Numerical Description of Data

Learning Objectives

- sample mean
- sample variance
- sample standard deviation
- sample median
- lower and upper quartiles
- interquartile range *(IQR)*
- sample mode

In the previous section we looked at some graphical and tabular techniques for describing a data set. We shall now consider some numerical characteristics of a set of measurements. Suppose that we have a sample with values $x_1, x_2, \ldots, x_n$. There are many characteristics associated with this data set, for example, the central tendency and variability. A measure of the central tendency is given by the sample mean, median, or mode, and the measure of dispersion or variability is usually given by the sample variance or sample standard deviation or interquartile range.

Definition 1. *Let* $x_1, x_2, \ldots, x_n$ *be a set of sample values. Then the* **sample mean** *(or* **empirical mean***)* $\overline{x}$ *is defined by*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The **sample variance** *is defined by*

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

The **sample standard deviation** *is*

$$s = \sqrt{s^2}.$$

The sample variance s^2 and the sample standard deviation s both are measures of the variability or "scatteredness" of data values around the sample mean $\overline{x}$. Larger the variance, more is the spread. We note that s^2 and s are both nonnegative. One question we may ask is "why not just take the sum of the differences $(x_i - \overline{x})$ as a measure of variation?" The answer lies in the following result which shows that if we add up all deviations about the sample mean, we always get a zero value.

Theorem 1. *For a given set of measurements* $x_1, x_2, \ldots, x_n$, *let* $\overline{x}$ *be the sample mean. Then*

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

Proof. Since $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$, we have $\sum_{i=1}^{n} x_i = n\overline{x}$. Now

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} \\ &= n\overline{x} - n\overline{x} = 0.\end{aligned}$$

Thus although there may be a large variation in the data values, $\sum_{i=1}^{n}(x_i - \overline{x})$ as a measure of spread would always be zero, implying no variability. So it is not useful as a measure of variability. $\square$

Sometimes we can simplify the calculation of the sample variance s^2 by using the following computational formula:

$$s^2 = \frac{\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}{(n-1)}.$$

If the data set has a large variation with some extreme values (called outliers), the mean may not be a very good measure of the center. For example, average salary may not be a good indicator of the financial well-being of the employees of a company if there is a huge difference in pay between support personnel and management personnel. In that case, one could use the median as a measure of the center, roughly 50% of data fall below and 50% above. The median is less sensitive to extreme data values.

Definition 2. *For a data set, the* **median** *is the middle number of the ordered data set. If the data set has an even number of elements, then the median is the average of the middle two numbers. The* **lower quartile** *is the middle number of the half of the data below the median, and the* **upper quartile** *is the middle number of the half of the data above the median. We will denote*

$$Q_1 = \textit{lower quartile}$$

$$Q_2 = M = \textit{middle quartile}\,(median)$$

$$Q_3 = \textit{upper quartile}$$

The difference between the quartiles is called **interquartile range** *(IQR).*

$$IQR = Q_3 - Q_1.$$

A possible outlier (mild outlier) will be any data point that lies below

$$Q_1 - 1.5(IQR) \textit{ or above } Q_3 + 1.5(IQR).$$

Note that the IQR is unaffected by the positions of those observations in the smallest 25% or the largest 25% of the data.

Definition 3. **Mode** *is the most frequently occurring member of the data set. If all the data values are different, then by definition, the data set has no mode.*

Example 1.

The following data give the time in months from hire to promotion to manager for a random sample of 25 software engineers from all software engineers employed by a large telecommunications firm.

5	7	229	453	12	14	18	14	14	483
22	21	25	23	24	34	37	34	49	64
47	67	69	192	125					

Calculate the mean, median, mode, variance, and standard deviation for this sample.

Solution

The sample mean is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 83.28 \textit{ months.}$$

To obtain the median, first arrange the data in ascending order:

5	7	12	14	14	14	18	21	22	23
24	25	34	34	37	47	49	64	67	69
125	192	229	453	483					

Now the median is the thirteenth number which is 34 months.

Since 14 occurs most often (thrice), the mode is 14 months.

The sample variance is

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \frac{1}{24}\left[(5 - 83.28)^2 + \cdots + (125 - 83.28)^2\right] \\
&= 16{,}478.
\end{aligned}
$$

and the sample standard deviation is, $s = \sqrt{s^2} = 128.36$ *months. Thus, we have sample mean* $\overline{x} = 83.28$ *months, median* $= 34$ *months, and mode = 14 months. Note that the mean is very much different from the other two measures of center because of a few large data values. Also, the sample variance* $s^2 = 16{,}478$ *months, and the sample standard deviation* $s = 128.36$ *months.*

Example 2.

For the data of Example 1, find lower and upper quartiles, median, and interquartile range (IQR). Check for any outliers.

Solution

Arrange the data in an ascending order.

