squares 7

8. Statistics

8.1. Measures of Centre: Mean, Median and Mode.

If we have a series of numbers then one thing we might ask ourselves is what is their 'average'. However there are several different measures of average and in this section we will look at the three most common.

The first is perhaps what most people think of when they say average and its definition is as follows.

Definition 8.1.1 (Arithmetic Mean)**.** Given a list of n numbers $x_1, x_2, \ldots, x_n$ (which don't have to be distinct), then their *arithmetic mean* is defined to be

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

Remark 8.1.2.

- This particular mean is called the arithmetic mean to distinguish it from several other means. Perhaps the most common of these is called the *geometric mean*. To calculate the geometric mean we multiply the numbers and then take the n'th root, rather than adding the numbers and then dividing by n. We will only deal with the arithmetic mean in this course however, so we will just call it the mean.

- The mean is usually denoted by a bar over the x.
- When we come to study probability distributions in MATH00040, we will refer to the mean of the distribution as the *expected value* of the distribution.

Here are a couple of examples.

Example 8.1.3. Find the mean of the numbers $2, 4, -2, 0, 4, 5$.
The mean is $\overline{x} = \frac{1}{6}(2 + 4 + (-2) + 0 + 4 + 5) = \frac{13}{6}$.

Example 8.1.4. Find the mean of the numbers $1, 2, 2, 1, 2, 1, 123456$.
The mean is $\overline{x} = \frac{1}{7}(1 + 2 + 2 + 1 + 2 + 1 + 123456) = \frac{123465}{7} = 17637\frac{6}{7}$.

In Example 8.1.4, I think you will agree that the mean does not really give a useful measure of what the 'average' of the numbers is.

It is to get over this sort of problem that we will define two other sorts of 'average'. While none of them is perfect, if we give all three, we will get a reasonable idea of what the 'average' is. The next definition is as follows.

Definition 8.1.5 (Median)**.** Given a list of n numbers $x_1, x_2, \ldots, x_n$ in ascending order, then their *median* is defined to be $m = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ if n is even and $m = x_{\frac{n+1}{2}}$ if n is odd.

Remark 8.1.6. This definition may seem a bit complicated but all it is saying is that the median is the number that is half way along the list, where we have to allow for the fact that if there are an even number of numbers there is no one number half way along the list, so we take the mean of the ones before and after half way.

For example if we have four numbers x_1, x_2, x_3, x_4 (in ascending order), then $n = 4$ (so it is even) and the median is defined to be $\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{\frac{4}{2}} + x_{\frac{4}{2}+1}}{2} = \frac{x_2 + x_3}{2}$, as we want.

On the other hand if n is odd, say $n = 5$, then the median of the numbers x_1, x_2, x_3, x_4, x_5 (which are in ascending order) is $x_{\frac{n+1}{2}} = x_{\frac{5+1}{2}} = x_3$, again as we want.

Here are some examples.

Example 8.1.7. Find the median of the numbers $-3, 4, -2, 0, 4, 5, 2, 1, 0, -2, 8$.
When asked to find the median of a list of numbers, then usually it is a good idea to first put them in ascending order, since this makes it much easier to see where the middle one (or middle two if there are an even number of numbers) lies. This is best done by crossing out the numbers as you write each new number down. In this case the new list is $-3, -2, -2, 0, 0, 1, 2, 4, 4, 5, 8$. Since there are 11 numbers (an odd number), the median is $m = x_{\frac{11+1}{2}} = x_6 = 1$.

Note that we don't have to write all the numbers down; since we only want x_6 we only need to write the first six down. If you are short of time then this may be a

good idea but it can also be useful to write them all down to check you have not missed any.

Example 8.1.8. Find the median of the numbers $10, 8, 2, -6, 4, 5, 2, -1, 0, -4, 5, 6$.
Here there are twelve numbers (an even number of numbers), so we are looking for $m = \dfrac{x_{\frac{12}{2}} + x_{\frac{12}{2}+1}}{2} = \dfrac{x_6 + x_7}{2}$. The first seven numbers in ascending order are $-6, -4, -1, 0, 2, 2, 4$. Thus the median is $m = \dfrac{2+4}{2} = 3$.

