Short Course on Wolfram Mathematica

by A.B. Golovin, Ph.D. in Physics and Mathematics,

agolovin@ccny.cuny.edu,

https://www.ccny.cuny.edu/profiles/andrii-golovin

Statistics is the art of learning from data

- It is about mode to collect data;
- It is about subsequent description and analysis;
- It is about final conclusions from data acquisition.

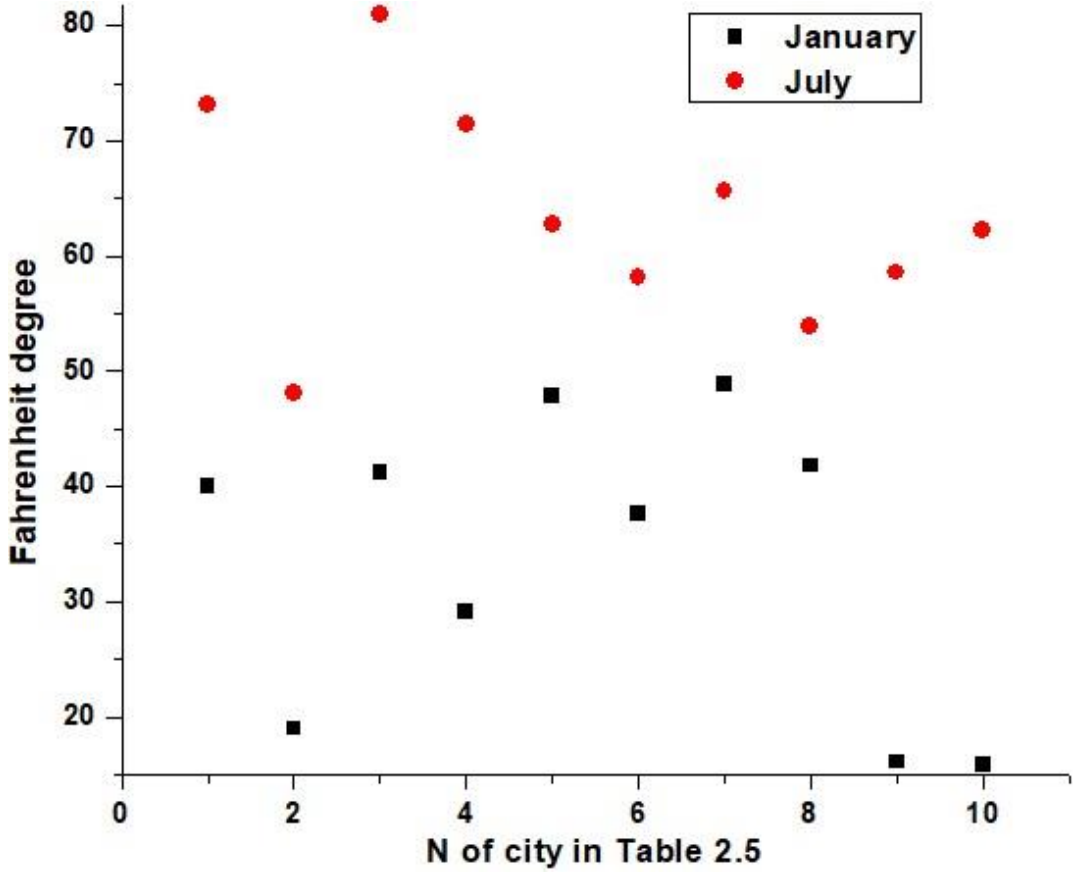
- The part of statistics dedicated to the description and summarization of data is called *descriptive statistics.*

Problem. Using data on the first 10 cities listed in Table, draw a scatter diagram and find the sample correlation coefficient between the January and July temperatures.

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5

Problem 33. Using data on the first 10 cities listed in Table 2.5, draw a scatter diagram and find the sample correlation coefficient between the January and July temperatures.

	A(X)	B(Y)
	January	July
	F	
1	40	73.2
2	19	48.1
3	41.2	81
4	29.1	71.5
5	47.8	62.8
6	37.7	58.1
7	48.9	65.7
8	41.8	53.9
9	16.1	58.6
10	15.8	62.2



Let's introduce some statistics that are used for describing the center of a set of data values $x_i = [x_1, x_2, \ldots, x_n]$:

Sample Mean

- Definition: The *sample mean* is the arithmetic average of the values of the variable x_i in a sample $[x_1, x_2, \ldots, x_n]$:

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

Sample Median

- Definition: The *sample median* is the middle value, when the data set is arranged in increasing order.

Thus, one should order the values of a data set of size *n* from smallest to largest.