5	7	12	14	14	14	18	21	22	23
24	25	34	34	37	47	49	64	67	69
125	192	229	453	483					

Then the median M *is the middle (13th) data value,* $M = Q_2 = 34$. *The lower quartile is the middle number below the median,* $Q_1 = [(14 + 18)/2] = 16$. *The upper quartile,* $Q_3 = [(67 + 69)/2] = 68$.
The interquartile range, $(IQR) = Q_3 - Q_1 = 68 - 16 = 52$.
To test for outliers, compute

$$Q_1 - 1.5(IQR) = 16 - 1.5(52) = -62$$

and

$$Q_3 + 1.5(IQR) = 68 + 1.5(52) = 146.$$

Then all the data that fall above 146 are possible outliers. None is below -62. *Therefore the outliers are 192, 229, 453, and 483.*

We have remarked earlier that the mean as a measure of central location is greatly affected by the extreme values or outliers. A robust measure of central location (a measure that is relatively unaffected by outliers) is the trimmed mean. For $0 \leq \alpha \leq 1$, a $100\alpha\%$ trimmed mean is found as follows: Order the data, and then discard the lowest $100\alpha\%$and the highest $100\alpha\%$of the data values. Find the mean of the rest of the data values. We denote the $100\alpha\%$ trimmed mean by $\bar{x}\alpha$. We illustrate the trimmed mean concept in the following example.

Example 3.

For the data set representing the number of children in a random sample of 10 families in a neighborhood, find the 10% trimmed mean ($\alpha = 0.1$).

$$1 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 9 \quad 1 \quad 6 \quad 2$$

Solution

Arrange the data in ascending order.

$$1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 6 \quad 9$$

The data set has 10 elements. Discarding the lowest 10% (10% of 10 is 1) and discarding the highest 10% of the data values, we obtain the trimmed data set as

$$1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 6$$

The 10% trimmed mean is

$$\bar{x}_{0.1} = \frac{1+2+2+2+2+3+3+6}{8} = 2.6.$$

Note that the mean for the data in the previous example without removing any observations is 3.1, which is different from the trimmed mean.

Example 4.

There are 9 members of the Community Youth Leadership Board. Find the range and interquartile range of their ages: 22, 16, 24, 17, 16, 25, 20, 19, 26.

greatest value – least value = 26 – 16 Find the range.

= 10

median Find the median.

16 16 17 19 20 22 24 25 26

$Q_1 = \frac{(16 + 17)}{2} = 16.5$ $Q_3 = \frac{(24 + 25)}{2} = 24.5$ Find Q_1 and Q_3.

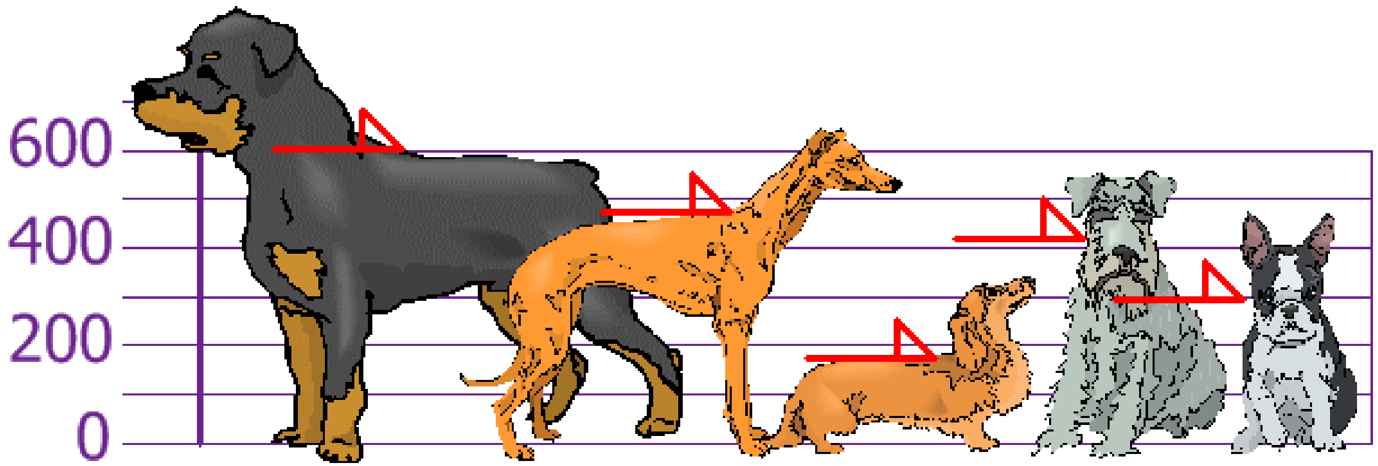
$Q_3 - Q_1 = 24.5 - 16.5$ Find the interquartile range.

$= 8$

The range is 10 years. The interquartile range is 8 years.

Example 5.