Example 8.1.9. Find the median of the numbers $1, 2, 2, 1, 2, 1, 123456$.
Here there are seven numbers (an odd number of numbers), so we are looking for $m = x_{\frac{7+1}{2}} = x_4$. The first four numbers in ascending order are $1, 1, 1, 2$. Thus the median is 2.

Remark 8.1.10. The numbers in Examples 8.1.4 and 8.1.9 are the same. I think you will agree that it could be argued that the median gives a better measure of a representative number in this case than the mean.

There is one other measure of 'average' that we will look at in this section.

Definition 8.1.11 (Mode)**.** Given a list of numbers, then their *mode* is defined to be the number (or numbers) that occur most frequently.

Remark 8.1.12.

- Note that in contrast to the mean and median, the mode of a list of numbers may not be unique. That is there may be more than one mode.
- The mode also makes sense for non-numeric data. For example, we could find the mode of the types of car that pass us in the street in one hour, or the mode of the first names of the people in a classroom. We won't do this in this course however.

Here are some examples.

Example 8.1.13. Find the mode of the numbers $1, -3, 0, 3, 4, 5, 5, 2$.
Here we have two fives but only one of each of the other numbers, so the mode of this list is 5.

Example 8.1.14. Find the mode of the numbers $1, 1, 2, 2, 3, 4, 5$.
Here we have two ones and two twos but only one of each of the other numbers, so 1 and 2 are both modes of this list.

Example 8.1.15. Find the mode of the numbers $-1, 0, 1, 2, 3, 4, 5$.
Here there is one of each number, so each of the given numbers is a mode of the list.

8.2. Measures of Spread: Standard Deviation and Interquartile Range.

In Section 8.1 we looked at various measures of the 'average' of a list of numbers; that is what numbers give us a good idea of what a typical number is. Another

important property of a list of numbers is how 'spread out' they are and it in this section we will look at some measures of this.

One way of measuring the spread of a list of numbers would be to calculate the mean and then calculate the mean of the distances of the numbers from the mean. However this measure is not usually used; instead the following measure is much more common.

Definition 8.2.1 (Standard Deviation)**.** Given a list of n numbers $x_1, x_2, \ldots, x_n$, then their *standard deviation* is defined to be

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

Remark 8.2.2.

- The standard deviation is usually denoted by the Greek letter sigma.
- Let us have a look at what this means in words. We first find the mean $\overline{x}$, then calculate the sum of the squares of the differences of the numbers from the mean, then divide by the number of numbers and finally take the square root.
- The is another form of standard deviation (the sample standard deviation) where we divide by $n-1$ rather than n. This is used when calculating the standard deviation of a sample rather than the whole population (in an opinion poll for example). We will not calculate this in this course.

The quantity $\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ also has a name.

Definition 8.2.3 (Variance)**.** Given a list of n numbers $x_1, x_2, \ldots, x_n$, then their *variance* is defined to be

$$\mathrm{Var}(x) = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

Here are a couple of examples of calculating the variance and standard deviation.

Example 8.2.4. Find the variance and standard deviation of the numbers $2, 4, -2, 0, 4, 5$.

Using Example 8.1.3, the mean is $\overline{x} = \dfrac{13}{6}$. Hence the variance is

$$\begin{aligned}\mathrm{Var}(x) &= \frac{\sum_{i=1}^{6}(x_i - \overline{x})^2}{6} \\ &= \frac{\left(2 - \frac{13}{6}\right)^2 + \left(4 - \frac{13}{6}\right)^2 + \left(-2 - \frac{13}{6}\right)^2 + \left(0 - \frac{13}{6}\right)^2 + \left(4 - \frac{13}{6}\right)^2 + \left(5 - \frac{13}{6}\right)^2}{6} \\ &= \frac{221}{36}.\end{aligned}$$

Thus the standard deviation is $\sigma = \sqrt{\mathrm{Var}(x)} = \sqrt{\dfrac{221}{36}} \simeq 2.478$.

