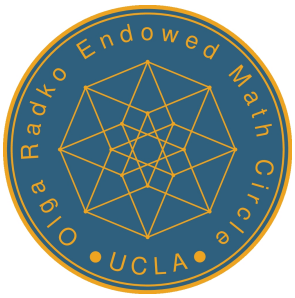


Oleg Gleizer
prof1140g@math.ucla.edu

Seth Tang
sethtang001@g.ucla.edu

Sample Standard Deviation

According to the Merriam-Webster dictionary, *statistics* is a branch of math dealing with collection, analysis, interpretation, and presentation of masses of numerical data.

The word *status* means a *state* in Latin.

Question 1 *Guess how the Latin word status evolved into the English word statistics.*

Key words in statistics:

Inference: a conclusion reached on the basis of evidence and reasoning.

Population: an entire set of similar objects or events about which one wants to draw an inference. For example, if one is studying the average height of the human population at a particular time in history, the population is the entire human population.

Sample: a subset of the population from which data is collected and analyzed. For example, one may use information about 30 students found within our class to try to infer the behavior of human heights.

The purpose of statistics is to use the information contained in a sample to make inferences about the population from which the sample is taken.

Problem 1 *A statistician working for a company manufacturing motor vehicles is researching the following question. What percentage of the US population currently looking to purchase a new vehicle in the price range $20,000 - $40,000 would buy a particular brand made by the company?*

- *What is the population of interest?*
- *What is the objective of the study?*
- *How would you collect a sample?*

In mathematics, the letter N is often used as a symbol of a large positive integer.

Let the real numbers $x_1, x_2, \ldots, x_N$ be the values of some quantity for a population. The following number is called the *population mean:*

$$\mu = \frac{1}{N} \sum_{k=1}^{N} x_k. \tag{1}$$

Let $\{y_1, y_2, \ldots y_n\} \subset \{x_1, x_2, \ldots, x_N\}$. The following number is called the *sample mean:*

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k. \tag{2}$$

Note that while μ is a fixed number, $\bar{y}$ can change from one sample to another. In addition, x_k may be repeated.

Problem 2 *Make an Excel or Google sheet file with the names and heights of all the students in your class. (For the reminder of this packet, everything that applies to Excel also applies to Google sheets.)*

- *Find an Excel function that computes mean values. Use the function to find the average height of a student in the class.*
- *Similarly, find the average height of the students at your table. If you are not seated around tables, find the average height of the six students nearest to you.*
- *How well do sample means represent the population mean in the study?*

The following number is called the *population standard deviation:*

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2}. \tag{3}$$

It measures the dispersion of data points around the mean within the entire population.

We know that the numbers μ and σ exist, but they are most often either impossible or not feasible to find out. For example, a company making light bulbs claims that each bulb is expected to work q hours on average. To find the population mean, the compony should burn out every light bulb they have produced, to measure the life span of each, to add up the numbers, and to divide by the number of the bulbs made. Once μ is known, they can proceed to figure out σ using formula 3. This approach is not quite practical, is it?

The following number is called the *sample standard deviation:*

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}. \tag{4}$$

It measures the dispersion of data points around the sample mean. The sample standard deviation found using formula (4) is also occasionally called the *unbiased* sample standard deviation.[1] This means that (4) is an estimate of the population standard deviation σ free from a systemic error, a.k.a. bias.

For the first look, it seems that the formula for the sample standard deviation should be similar to the formula for the population standard deviation:

$$s_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2}. \tag{5}$$

The sample standard deviation s_n found using formula (5) is called *biased*, meaning that it has a systemic error and thus is a less accurate estimate of

[1] Assuming the sample it unbiased. To get an unbiased sample, every person in the population must have an equal chance of being selected. For example, the problem 2 sample is highly biased.

σ than s. It is the goal of this packet to explain why s is a more accurate estimate of σ than s_n.

Problem 3 *Find Excel functions that compute s and s_n. Do the following for the set of students' heights from problem 2.*

- *Find the population standard deviation σ.*

- *Considering the 6 students near you as a sample, find the unbiased sample standard deviation s.*

- *Considering the 6 students near you as a sample, find the biased sample standard deviation s_n.*

Problem 4 *Show that*

$$\sum_{k=1}^{n}(y_k - \bar{y})^2 = \sum_{k=1}^{n} y_k^2 - n\bar{y}^2. \tag{6}$$

Hint: $n \times \dfrac{1}{n} = 1$. Also the right hand side is composed of 2 parts.