You and your friends have just measured the heights of your dogs (in millimeters):



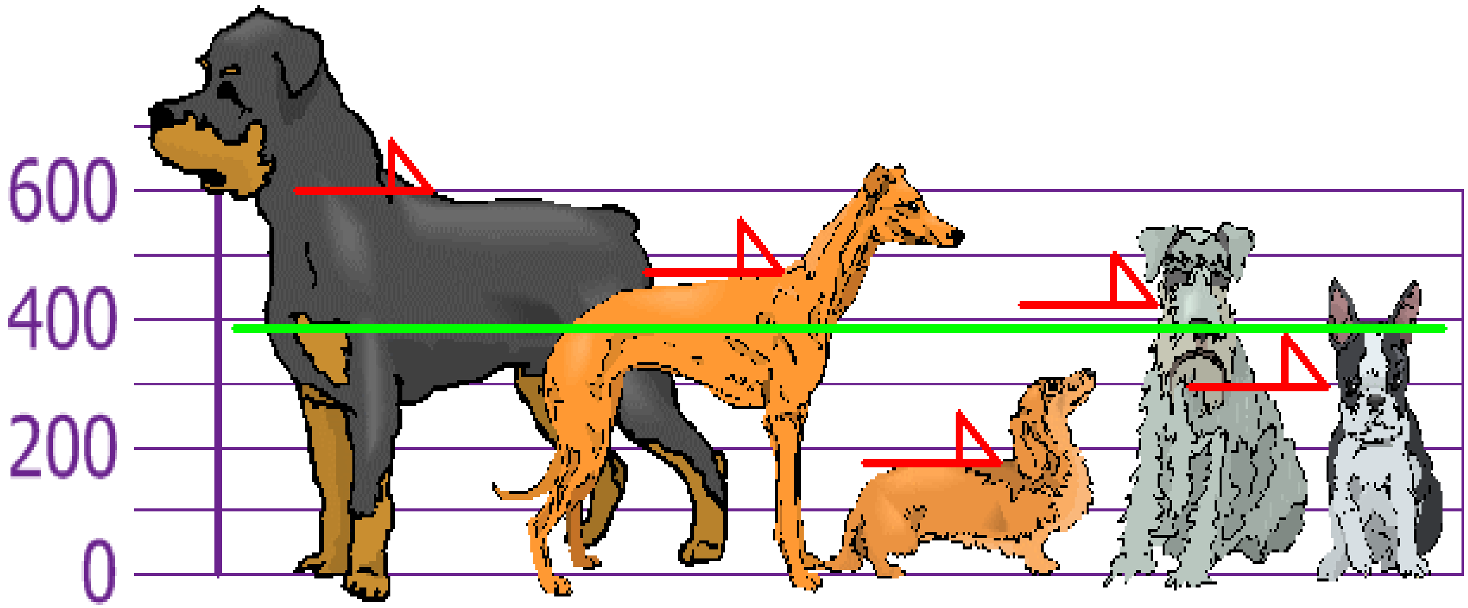
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

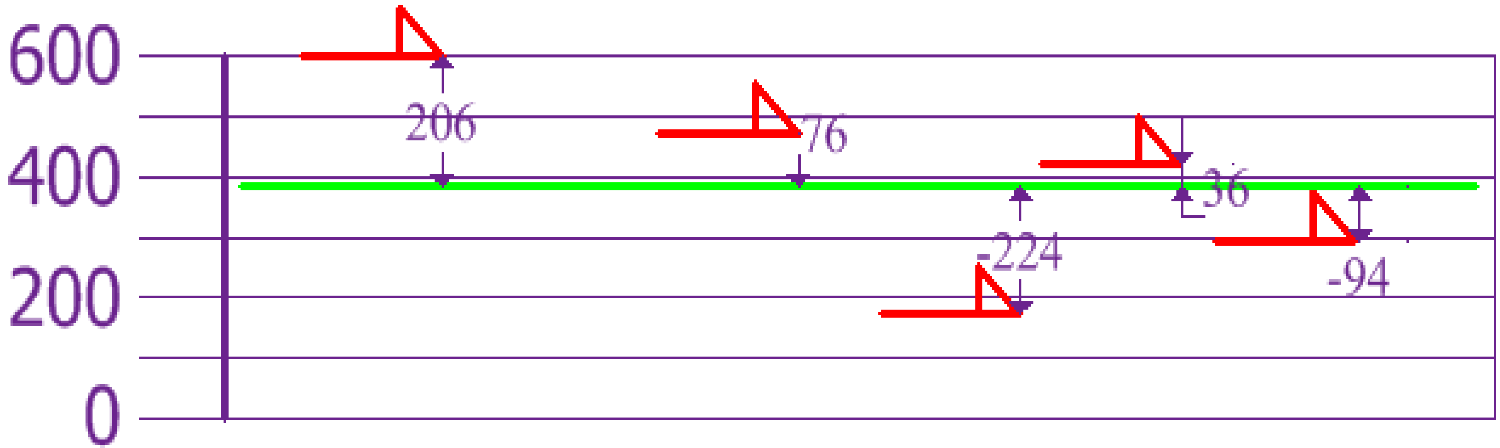
Your first step is to find the Mean:

$$\bar{x} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



Now, we calculate each dogs difference from the Mean:



To calculate the Variance, take each difference, square it, and then average the result:

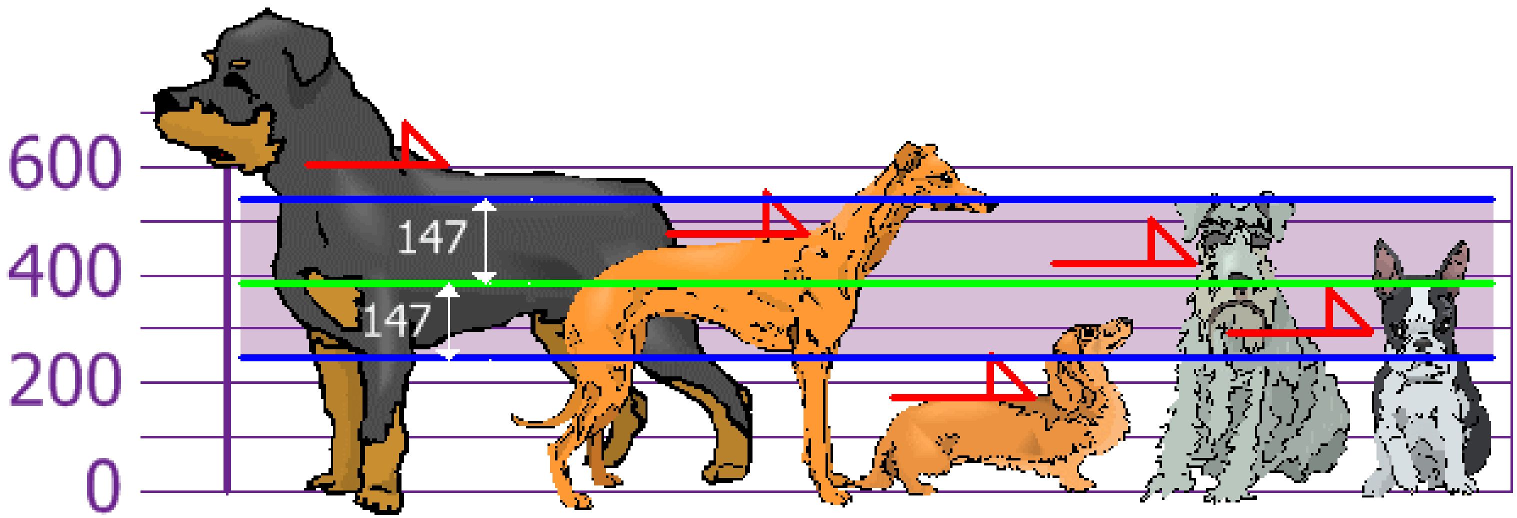
$$
\begin{aligned}
s^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\
&= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\
&= \frac{108,520}{5} = 21,704
\end{aligned}
$$

So, the Variance is 21,704.

And the Standard Deviation is just the square root of Variance, so:

$$s = \sqrt{21{,}704} = 147.32... = 147 \text{ (to the nearest mm)}$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small. Rottweilers **are** tall dogs. And Dachshunds **are** a bit short ... but don't tell them!

Steps to Finding Standard Deviation

1. Find the mean of the set of data: $\overline{x}$
2. Find the difference between each value and the mean: $x - \overline{x}$
3. Square the difference: $(x - \overline{x})^2$
4. Find the average (mean) of these squares:

$$\frac{\sum (x - \overline{x})^2}{n - 1}$$

The Population: divide by ***n*** when calculating Variance
A Sample: divide by ***n-1*** when calculating Variance

5. Take the square root to find the standard deviation

$$\sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$$

Example 6.

Consider two students, and their scores for the five exams.
Student A has scores 84, 86, 83, 85, and 87.
Student B has scores 90, 75, 94, 68, and 98.

Calculate the mean and Standard Deviation for both Student A and Student B

Calculation for Student A.

x	$(x - \overline{x})$	$(x - \overline{x})^2$
84	(84 – 85)	$(-1)^2 = 1$
86	(86 – 85)	$(1)^2 = 1$
83	(83 – 85)	$(-2)^2 = 4$
85	(85 – 85)	$(0)^2 = 0$
87	(87 – 85)	$(2)^2 = 4$
$\sum = 425$		$\sum = 10$
$\overline{x} = 425/5=85$		$s^2=10/(5-1)=2.5$

$$s = \sqrt{2.5} \approx 1.58$$

Calculation for Student B.

x	$(x - \overline{x})$	$(x - \overline{x})^2$
90		
75		
94		
68		
98		
$\sum =$		$\sum =$
$\overline{x} =$		$s^2 =$

$s =$

Calculation for Student B.

x	$(x-\overline{x})$	$(x-\overline{x})^2$
90	5	25
75	-10	100
94	9	81
68	-17	289
98	13	169
$\sum$ = 425		$\sum$ = 664
$\overline{x}$ = 425/5=85		s^2=664/(5-1)=166

$$s = \sqrt{166} \approx 12.88$$

Example 7.