Example 8.2.5. Find the variance and standard deviation of the numbers $1, 2, 2, 1, 2, 1, 123456$.

Using Example 8.1.4, the mean is $\overline{x} = \dfrac{123465}{7}$. Hence the variance is

$$\begin{aligned}\mathrm{Var}(x) =& \frac{\sum_{i=1}^{7}(x_i - \overline{x})^2}{7} \\ =& \frac{\left(1 - \frac{123465}{7}\right)^2 + \left(2 - \frac{123465}{7}\right)^2 + \left(2 - \frac{123465}{7}\right)^2 + \left(1 - \frac{123465}{7}\right)^2}{7} \\ &+ \frac{\left(2 - \frac{123465}{7}\right)^2 + \left(1 - \frac{123465}{7}\right)^2 + \left(123456 - \frac{123465}{7}\right)^2}{7} \\ \simeq& 1599678780.\end{aligned}$$

Thus the standard deviation is $\sigma = \sqrt{\mathrm{Var}(x)} \simeq \sqrt{1599678780} \simeq 39996$.

While the standard deviation in Example 8.2.5 does show that the numbers are more spread out than the numbers in Example 8.2.4, it could be argued that this is all due to one number. If the numbers are data from some experiment, for example, it is probable that the number 123456 is due to some sort of error and in this case we really need a better measure of spread. If we compare Example 8.1.9 to Example 8.1.4, we see that in this case the median gives a better measure of centre than the mean. So it might be expected that a better measure of spread will be related to the median, rather than the mean (which the standard deviation is related to). In fact this will be the case but before we give the definition of this new measure of spread, we need the following.

Definition 8.2.6 (Lower and Upper Quartile)**.** Given a list of n numbers, then their *lower quartile* and *upper quartile* are calculated as follows:

(1) List the numbers in ascending order.
(2) • If there are an even number of numbers, then split the numbers into a lower half and and upper half.

(3)
- If there are an odd number of numbers, then discard the median and split the remaining numbers into a lower half and an upper half.
- The *lower quartile*, denoted Q_1, is the median of the lower half of numbers.
- The *upper quartile*, denoted Q_3, is the median of the upper half of numbers.

Warning 8.2.7. Unfortunately there is no generally accepted way to calculate the lower and upper quartile (sometimes also called the first and third quartiles). So, if you are reading any particular book, you first have to make sure exactly what methods are being used to calculate them

Remark 8.2.8. Sometimes the median is called the second quartile and is denoted Q_2.

We can now define our new measure of spread.

Definition 8.2.9 (Interquartile Range)**.** Given a list of numbers, then their *interquartile range* is defined to be $Q_3 - Q_1$.

Here are some examples.

Example 8.2.10. Find the interquartile range of the numbers $1, 2, 2, 1, 2, 1, 123456$.
We first write the numbers in ascending order: $1, 1, 1, 2, 2, 2, 123456$. Since there are seven numbers (an odd number) the median is given by $x_{\frac{7+1}{2}} = x_4 = 2$ (note we don't actually need to know what the median is, just where it lies in the list). Again, since we have an odd number of numbers, we discard the median and split the remaining numbers into a lower half $1, 1, 1$ and an upper half $2, 2, 123456$. There are three numbers in each of these new groups (an odd number), so in each case the median is $x_{\frac{3+1}{2}} = x_2$. Thus the lower quartile is $Q_1 = 1$ and the upper quartile is $Q_3 = 2$. Hence the interquartile range is $Q_3 - Q_1 = 2 - 1 = 1$.

Remark 8.2.11. The numbers in Examples 8.2.5 and 8.2.10 are the same and I think you will agree that in some situations, the interquartile range is a better measure of spread than the standard deviation.