A *random variable* Y is a function from the set of all the possible outcomes of an experiment to the set of real numbers.

$$Y : \Omega \to \mathbb{R}$$

Problem 5 *Let Y be the height of a randomly chosen student in the class. What is the domain Ω in this case?*

Let $x_1, x_2, \ldots, x_N$ be the values the random variable Y takes on all the elements of the population set. The *expectation* $E(Y)$ of the random variable Y equals to the population mean:

$$E(Y) = \frac{1}{N}\sum_{k=1}^{N} x_k = \mu. \tag{7}$$

The *variance* of a random variable Y

$$V(Y) = E((Y-\mu)^2) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 = \sigma^2 \tag{8}$$

equals to the population standard deviation.

A function that takes functions as inputs is often called a *functional*. Expectation and variance of a random variable are functionals. A functional F is called *linear*, if for any two input functions A and B and for any constant c, the following holds:

- $F(A+B) = F(A) + F(B)$
- $F(cA) = cF(A)$

Problem 6 *Show that expectation E is a linear functional.*

Problem 7 *Let the random variable Y be a constant, $Y \equiv c$. Find $E(Y)$ and $V(y)$.*

Problem 8 *$E(Y^2)$ is defined as follows:*

$$E(Y^2) = \frac{1}{N}\sum_{k=1}^{N} x_k^2. \tag{9}$$

Show that $V(Y) = E(Y^2) - \mu^2$.

Note that combining formula 8 and problem 8 together brings about the following useful formula:

$$E(Y^2) = \mu^2 + \sigma^2. \tag{10}$$

Sample mean and sample standard deviation are known as estimators of the population parameter such as μ and σ. To say that an estimator is unbiased, is equivalent to say:

$$E(estimator) = parameter. \tag{11}$$

For example, saying that the sample standard deviation s is unbiased is equivalent to

$$E(s) = \sigma \quad \text{and} \quad E(\bar{y}) = \mu \tag{12}$$

A distribution is called *uniform*, if every outcome of a random variable is equally likely. The uniform distribution is a fun example to play with an unbiased estimator. Let (0,b) be the range of outcome for distribution Y.

One estimator of the parameter b for the uniform distribution y is given by :

$$\widetilde{b} = 2\bar{y} \tag{13}$$

It is also known the following relations (note the lack of squiggle on b):

$$\bar{y} = \frac{b}{2} \tag{14}$$

Problem 9 *Calculate $E(\widetilde{b})$ and determine, if this estimator is biased or not. Hint: the end result should be $E(\widetilde{b})$ is expressed in terms of the real parameter b.*

Let Y_1 be the height of the first randomly chosen student in the class, Y_2 be that of the second, etc., and let $n < N$ be the sample size where N is the total number of students in the class. Note that $E(Y_1) = E(Y_2) = \ldots = E(Y_n) = \mu$ and $V(Y_1) = V(Y_2) = \ldots = V(Y_n) = \sigma^2$. Further note that the random variables Y_i and Y_j, $i \neq j$, are *independent*: measuring the value of one does not affect possible values of the other. For example, knowing the height of one student in the class does not in any way affect possible heights of other students. Hence, for independent variables Y_i and Y_j, $i \neq j$, $E(Y_iY_j) = E(Y_i)E(Y_j)$. Generalizing from here, consider n identical and mutually independent random variables $Y_1, Y_2, \ldots, Y_n$.

Problem 10 *Show that for $i \neq j$, $E((Y_i - \mu)(Y_j - \mu)) = 0$.*

Similar to 2, let us define

$$\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k. \tag{15}$$

Note that $\bar{Y}$ is a random variable. Its value can change from one sample to another.

Problem 11 *What is the domain Ω of the function $\bar{Y}$?*

Problem 12 *Show that $E(\bar{Y}) = \mu$. Is $\bar{Y}$ an unbiased estimator of μ?*

Problem 13 *Show that*

$$\left(\sum_{k=1}^{n} p_k\right)^2 = \sum_{k=1}^{n} p_k^2 + 2\sum_{1\le i<j\le n} p_i p_j.$$

for:

- $n = 2$
- $n = 3$
- *any positive integer* $n > 1$.

Problem 14 *Use independence of the random variables* Y_i *and* Y_j, $1 \le i < j \le n$, *to show that*

$$V(\bar{Y}) = \frac{\sigma^2}{n} \tag{16}$$

Hint: $Var(nX) = n^2 Var(X)$.