The weights in pounds of the five players of the AUCA football team are 210, 245, 220, 230, and 225. Find the standard deviation of the weights.

x	$(x-\overline{x})$	$(x-\overline{x})^2$
210		
245		
220		
230		
225		
$\sum =$		$\sum =$
$\overline{x} =$		$s^2 =$

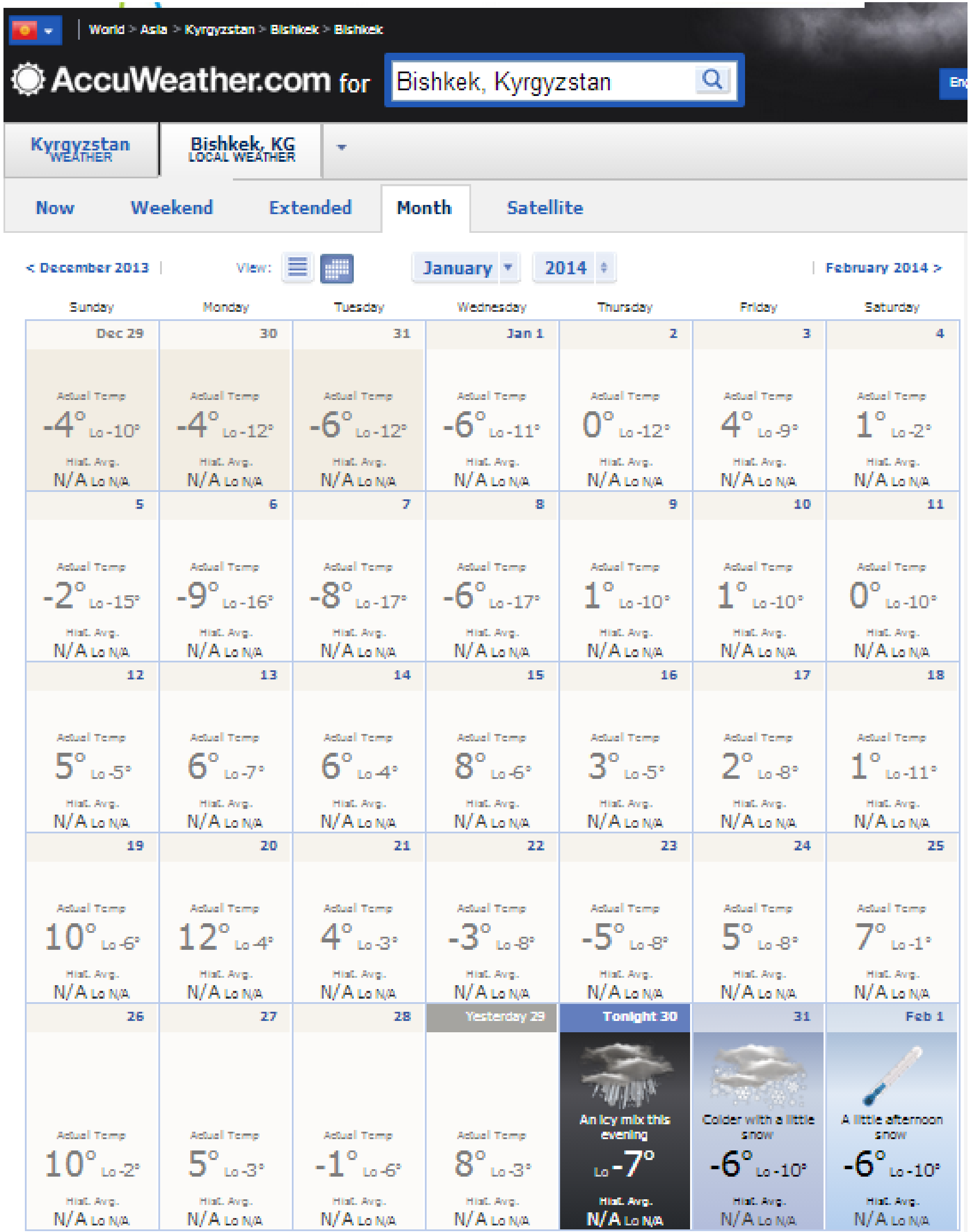
$$s =$$

Example 7.

The weights in pounds of the five players of the AUCA football team are 210, 245, 220, 230, and 225. Find the standard deviation of the weights.

x	$(x-\overline{x})$	$(x-\overline{x})^2$
210	-16	256
245	19	361
220	-6	36
230	4	16
225	-1	1
$\sum$ = 1130		$\sum$ = 670
$\overline{x}$ = 1130/5=226		s^2=670/(5-1)=167.5