Example 8.2.12. Find the interquartile range of the numbers
$-2, 3, 7 - 1, -7, 0, 1, 3, 4$.
We first write the numbers in ascending order: $-7, -2, -1, 0, 1, 3, 3, 4, 7$. Since there are nine numbers (an odd number) the median is given by $x_{\frac{9+1}{2}} = x_5 = 1$. Again, since we have an odd number of numbers, we discard the median and split the remaining numbers into a lower half $-7, -2, -1, 0$ and an upper half $3, 3, 4, 7$. This time there are four numbers in each of these new groups (an even number), so in each case the median is $\dfrac{x_{\frac{4}{2}} + x_{\frac{4}{2}+1}}{2} = \dfrac{x_2 + x_3}{2}$. Thus the lower quartile is $Q_1 = \dfrac{-2 + (-1)}{2} = -\dfrac{3}{2}$ and the upper quartile is $Q_3 = \dfrac{3+4}{2} = \dfrac{7}{2}$. Hence the interquartile range is $Q_3 - Q_1 = \dfrac{7}{2} - \left(-\dfrac{3}{2}\right) = 5$.

Example 8.2.13. Find the interquartile range of the numbers
$-9, 8, 3, 5, -2, -4, -8, -3$.
We first write the numbers in ascending order: $-9, -8, -4, -3, -2, 3, 5, 8$. Since there are eight numbers (an even number) we just split the numbers into a lower half $-9, -8, -4, -3$ and an upper half $-2, 3, 5, 8$. Again there are four numbers in each of these new groups (an even number), so the median is $\dfrac{x_{\frac{4}{2}} + x_{\frac{4}{2}+1}}{2} = \dfrac{x_2 + x_3}{2}$. Hence the lower quartile is $Q_1 = \dfrac{-8 + (-4)}{2} = -6$ and the upper quartile is $Q_3 = \dfrac{3+5}{2} = 4$. Thus the interquartile range is $Q_3 - Q_1 = 4 - (-6) = 10$.

Example 8.2.14. Find the interquartile range of the numbers
$0, 3, -2, -11, 14, 4, 1, -1, -1, 4$.
We first write the numbers in ascending order: $-11, -2, -1, -1, 0, 1, 3, 4, 4, 14$. Since there are ten numbers (an even number) we just split the numbers into a lower half $-11, -2, -1, -1, 0$ and an upper half $1, 3, 4, 4, 14$. This time there are five numbers in each of these new groups (an odd number), so the median is $x_{\frac{5+1}{2}} = x_3$. Hence the lower quartile is $Q_1 = -1$ and the upper quartile is $Q_3 = 4$. Thus the interquartile range is $Q_3 - Q_1 = 4 - (-1) = 5$.

8.3. Line of Best Fit: Least Squares.

So far in this chapter we have just looked at situations where our data is a list of numbers. In this section we will look at the case where our data is a list of points in the x-y plane. Given a list of points like this, we might ask ourselves what is the 'best' line we can draw to represent these points.

There are several ways we could calculate the line of 'best' fit. For example, if we have n points which we will denote by (x_i, y_i), $i = 1, 2, \ldots, n$, then one way to find the 'best' line would be to find a line $y = mx + c$ such that the quantity $\sum_{i=1}^{n} |mx_i + c - y_i|$ is minimized. That is the line such that the sum of the vertical distances of the points from the line is minimized.

However a much more common method is to minimize the sum of the squares of these distances. Note that this is somewhat similar to the calculation of the variance, where we find the sum of the squares of the distances from the mean. The actual derivation of the formulae that allow us to calculate m and c using this method is quite complicated, so I will just state them.

Theorem 8.3.1 (Line of best fit: least squares)**.** *Given a list of n points (x_i, y_i), $i = 1, 2, \ldots, n$, then the line of best fit $y = mx + c$ (calculated using the least squares method) is found using the formulae:*

$$m = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \textit{and} \quad c = \overline{y} - m\overline{x}.$$

Here are a couple of examples.