Since $\bar{Y}$ is a random variable just like Y, it follows from problem 8 that $V(\bar{Y}) = E(\bar{Y}^2) - \mu^2$. The latter implies that

$$E(\bar{Y}^2) = \mu^2 + \frac{\sigma^2}{n}. \tag{17}$$

Similar to formula 6,

$$\sum_{k=1}^{n}(Y_k - \bar{Y})^2 = \sum_{k=1}^{n} Y_k^2 - n\bar{Y}^2. \tag{18}$$

Problem 15 *Show that*

$$E\left(\sum_{k=1}^{n}(Y_k - \bar{Y})^2\right) = (n-1)\sigma^2.$$

Hint: formula 10 helps.

Problem 15 shows that

$$E\left(\frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \bar{Y})^2\right) = \sigma^2,$$

i.e. s is an unbiased estimate of σ.

Now that we have learned about Sample SD and the Excel functions, let us combine those information and put them into action, and gain some additional insight into variances. Open the Excel file from earlier, and execute the following.

Problem 16 • ***16-1:*** *Set up data according to figure 16-1. Column I will be your data, column J will be the difference between Y_n and the sample mean, and column K will be the squared difference based on column J. Finally, also include the mean and variance at the bottom.*

• ***16-2:*** *Now copy and paste graph 16-1 five times. Alter the new 5 graphs by changing the mean to some number not equal to the actual mean (make sure some are less and some are greater than mean), and notice the change in the difference row and squared difference row. Be sure to adjust variance equation accordingly* ***(hint: use the literal equation of variance)****.*

• ***16-3:*** *Organize your results as such, with mean in ranked order matching variance. We will now graph it.*

• ***16-4:*** *Now highlight the data, and click the insert bottom on the top left corner, click charts, and voila! You have created a graph in Excel. Do you notice anything interesting about the graph?*

I	J	K
1	-3.769230769	14.20710059
3	-1.769230769	3.130177515
4	-0.7692307692	0.5917159763
5	0.2307692308	0.05325443787
6	1.230769231	1.514792899
7	2.230769231	4.976331361
8	3.230769231	10.43786982
9	4.230769231	17.89940828
0	-4.769230769	22.74556213
4	-0.7692307692	0.5917159763
5	0.2307692308	0.05325443787
4	-0.7692307692	0.5917159763
6	1.230769231	1.514792899
Mean		Variance
4.769230769		6.023668639

(a) 16-1

1	-2	4
3	0	0
4	1	1
5	2	4
6	3	9
7	4	16
8	5	25
9	6	36
0	-3	9
4	1	1
5	2	4
4	1	1
6	3	9
Mean		Variance
3		9.153846154

(b) 16-2

Mean (Ordered)	Variance
3	9.153846154
4	6.615384615
4.769230769	6.023668639
5	6.076923077
6	7.538461538

(c) 16-3

(d) 16-4

Figure 1: Relevant graphs to Problem 16

Additional information: As we observed in Problem 16, the population mean is the value that minimizes the variance of the entire population. By the same logic, the sample mean serves as the value that minimizes the sample variance within a given sample. This means that if we were to compute the sample variance using the true population mean instead of the sample mean, the result would always be greater than or equal to the actual sample variance. Or we can write it as thus:

$$\sum_{k=1}^{N}(x_k - \mu)^2 \geq \sum_{k=1}^{n}(x_k - \bar{x})^2. \tag{19}$$

From this, it is obvious that if we multiply both sides by factor of $\frac{1}{n}$, this relation will be unchanged. Thus to balance out, the right hand side must be multiplied by a factor greater than the right hand side; for example $\frac{1}{n} < \frac{1}{n-1}$.

While we've proven earlier algebraically that $\frac{1}{n-1}$ is correct, is there any other way to solve this? The answer is yes, and it will be introduced below with some more advanced Excel tool!