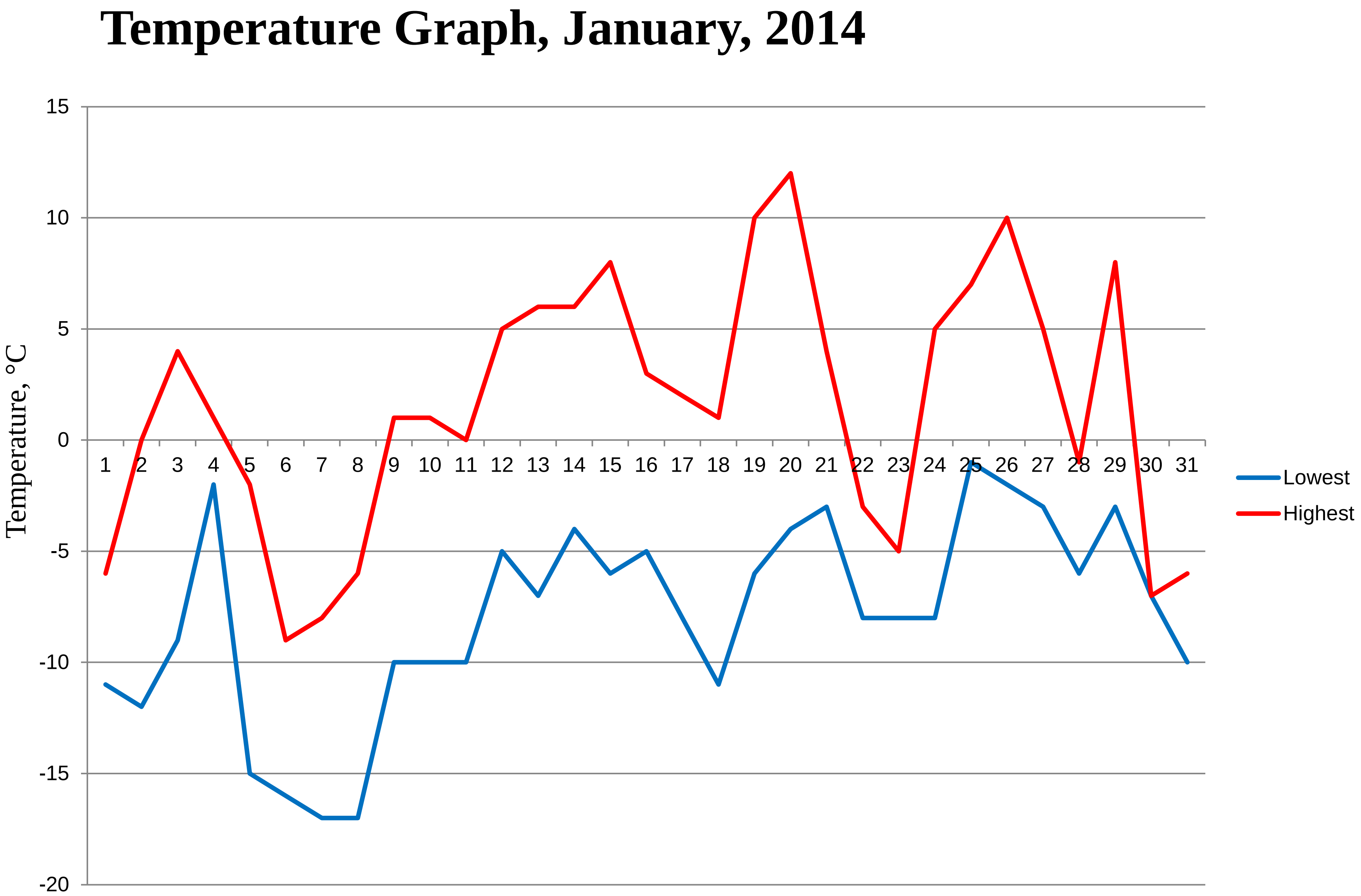
$$s=\sqrt{167.5}\approx 12.94$$



Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Dec 29 Actual Temp -4° Lo -10° Hist. Avg. N/A Lo N/A	30 Actual Temp -4° Lo -12° Hist. Avg. N/A Lo N/A	31 Actual Temp -6° Lo -12° Hist. Avg. N/A Lo N/A	Jan 1 Actual Temp -6° Lo -11° Hist. Avg. N/A Lo N/A	2 Actual Temp 0° Lo -12° Hist. Avg. N/A Lo N/A	3 Actual Temp 4° Lo -9° Hist. Avg. N/A Lo N/A	4 Actual Temp 1° Lo -2° Hist. Avg. N/A Lo N/A
5 Actual Temp -2° Lo -15° Hist. Avg. N/A Lo N/A	6 Actual Temp -9° Lo -16° Hist. Avg. N/A Lo N/A	7 Actual Temp -8° Lo -17° Hist. Avg. N/A Lo N/A	8 Actual Temp -6° Lo -17° Hist. Avg. N/A Lo N/A	9 Actual Temp 1° Lo -10° Hist. Avg. N/A Lo N/A	10 Actual Temp 1° Lo -10° Hist. Avg. N/A Lo N/A	11 Actual Temp 0° Lo -10° Hist. Avg. N/A Lo N/A
12 Actual Temp 5° Lo -5° Hist. Avg. N/A Lo N/A	13 Actual Temp 6° Lo -7° Hist. Avg. N/A Lo N/A	14 Actual Temp 6° Lo -4° Hist. Avg. N/A Lo N/A	15 Actual Temp 8° Lo -6° Hist. Avg. N/A Lo N/A	16 Actual Temp 3° Lo -5° Hist. Avg. N/A Lo N/A	17 Actual Temp 2° Lo -8° Hist. Avg. N/A Lo N/A	18 Actual Temp 1° Lo -11° Hist. Avg. N/A Lo N/A
19 Actual Temp 10° Lo -6° Hist. Avg. N/A Lo N/A	20 Actual Temp 12° Lo -4° Hist. Avg. N/A Lo N/A	21 Actual Temp 4° Lo -3° Hist. Avg. N/A Lo N/A	22 Actual Temp -3° Lo -8° Hist. Avg. N/A Lo N/A	23 Actual Temp -5° Lo -8° Hist. Avg. N/A Lo N/A	24 Actual Temp 5° Lo -8° Hist. Avg. N/A Lo N/A	25 Actual Temp 7° Lo -1° Hist. Avg. N/A Lo N/A
26 Actual Temp 10° Lo -2° Hist. Avg. N/A Lo N/A	27 Actual Temp 5° Lo -3° Hist. Avg. N/A Lo N/A	28 Actual Temp -1° Lo -6° Hist. Avg. N/A Lo N/A	Yesterday 29 Actual Temp 8° Lo -3° Hist. Avg. N/A Lo N/A	Tonight 30 An icy mix this evening Lo -7° Hist. Avg. N/A Lo N/A	31 Colder with a little snow -6° Lo -10° Hist. Avg. N/A Lo N/A	Feb 1 A little afternoon snow -6° Lo -10° Hist. Avg. N/A Lo N/A