Example 8.3.2. Find the line of best fit using the least squares method with the points $(-3,-1),(-2,0),(1,1),(3,2),(4,3),(6,3),(8,2),(11,2),(12,5)$ and $(14,5)$. Plot the line of best fit and the points on a graph.
In this case $n = 10$ and

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{10} x_i = -3 + (-2) + 1 + 3 + 4 + 6 + 8 + 11 + 12 + 14 = 54$$
$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{10} y_i = -1 + 0 + 1 + 2 + 3 + 3 + 2 + 2 + 5 + 5 = 22$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i =& \sum_{i=1}^{10} x_i y_i \\ =&(-3)(-1) + (-2)(0) + (1)(1) + (3)(2) + (4)(3) \\ &+ (6)(3) + (8)(2) + (11)(2) + (12)(5) + (14)(5) \\ =&3 + 0 + 1 + 6 + 12 + 18 + 16 + 22 + 60 + 70 \\ =&208.\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 =& \sum_{i=1}^{10} x_i^2 \\ =&(-3)^2 + (-2)^2 + 1^2 + 3^2 + 4^2 + 6^2 + 8^2 + 11^2 + 12^2 + 14^2 \\ =&9 + 4 + 1 + 9 + 16 + 36 + 64 + 121 + 144 + 196 \\ =&600.\end{aligned}$$

Hence

$$m = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{10(208) - (54)(22)}{10(600) - 54^2} = \frac{892}{3084} = \frac{223}{771} \simeq 0.289$$

and

$$c = \overline{y} - m\overline{x} = \frac{\sum_{i=1}^{10} y_i}{10} - m\frac{\sum_{i=1}^{10} x_i}{10} = \frac{22}{10} - \frac{223}{771} \times \frac{54}{10} = \frac{164}{257} \simeq 0.638.$$

Thus the line of best fit is $y = \frac{223}{771}x + \frac{164}{257}$.
The points and the graph are shown in Figure 1.

Remark 8.3.3. As you can see in Figure 1, the line does fit the points quite well.

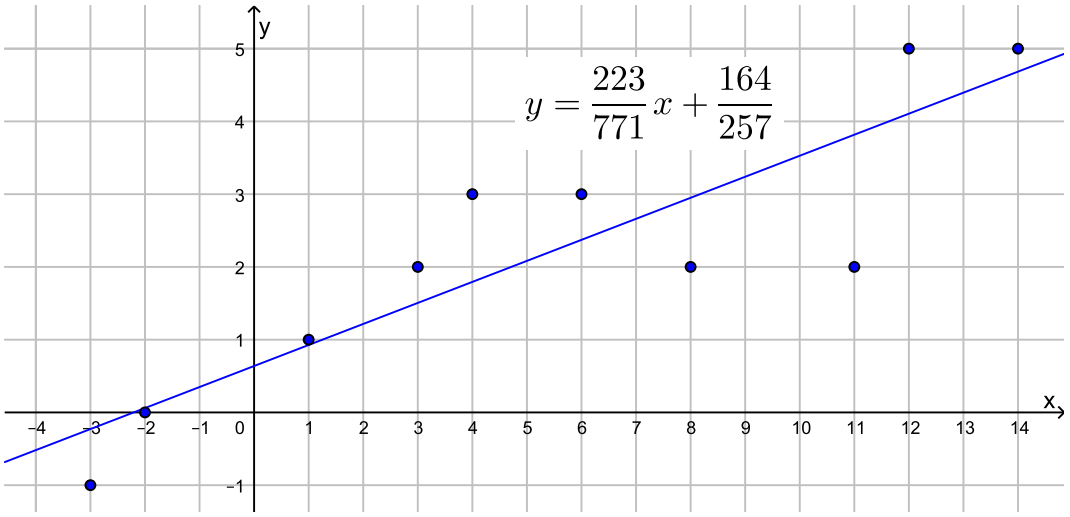

FIGURE 1. The Line of Best Fit and Points From Example 8.3.2.

