On the Usefulness of Vector Representations in Statistics Training

Atsushi Terao
Department of Systems Science
Tokyo Institute of Technology
Ohokayama Meguro-ku Tokyo 152 Japan
terao@tp.titech.ac.jp

The students who are majoring in psychology or other relevant disciplines have to study statistics. It seems to be a difficult task to understand statistical concepts for these students. Generally speaking, they are not very good at mathematics. All of the statistical theories, however, are constructed on the bases of mathematics. The teacher often has to try to help them construct or understand statistical concepts without high level mathematics. Thus, it is quite meaningful to examine how to teach statistical concepts for these students in statistics training.

Are there useful materials that help students who are not good at mathematics understand statistical concepts deeply ? It seems to be useful to use vector representations in the training of these students. The vector representation is a kind of diagram. Diagrams often help students memorize, understand, and solve problems [Larkin & Simon 1987]. For example, suppose that a student has to understand the proof which indicates that the correlation coefficient ranges from minus 1 to plus 1. If algebraic method is used in order to prove it, it is probably difficult for him (or her) to understand the proof. If an appropriate vector representation is used in the proof, the student may be able to understand it, because the correlation coefficient is consistent with the value of cosine for two vectors which represent variables and it is clear that the value of cosine ranges from minus 1 to plus 1. The only thing the student has to do is to understand that the variables are represented by vectors. Thus, it may be easy for him or her.

There are some textbooks which explain statistics by using vector representations [e.g., Takeuchi,Yanai, & Mukherjee 1982; Eaton 1984]. It is expected that vector representations support or improve students' understanding of statistical concepts. However, no research has been done to prove its validity.

Moreover, vector representations have been only used by teachers in order to explain statistical concepts for students. However the representations can be used by students who have to solve some statistical problems. If students are taught statistical concepts by using vector representations, are they able to solve some statistical problems by using the representations ? I consider the representations as materials not only for teacher's explanation but also for student's problem solving.

When a student learns from a textbook, he or she must have some questions about the explanations in the textbook. He or she may try to find the statements that mention the questions. However, the textbook is often inadequate because it omits statements about the questions. In this case, the student has to generate new pieces of knowledge by "self-explanations" [Chi,Bassok,Lewis,Reimann, & Glaser 1989]. Self-explanations are considered to be inferences that went beyond the text [Chi, de Leeuw, Chiu, & LaVancher 1994]. There are various forms of self-explanations [Ferguson-Hessler & de Jong, 1990]. I will focus on the forms of problem solving. For example, the multiple correlation coefficient is defined as an index that is used for multiple regression. Suppose that a student has a question of why this index is not used in the case of simple regression. Although this question seems to be very natural, no textbook explains it. Thus the student has to solve the question by himself (or herself). In other words, he or she has to do problem solving because he or she will try to define the multiple correlation coefficient in the case of multiple regression. (This task is used in experiment 1 and 2.)

The purpose of this study is to examine the usefulness of vector representations in statistics training. I will explore this from the two viewpoints, that is, understanding the textbook and problem solving, for the reasons mentioned above. Two simple experiments have been carried out.

To give tasks that can be solved by using vector representations is also the purpose of this study. Although this purpose is a supplementary one, it must be useful for teachers who intend to teach statistics by using vector representations because no textbook shows what tasks can be resolved by using the representations and how these tasks are resolved.

Experiment 1

In the experiment 1, the usefulness of vector representations was examined from the viewpoints of problem solving. Two conditions were compared. In experimental condition, they were required to do it by using vector representations. In control condition, subjects were required to solve a task by using numerical formulas.

Method

Subjects. Subjects were four graduate students who were studying statistics or other relevant disciplines. This study originally aimed to examine statistics training for students who are not good at mathematics. However the students in experiment 1 were considered to be good at mathematics. I have started this study for these students because of availability. In experiment 2, the subjects who are not good at mathematics will be used.

Design. As mentioned above, two conditions were compared. We will call them *vector representation* condition and *numerical formula* condition respectively. These conditions were within-subject.

Material. The subjects were required to solve the following problem: Consider the multiple correlation coefficient "R" in the case of simple regression. Please explain the relationship between this "R" and "r" that is the simple correlation coefficient for two variables.

As mentioned above, this question seems to be natural. No textbook, however, contains statements that answer the question. The answer of this task is "$R = |r|$". This conclusion can be drawn in both conditions. In the numerical formula condition, the solution is as follows: the "R" is defined as the formula $R = \frac{\frac{1}{n}\sum(y_i-\overline{y})(Y_i-\overline{Y})}{\sqrt{\frac{1}{n}\sum(y_i-\overline{y})^2}\sqrt{\frac{1}{n}\sum(Y_i-\overline{Y})^2}}$. The capital letter "$Y$" represents expected value, and the small letter "y" represents actual value or data. This formula means that "R" is defined as the correlation coefficient between two variables that are represented by "Y" and "y". In the case of simple regression, $Y_i = a_0 + a_1 x_i$, $\bar{Y} = a_0 + a_1 \bar{x}$. ($a_0$ represents the intercept, and a_1 represents the inclination.) If these relations are substituted for the R's formula and the formula is put in order, then the conclusion $R = |r|$ is obtained. In the vector representation condition, the "R" is defined as the value of cosine for two vectors "Y" and "y". Therefore the conclusion is obtained from the figures that are presented in Figure 1.