Example 8.

The table gives the temperature for January 2014 in Bishkek, Kyrgyzstan.
Find the average highest temperature in January and the average lowest temperature in January. Also find the standard deviation from the mean temperature in both cases.


Temperature Graph, January, 2014
Temperature, °C
15
10
5
0
-5
-10
-15
-20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Lowest
Highest

Example 8. Calculations for highest temperature

day	x	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-6		
2	0		
3	4		
4	1		
5	-2		
6	-9		
7	-8		
8	-6		
9	1		
10	1		
11	0		
12	5		
13	6		
14	6		
15	8		
16	3		

day	x	$(x-\bar{x})$	$(x-\bar{x})^2$
17	2		
18	1		
19	10		
20	12		
21	4		
22	-3		
23	-5		
24	5		
25	7		
26	10		
27	5		
28	-1		
29	8		
30	-7		
31	-6		
	$\sum =$		$\sum =$

$\bar{x} =$

$s^2 =$

$s =$

Example 8. Calculations for highest temperature

day	x	$(x-\overline{x})$	$(x-\overline{x})^2$
1	-6	-7.48	56.01
2	0	-1.48	2.20
3	4	2.52	6.33
4	1	-0.48	0.23
5	-2	-3.48	12.14
6	-9	-10.48	109.91
7	-8	-9.48	89.94
8	-6	-7.48	56.01
9	1	-0.48	0.23
10	1	-0.48	0.23
11	0	-1.48	2.20
12	5	3.52	12.36
13	6	4.52	20.40
14	6	4.52	20.40
15	8	6.52	42.46
16	3	1.52	2.30