- If *n* is odd, the *sample median* is the value in position $\frac{n+1}{2}$;
- if *n* is even, it is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

EXAMPLE

The winning scores in the U.S. Masters golf tournament in the years from 1999 to 2008 were as follows:

$$x_i = [\ 280, 278, 272, 276, 281, 279, 276, 281, 289, 280\ ]$$

Sample Mean is 279.2

```
In[ ]:= (280 + 278 + 272 + 276 + 281 + 279 + 276 + 281 + 289 + 280)/10

Out[ ]= 279.2
```

Sample Median is 279.5

n=10 **is even, it is the average of the values in positions** ***n*****/2=5 and** ***n*****/2+1=6 :**

[272, 276, 276, 278, 279, 280, 280, 281, 281, 289]

$$\frac{279+280}{2}=279.5$$

Sample Mean and Sample Median

- **The *sample mean* and *sample median* are both useful statistics for describing the central tendency of a data set.**
- **The *sample mean* makes use of all the data values and is affected by extreme values that are much larger or smaller than the others.**
- **The *sample median* makes use of only one or two of the middle values and is thus not affected by extreme values.**
- **Which of them is more useful depends on what one is trying to learn from the data.**

One also might be interested in the parameters that describe the spread or variability of the data values.

Sample Variance

- **Definition: The *sample variance* is the measure of average value of squares of the distances between the data values and the sample mean:**

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

Example 2.3f. Find the sample variances of two data sets:
$x_i = [\,3, 4, 6, 7, 10\,]$ **and** $y_i = [-20, 5, 15, 24\,]$

Solution: $\overline{x} = \frac{3+4+6+7+10}{5} = 6,\ \overline{y} = \frac{-20+5+15+24}{4} = 6;$

$$s_x^2 = \frac{(3-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(10-6)^2}{5-1} = 7.5,$$

$$s_y^2 = \frac{(-20-6)^2+(5-6)^2+(15-6)^2+(24-6)^2}{4-1} = 360.67.$$

- **Conclusion: both data sets have the same *sample mean*, but there is a much greater variability in the values of the Y set than in the X set.**

Sample Standard Deviation

- **Definition: The positive square root of the sample variance is called the *sample standard deviation:***

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

The *sample standard deviation* is measured in the same units as the data.

Paired Data Sets

Let's consider data sets that consist of pairs of values that have some relationship to each other. For example, the daily midday temperature [°C] and the number of defective parts produced during that day, a company recorded the data presented in Table 2.8. For this data set, x_i represents the temperature and y_i the number of defective parts produced on day i.

(x_i, y_i)

	A(X)	B(Y)
	T, C	N of Defect
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