A **Bariance** is an alternative way to calculate the standard deviation. Imagine instead of summing only the squared difference between the mean and all other points (i.e., there are n differences), we summed the pairwise difference between every single point (i.e., for n points, there will be a total of n^2 differences). It can be proven that $Var(x) = \frac{Bar(x)}{2}$; we will accept as fact for now. Below is a graphical representation:

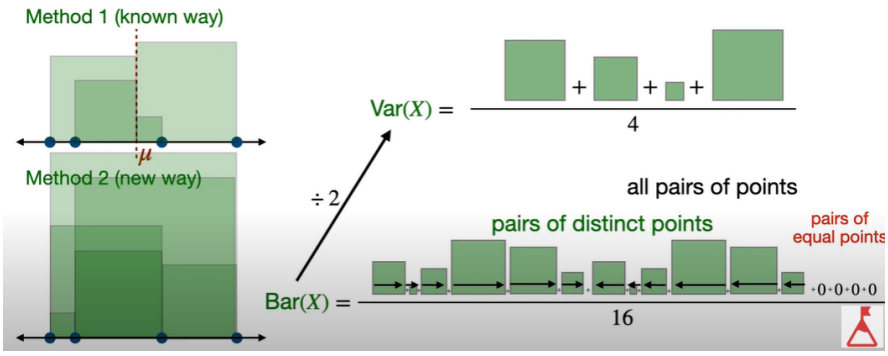


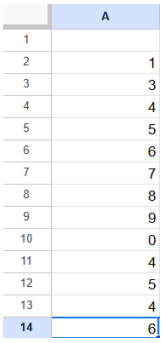
Figure 2: Variance vs Bariance(credit to Youtuber Serrano.Academy)

Now that we have learned about Bariance, let us use Excel to automate the calculation process of pairwise differences!

Problem 17 • ***17-1:*** *Set up data according to figure 17-1. Grid A2 will be your first data, and you will enter your data in a row described in figure 17-1.*

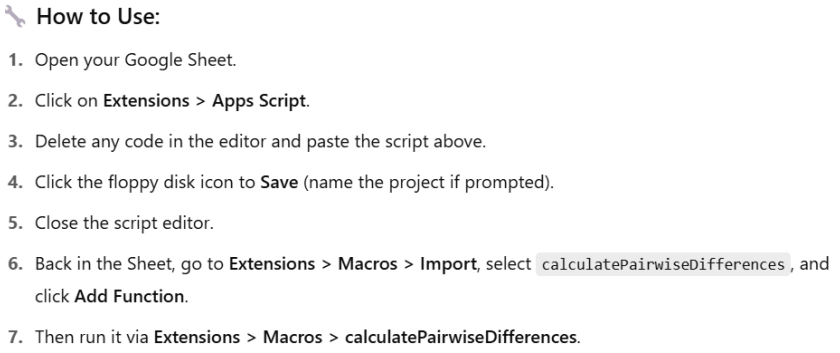
• ***17-2:*** *Go to ChatGPT and enter the following question: "write me a google Excel script to perform calculation of pair wise differences". Rumor has it many programmer are relying on ChatGPT for coding: see for yourself!*

• ***17-3:*** *Follow all the instructions provided, and you should obtain something thing similar to this! (Be sure to modify so it shows the pairwise difference* ***squared****)*



	A
1	
2	1
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	0
11	4
12	5
13	4
14	6

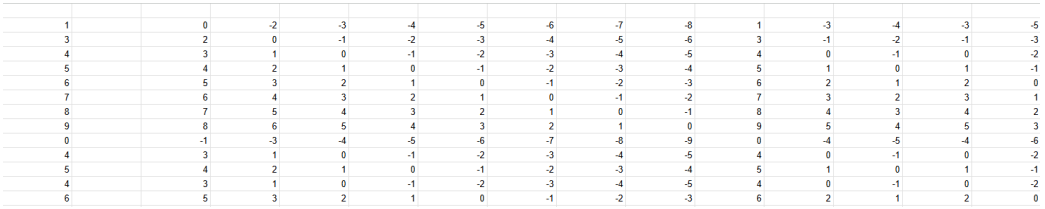
(a) 17-1



How to Use:

1. Open your Google Sheet.
2. Click on **Extensions** > **Apps Script**.
3. Delete any code in the editor and paste the script above.
4. Click the floppy disk icon to **Save** (name the project if prompted).
5. Close the script editor.
6. Back in the Sheet, go to **Extensions** > **Macros** > **Import**, select `calculatePairwiseDifferences`, and click **Add Function**.
7. Then run it via **Extensions** > **Macros** > **calculatePairwiseDifferences**.