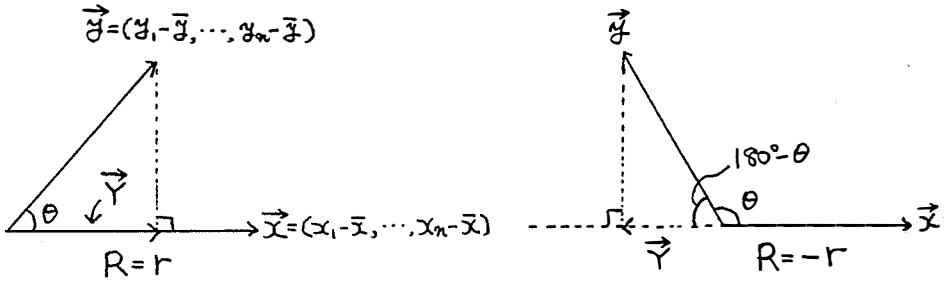

Figure 1. The vector representations that indicate $R = |r|$.

Procedure. At first, subjects were required to solve the problem in the numerical formula condition. They were allowed to refer a textbook. [Tanaka & Wakimoto 1983] that explains multivariate statistical analysis by using numerical formulas and contains all of the formulas that are needed to solve the problem. The time for solving the problem was not limited. After he or she gave up trying to solve it or reached the conclusion, they were shown a vector representation that explained the multiple and the simple correlation coefficient (see Figure 2), and then were required to solve the task again by using vector representations. All subjects were given a paper-and-pencil version of this problem.

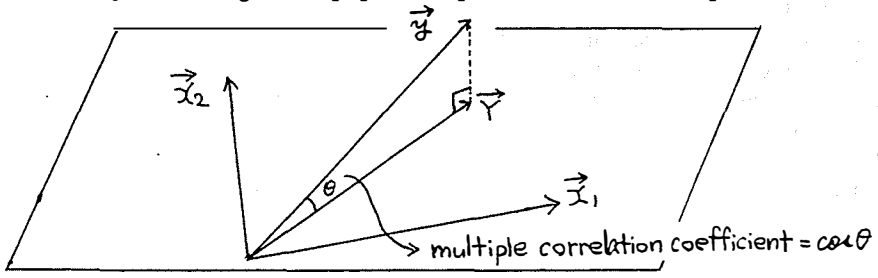

Figure 2. The vector representation that explains the correlation coefficient.

Results and Discussion

Only one subject succeeded in solving the problems in the numerical formula condition. Three of four subjects failed to solve it in this condition. Note that the subjects are graduate students majoring in statistics or other relevant disciplinary. The result suggests that it is difficult for students to solve this problem by using numerical formulas even if they are good at mathematics.

In the vector representation condition, all the subjects succeeded in solving the problems. Note that all of the subjects who gave up solving this task in the numerical formula condition succeeded in solving it in the vector representation condition.

The results indicate that vector representations are useful in the context of problem solving. However this experiment is seriously flawed because all subjects attempted an algebraic solution before trying the vector representation. I will resolve this problem in the second experiment by using between-subject design.

Experiment 2

In this experiment, subjects were changed from graduate students to undergraduate students who were majoring in psychology. They were considered to be not good at mathematics. In experiment 1, the results supported the usefulness of vector representations. However, we have to confirm these results again for these subjects in experiment 2.

In experiment 1, we have examined the usefulness of vector representations in the context of problem solving. In addition to this context, we will explore it in the context of understanding textbooks.

Moreover, we will examine whether vector representations are always useful for students. If these are not always useful, we can find the principles why and when these representations are useful.

Method

Subjects. Subjects were eight undergraduate students who were majoring in psychology.

Design. Before this experiment, subjects received a simple examination (pre-test) that measured basic knowledge of statistics. This examination inquired about the formula of mean, variance, SD, covariance, and correlation coefficient, and was marked on a maximum scale of five points. Based on the results of this examination, subjects were assigned to two conditions, that is, *vector representation* condition (four subjects) and *numerical formula* condition (four subjects), so that there would be no difference between the two conditions with respect to subjects' knowledge about elementary statistics.

In experiment 1, the design is within-subject. Because of this, it is possible the good results in the vector representation condition are caused by the order effect. In order to resolve this problem, the design is between-subject in experiment 2.