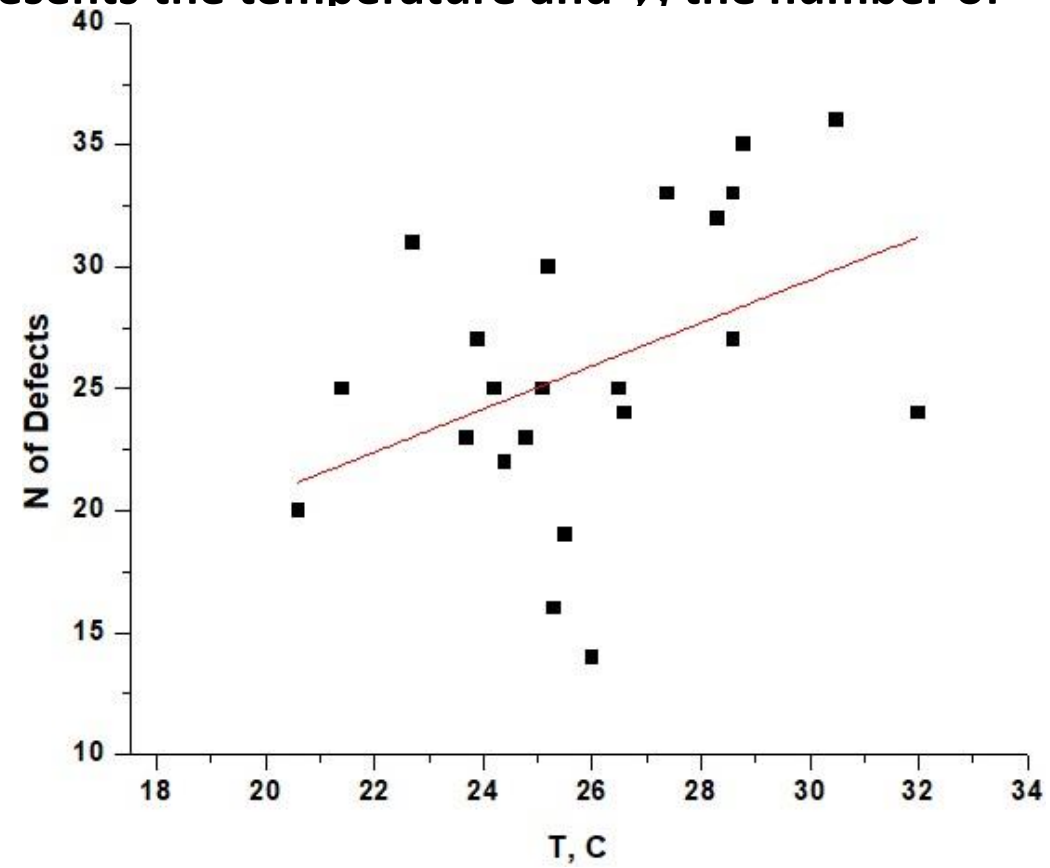


Figure 2.13 indicates that there is some connection between high temperatures and large numbers of defective items.

- If $x_i - \overline{x}$ and $y_i - \overline{y}$ both have the same sign (either positive or negative), then their product $(x_i - \overline{x})(y_i - \overline{y})$ will be positive. Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ will be a large positive number.
- Also, it similarly follows that when large values of x_i tend to be paired with small values of y_i, then the signs of $x_i - \overline{x}$ and $y_i - \overline{y}$ will be opposite and so $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ will be a large negative number.
- To determine what it means for $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ to be "large," we standardize this sum first by dividing by *n−1* and then by dividing by the product of the two sample standard deviations.

- **Definition:** The ***sample correlation coefficient*** r, of the data pairs (x_i, y_i), where $i = 1, \ldots, n$, is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \left| s = \sqrt{\sum_{i=1}^{n}\frac{(x_i - \overline{x})^2}{n-1}} \right| = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- When $r > 0$ the sample data pairs are ***positively correlated***, and when $r < 0$ they are ***negatively correlated***.

Paired Data Sets

Let's consider data sets that consist of pairs of values that have some relationship to each other. For example, the daily midday temperature [°C] and the number of defective parts produced during that day, a company recorded the data presented in Table 2.8. For this data set, x_i represents the temperature and y_i the number of defective parts produced on day i.

Table 2.8.

	A(X)	B(Y)
	T, C	N of Defect
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

EXAMPLE 2.6a. **Find the sample *correlation coefficient* for the data presented in Table 2.8.**

Solution:

```
In[1]:= data1 = {24.2, 22.7, 30.5, 28.6, 25.5, 32.0, 28.6, 26.5, 25.3, 26.0, 24.4,
          24.8, 20.6, 25.1, 21.4, 23.7, 23.9, 25.2, 27.4, 28.3, 28.8, 26.6};
        data2 = {25, 31, 36, 33, 19, 24, 27, 25, 16, 14, 22, 23, 20, 25, 25, 23, 27,
          30, 33, 32, 35, 24};

In[3]:= a1 = Mean[data1] // N;
        a2 = Mean[data2] // N;

In[5]:= sum1 = Sum[Take[((data1 - a1) × (data2 - a2)), {n}], {n, 1, 22, 1}];
        sum2 = Sum[Take[(data1 - a1)^2, {n}], {n, 1, 22, 1}];
        sum3 = Sum[Take[(data2 - a2)^2, {n}], {n, 1, 22, 1}];
        sum1
        ---------
        √(sum2 × sum3)

Out[8]= {0.418944}
```

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \approx 0.4$$

Problem. Using data on the first 10 cities listed in Table, draw a scatter diagram and find the sample correlation coefficient between the January and July temperatures.

	A(X)	B(Y)
	January	July
	F	
1	40	73.2
2	19	48.1
3	41.2	81
4	29.1	71.5
5	47.8	62.8
6	37.7	58.1
7	48.9	65.7
8	41.8	53.9
9	16.1	58.6
10	15.8	62.2

Answer: 0.37

```
In[1]:= SetDirectory["F:/11 Classes CUNY/EE311/02_Online-Lectures/Chapter_02/Lib"];

In[2]:= data1 = Flatten[Transpose[Import["Pr_33_Jan.csv"]]];

In[3]:= data2 = Flatten[Transpose[Import["Pr_33_Jul.csv"]]];

In[4]:= Correlation[data1, data2]   (* sample correlation coefficient *)

Out[4]= 0.370405
```

Problem. Using data on the first 10 cities listed in Table 2.5, draw a scatter diagram and find the sample correlation coefficient between the January and July temperatures.

	A(X)	B(Y)
	January	July
	F	
1	40	73.2
2	19	48.1
3	41.2	81
4	29.1	71.5
5	47.8	62.8
6	37.7	58.1
7	48.9	65.7
8	41.8	53.9
9	16.1	58.6
10	15.8	62.2