(b) 17-2



1	0	-2	-3	-4	-5	-6	-7	-8	1	-3	-4	-3	-5
3	2	0	-1	-2	-3	-4	-5	-6	3	-1	-2	-1	-3
4	3	1	0	-1	-2	-3	-4	-5	4	0	-1	0	-2
5	4	2	1	0	-1	-2	-3	-4	5	1	0	1	-1
6	5	3	2	1	0	-1	-2	-3	6	2	1	2	0
7	6	4	3	2	1	0	-1	-2	7	3	2	3	1
8	7	5	4	3	2	1	0	-1	8	4	3	4	2
9	8	6	5	4	3	2	1	0	9	5	4	5	3
0	-1	-3	-4	-5	-6	-7	-8	-9	0	-4	-5	-4	-6
4	3	1	0	-1	-2	-3	-4	-5	4	0	-1	0	-2
5	4	2	1	0	-1	-2	-3	-4	5	1	0	1	-1
4	3	1	0	-1	-2	-3	-4	-5	4	0	-1	0	-2
6	5	3	2	1	0	-1	-2	-3	6	2	1	2	0

(c) 17-3

Figure 3: Relevant graphs to Problem 17

Although it is not immediately apparent how helps us pinpoint $\frac{1}{n-1}$, a smaller pairwise difference chart may be more illustrative. Take the following sample:

	Y	Y1	Y2	Y3	Y4	Y5
X		1	3	4	5	6
X1	1	0	-2	-3	-4	-5
X2	3	2	0	-1	-2	-3
X3	4	3	1	0	-1	-2
X4	5	4	2	1	0	-1
X5	6	5	3	2	1	0

Figure 4: Sample pairwise difference chart

If written as a double summation, we can calculate the bariance as thus:

$$Bar(x) = \frac{1}{5^2} \sum_{j=1}^{5} \sum_{k=1}^{5} (x_j - y_k)^2. \tag{20}$$

However, notice a curious pattern:

	Y	Y1	Y2	Y3	Y4	Y5
X		1	3	4	5	6
X1	1	0	-2	-3	-4	-5
X2	3	2	0	-1	-2	-3
X3	4	3	1	0	-1	-2
X4	5	4	2	1	0	-1
X5	6	5	3	2	1	0

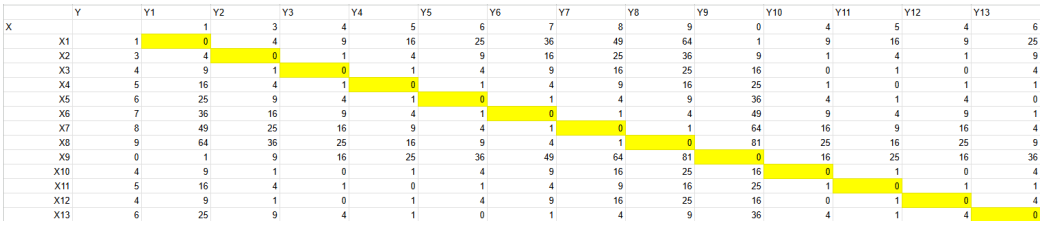
Figure 5: Sample pairwise difference chart

Specifically, what if we alter the equation further, such as:

$$Bar(x) = \frac{1}{5(5-1)} \sum_{\substack{j=1 \\ j \neq k}}^{5} \sum_{k=1}^{5} (x_j - y_k)^2. \tag{21}$$

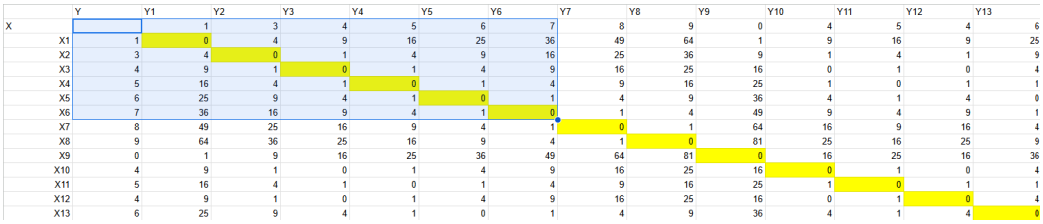
Problem 18 *What is the difference between (20) and (21)? Are they the same or are they different?*

The million dollar question is: should we modify the Bar(x) as such? Here is an example that will better illustrate the question. Consider the following