Materials. We examined the usefulness of vector representations in the two contexts in experiment 2, that is, understanding the text material and problem solving. Therefore, the subjects were given a text material and some problems.

The text material was made by the author because I was not able to find an appropriate one. Two types of text materials were used. I will briefly explain the contents of the text materials.

In the numerical formula condition, the text explained mean, SD, and variance at first. Then the scatter diagram was presented, and the correlation coefficient was defined in the forms of numerical formula based on the diagram. Then the proof which indicates that the correlation coefficient ranges from minus 1 to plus 1 was presented by using algebraic method (using numerical formula). At last, simple and multiple regression analysis was explained from the viewpoint of the least squares method that was based on the scatter diagram.

In the vector condition, mean, SD, variance were explained as same as numerical formula condition. Then a vector representation that is constructed with two vectors $\vec{X} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})$ and $\vec{Y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})$ was presented instead of scatter diagram. The correlation coefficient was defined as the value of the cosine for this two vectors. The range of the correlation coefficient was explained as the range of the cosine, that is, from minus1 to plus 1. At last simple and multiple regression analysis was explained from the viewpoint of projection on the linear space [Takeuchi, Yanai, & Mukherjee 1982].

In addition to the problem used in experiment 1 (We will call it problem 1.), two problems were used in this experiment (We will call them problem 2 and problem 3, respectively.).

Problem 2 is as follows: In the case of simple regression, if the number of paired data (for example x_1 and y_1) is two, we can describe one variance by using the other variance without any margin of error. Give an explanation for the reason this can be done.

In the numerical formula condition, the answer is as follows: In this case, two dots are plotted on the scatter diagram. The regression straight line is uniquely specified because the straight line that connects the two dots is unique.

In the vector representation condition, the answer is as follows: See the vector representation presented in figure 3. The vector $\vec{X}$ lies at right angles to the vector $(\bar{x}, \bar{x})$, and the vector $\vec{Y}$ lies at right angles to the vector $(\bar{y}, \bar{y})$. (This is found by calculating the inner product.) Thus, the vector $\vec{X}$ is parallel with vector $\vec{Y}$. This means the vector $\vec{Y}$ is described as "$\alpha\vec{X}$".

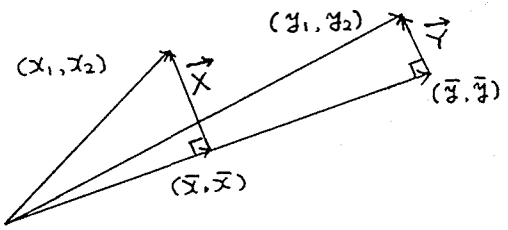

Figure 3. The vector representation used to solve problem 2.

Problem 3 is as follows: Explain the relationship between regression coefficient and the correlation coefficient in the case of simple regression by using only S_{xx} and S_{yy}.("S" represents covariance.) In the numerical formula condition, subjects can solve the problem as follows:

$$\hat{a}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r_{xy} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}$$

In the vector representation condition, subjects can solve the problem as follows: If we see the vector representation presented in figure 4, we can find $r = \cos\theta = \frac{a}{b}$. If the relations $a = \hat{a}_1\sqrt{\sum(x_i - \bar{x})^2}$, $b = \sqrt{\sum(y_i - \bar{y})^2}$ are substituted, we will reach the same conclusion as numerical formula condition.

All of the elementary statistical concepts that were needed to solve the problems were explained in the text materials.

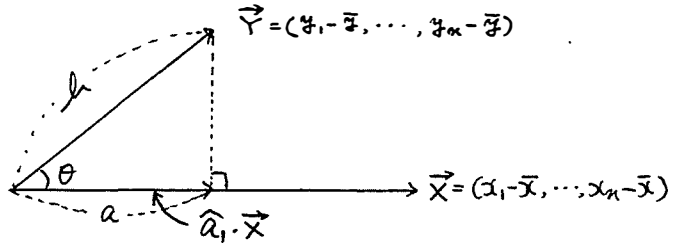

Figure 4. The vector representation used to solve problem 3.

Procedure. There were two sessions in this experiment, that is, understanding the text material and problem solving. In the first session, subjects were required to understand the text material. If they found a description that was difficult to understand, they were also required to underline that point and to note the reason in the margin why it is difficult to understand. The time was not limited in this section. It took about one hour for all subjects to finish this section.

After that, each subject had the descriptions that had been difficult for him or her to understand explained (by the author). Thus, all of the questions were resolved.

In the second session, subjects were required to solve the three problems mentioned above. Subjects were given fifteen minutes for each problem. All subjects were given a paper-and-pencil version to solve these problems. They were allowed to refer to the text material.