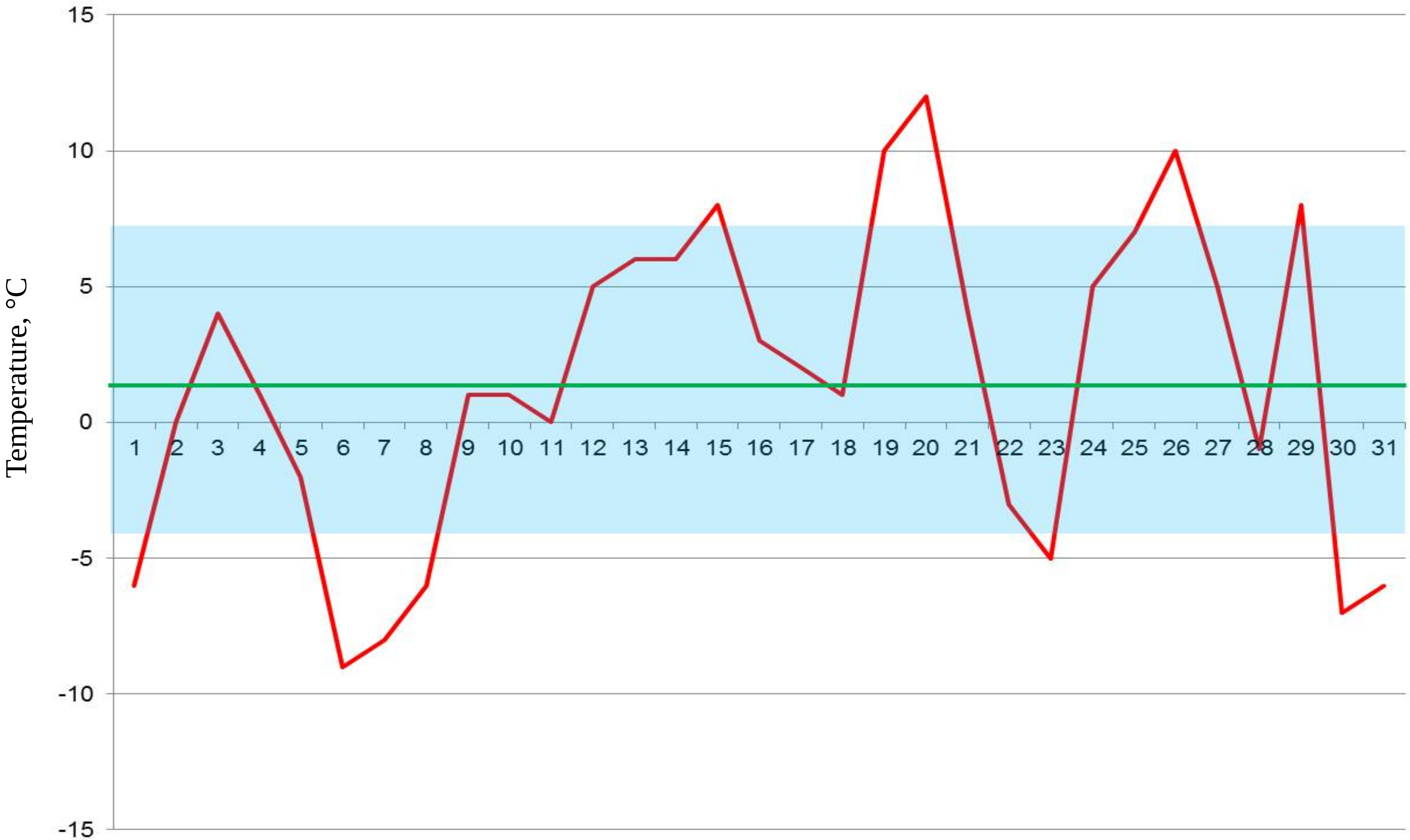
day	x	$(x-\overline{x})$	$(x-\overline{x})^2$
17	2	0.52	0.27
18	1	-0.48	0.23
19	10	8.52	72.52
20	12	10.52	110.59
21	4	2.52	6.33
22	-3	-4.48	20.11
23	-5	-6.48	42.04
24	5	3.52	12.36
25	7	5.52	30.43
26	10	8.52	72.52
27	5	3.52	12.36
28	-1	-2.48	6.17
29	8	6.52	42.46
30	-7	-8.48	71.98
31	-6	-7.48	56.01
	$\sum = 46$		$\sum = 989.74$

$\overline{x} = 46/31 = 1.48 \approx 1.5$ The average highest temperature is 1.5 ° in January, 2014.

$s^2 = 989.74/(31-1) = 32.99$

$s = \sqrt{32.99} \approx 5.7$ The standard deviation for highest temperature is 5.7 ° in January, 2014.

Highest Temperature Graph, January, 2014



Example 8. Calculations for lowest temperature

day	x	$(x - \overline{x})$	$(x - \overline{x})^2$
1	-11		
2	-12		
3	-9		
4	-2		
5	-15		
6	-16		
7	-17		
8	-17		
9	-10		
10	-10		
11	-10		
12	-5		
13	-7		
14	-4		
15	-6		
16	-5		

day	x	$(x - \overline{x})$	$(x - \overline{x})^2$
17	-8		
18	-11		
19	-6		
20	-4		
21	-3		
22	-8		
23	-8		
24	-8		
25	-1		
26	-2		
27	-3		
28	-6		
29	-3		
30	-7		
31	-10		
	$\sum =$		$\sum =$

$\overline{x} =$

$s^2 =$

$s =$

Example 8. Calculations for lowest temperature

day	x	$(x-\overline{x})$	$(x-\overline{x})^2$
1	-11	-3.13	9.79
2	-12	-4.13	17.05
3	-9	-1.13	1.27
4	-2	5.87	34.47
5	-15	-7.13	50.82
6	-16	-8.13	66.08
7	-17	-9.13	83.34
8	-17	-9.13	83.34
9	-10	-2.13	4.53
10	-10	-2.13	4.53
11	-10	-2.13	4.53
12	-5	2.87	8.24
13	-7	0.87	0.76
14	-4	3.87	14.98
15	-6	1.87	3.50
16	-5	2.87	8.24

day	x	$(x-\overline{x})$	$(x-\overline{x})^2$
17	-8	-0.13	0.02
18	-11	-3.13	9.79
19	-6	1.87	3.50
20	-4	3.87	14.98
21	-3	4.87	23.73
22	-8	-0.13	0.02
23	-8	-0.13	0.02
24	-8	-0.13	0.02
25	-1	6.87	47.21
26	-2	5.87	34.47
27	-3	4.87	23.73
28	-6	1.87	3.50
29	-3	4.87	23.73
30	-7	0.87	0.76
31	-10	-2.13	4.53
	$\sum$ = -244		$\sum$ = 585.48

$\overline{x} = -244/31 = -7.87 \approx -8$ The average lowest temperature is -8 ° in January, 2014.

$s^2 = 585.48/(31-1) = 19.52$

$s = \sqrt{19.52} \approx 4.4$ The standard deviation for lowest temperature is 4.4 ° in January, 2014.

Lowest Temperature Graph, January, 2014