	Y	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13
X		1	3	4	5	6	7	8	9	0	4	5	4	6
X1	1	0	4	9	16	25	36	49	64	1	9	16	9	25
X2	3	4	0	1	4	9	16	25	36	9	1	4	1	9
X3	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X4	5	16	4	1	0	1	4	9	16	25	1	0	1	1
X5	6	25	9	4	1	0	1	4	9	36	4	1	4	0
X6	7	36	16	9	4	1	0	1	4	49	9	4	9	1
X7	8	49	25	16	9	4	1	0	1	64	16	9	16	4
X8	9	64	36	25	16	9	4	1	0	81	25	16	25	9
X9	0	1	9	16	25	36	49	64	81	0	16	25	16	36
X10	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X11	5	16	4	1	0	1	4	9	16	25	1	0	1	1
X12	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X13	6	25	9	4	1	0	1	4	9	36	4	1	4	0

(a) 18-1



	Y	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13
X		1	3	4	5	6	7	8	9	0	4	5	4	6
X1	1	0	4	9	16	25	36	49	64	1	9	16	9	25
X2	3	4	0	1	4	9	16	25	36	9	1	4	1	9
X3	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X4	5	16	4	1	0	1	4	9	16	25	1	0	1	1
X5	6	25	9	4	1	0	1	4	9	36	4	1	4	0
X6	7	36	16	9	4	1	0	1	4	49	9	4	9	1
X7	8	49	25	16	9	4	1	0	1	64	16	9	16	4
X8	9	64	36	25	16	9	4	1	0	81	25	16	25	9
X9	0	1	9	16	25	36	49	64	81	0	16	25	16	36
X10	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X11	5	16	4	1	0	1	4	9	16	25	1	0	1	1
X12	4	9	1	0	1	4	9	16	25	16	0	1	0	4
X13	6	25	9	4	1	0	1	4	9	36	4	1	4	0

(b) 18-2

Let 18-1 represent the entire population, and let the shaded area in 18-2 represent a sample of the population. Notice that in 18-1, exactly $\frac{1}{13}$ of squares are zero; in comparison, in 18-2, a proportion of $\frac{1}{6}$ of squares are zero. This means, in the calculation of estimated bariance, that the proportion of zeros within the numerator part is affected by the size of the sample.

In other words, as the sample size decreases, the estimated bariance will also decrease due to the larger proportions of zeros in the numerator.The key to counteracting it is to adjust the denominator accordingly, specifically to not allow the empty zeros to reduce the bariance.

Thus, we modify the Bar estimator by multiply it by a factor of $\frac{n}{n-1}$, thus reducing the denominator from n^2 to $n(n-1)$, matching what have done in the previous page. Notice that while n is small, this modification will in enlarge bariance estimator, and as n approaches the real N (in some cases, $N = \infty$), the -1 part becomes negligible as the proportion taken up be zero becomes increasingly small as well.

To top it all off, let us present the now newly modified bariance equations, generalized for sample size n, where population = N.

For population of size N, we calculate the bariance as thus.

$$Bar(x) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} (x_j - y_k)^2. \tag{22}$$

Deriving variance is simply a division of 2, so we will move on to sample of size n.

$$Bar(x) = \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq k}}^{n} \sum_{k=1}^{n} (x_j - y_k)^2. \tag{23}$$

Whether it is from algebraic calculations or from the observation that we wish not to overcount the zeros in bariance calculation, we can see that the smaller size of sample relative to population inevitably creates a smaller variance estimator then the actual population variance, and that an adjustment is needed in order for the estimator to become unbiased, especially when n is so much smaller then N.

Problem 19 *Consider the equation of Bar(x), and the accompanying pairwise distance graph. Observing figure 7, we can see that our distance graph can be transformed into a geometric graph consisting of nodes(representing each x_is and pairwise distances.)*

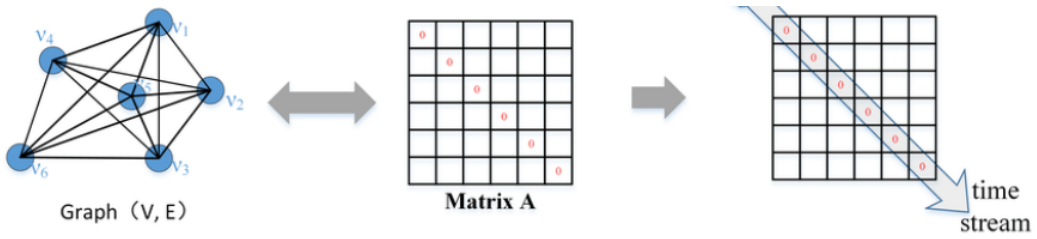


Figure 7: The idea

Using our pairwise distance graph, try to create a graph consisting of nodes and edges.