Example 8.3.4. Find the line of best fit using the least squares method with the points $(-3,4),(-1,4),(-1,3),(1,2),(2,3),(3,1),(6,0),(8,1),(9,-1),(11,-1)$ and $(13,-2)$. Plot the line of best fit and the points on a graph.
In this case $n = 11$ and

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{11} x_i = -3 + (-1) + (-1) + 1 + 2 + 3 + 6 + 8 + 9 + 11 + 13 = 48$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{11} y_i = 4 + 4 + 3 + 2 + 3 + 1 + 0 + 1 + (-1) + (-1) + (-2) = 14$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i =& \sum_{i=1}^{11} x_i y_i \\ =&(-3)(4) + (-1)(4) + (-1)(3) + (1)(2) + (2)(3) + (3)(1) \\ &+ (6)(0) + (8)(1) + (9)(-1) + (11)(-1) + (13)(-2) \\ =&-12 + (-4) + (-3) + 2 + 6 + 3 + 0 + 8 + (-9) + (-11) + (-26) \\ =&-46.\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 =& \sum_{i=1}^{11} x_i^2 \\ =&(-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 2^2 + 3^2 + 6^2 + 8^2 + 9^2 + 11^2 + 13^2 \\ =&9 + 1 + 1 + 4 + 9 + 36 + 64 + 81 + 121 + 169 \\ =&496.\end{aligned}$$

Hence

$$\begin{aligned} m &= \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{11(-46) - (48)(14)}{11(496) - 48^2} \\ &= \frac{-1178}{3152} \\ &= -\frac{589}{1576} \\ &\simeq -0.374, \end{aligned}$$

and

$$c = \overline{y} - m\overline{x} = \frac{\sum_{i=1}^{11} y_i}{11} - m\frac{\sum_{i=1}^{11} x_i}{11} = \frac{14}{11} - \left(-\frac{589}{1576}\right) \times \frac{48}{11} = \frac{572}{197} \simeq 2.904.$$

Thus the line of best fit is $y = -\frac{589}{1576}x + \frac{572}{197}$.
The points and the graph are shown in Figure 2.

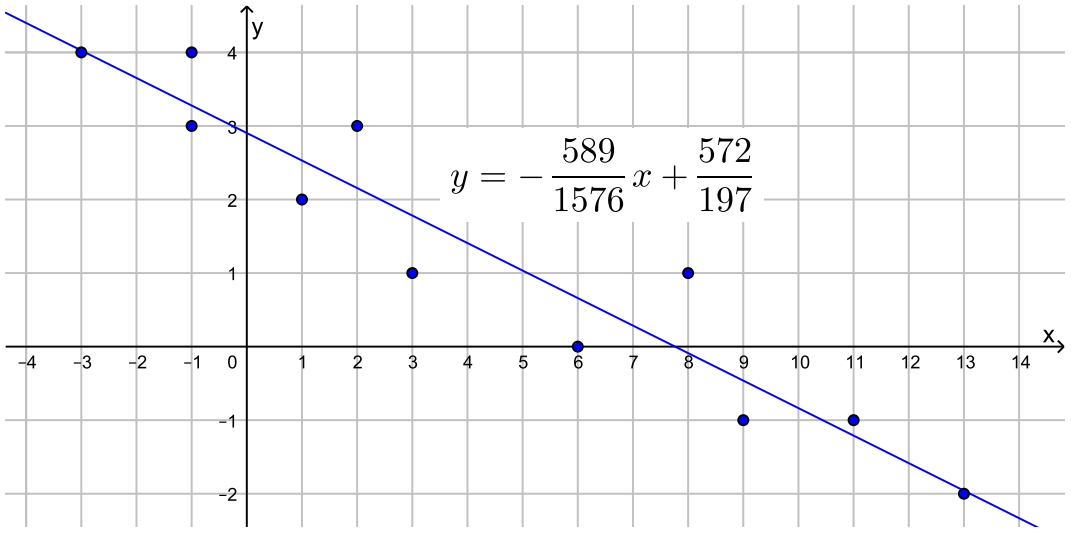

FIGURE 2. The Line of Best Fit and Points From Example 8.3.4.