The subjects received the correct solution of each problem after they had finished trying to solve all the problems. They were required to evaluate to what degree they could understand each solution and to what degree they could accept each solution on a 5-point scale ranging from "very difficult to understand (or accept)" to "very easy to understand (or accept)" .

Results and Discussion

Understanding the text material.

Table 1 shows the results of experiment 2. In the numerical formula condition, three of four subjects have underlined in the text material. Two subjects (subject G and H, see Table 1) were not able to understand the proof about range of the correlation coefficient, and two subjects (subject F and H)

were not able to understand the formula of the correlation coefficient (Especially, they were not able to understand why covariance is divided by SDs.). We can find these descriptions in the many textbooks. However, the results suggest that it is difficult for students to understand the descriptions.

In contrast to the results in numerical formula condition, only one subject (subject B) had underlined in the text material in vector representation condition. She was not able to understand the concept of "projection" because she had not learned this concept.

Although we must be careful to reach any conclusion because this experiment was done with a small number of subjects (this work is still in progress), these results support the usefulness of the vector representations.

Do the explanations using vector representation always help students understand statistics ? The answer is "No". Let us see the scores of evaluating the degree of understanding and accepting the correct solution of problem 2 (see Table 1). We will realize that the overall scores are less high in the vector representation condition than in the numerical formula condition. This result indicates that it is difficult to understand and accept the explanation about correct solution of problem 2.

Why do vector representations sometimes not help students understand statistics ? I believe that the usefulness of vector representations are caused by easy geometric manipulation (e.g., changing the angle between two vectors). In the case of understanding the proof about the range of the correlation coefficient, students found it fairly easy to manipulate the vector representation geometrically. The only thing they have to do is to understand that the variables are represented by vectors. In contrast to this case, the students are required other tasks in addition to the geometric manipulation in order to understand the correct solution of problem 2. For example, they have to understand to calculate the inner product. They also must understand why vector $\vec{X}$ must be divided into two vectors (x_1, x_2) and $(\bar{x}, \bar{x})$. Based on this discussion , I maintain that the usefulness of vector representations are caused by easy geometric manipulation. I will insist on this again in the next section.

Table 1. The Results of Experiment 2.

				Problem 1			Problem 2			Problem 3		
Conditions	Subjects	Pre-test	Difficulties	S/F	Un	Ac	S/F	Un	Ac	S/F	Un	Ac
Vector	A	3	0	F	5	2	F	2	1	F	4	4
	B	1	1	F	1	1	F	4	2	F	4	2
	C	1	0	S	5	5	F	2	3	F	5	5
	D	1	0	S	4	4	F	4	4	F	4	4
Formula	E	3	0	F	4	3	F	5	3	F	5	4
	F	1	1	F	5	5	F	5	4	S	5	5
	G	1	1	F	4	4	S	5	5	F	4	2
	H	0	2	F	4	2	S	5	5	F	4	4

Notes. Pre-test: the score of pre-test (1–5)
Difficulties: the number of descriptions in the text which were difficult to understand
S/F: whether succeeding in solving or failing to solve each problem
Un: the score of evaluating the degree of understanding the correct solution (1–5)
Ac: the score of evaluating the degree of accepting the correct solution (1–5)

Problem solving.

In the vector representation condition, one subject (subject D, see Table 1) succeeded in solving problem1, and another subject (subject C) reached the conclusion $R = r$ in the case of $r \geq 0$. It would be relatively easy to draw diagrams showed in Figure 1 and to manipulate them geometrically. No subject succeeded in solving the problem in the numerical formula condition. These results, which are similar to the results in experiment 1, suggest that vector representation is useful for this problem.

In contrast to the results for problem 1, no subject was unable to solve problem 2 in vector representation condition. Two of four subjects in numerical formula condition (subject G and H, see Table 1) succeeded in solving the problem. It seems to be difficult for students to draw an appropriate vector representation needed to solve this problem. In other words, the subjects were required to do other tasks in addition to geometric manipulation. A subject in vector representation condition (subject A, see Table 1) was interviewed after the experiment. He said that it was difficult to know how to use the vector $(\bar{x}, \bar{x})$ and $(\bar{y}, \bar{y})$.

Only one subject (subject F, see table) was able to solve problem 3. The vector representation needed to solve the problem is similar to the one used to solve problem 1. Thus, it must not be difficult to draw an appropriate diagram. However, no subject in vector representation condition succeeded in solving this problem regardless of the good results for problem 1. The role of vector representation for problem

3 is to help students calculate. In addition to geometric manipulation, the subjects were required to do another task, that is, calculation.

Based on this discussion above, I maintain that the usefulness of vector representations are caused by easy geometric manipulation. If students were required to perform other tasks in addition to the manipulation, the vector representations would not be useful.

General Discussion

Self-explanations.

Each problem used in this study represents the questions that students may have when they study statistics. If the textbook is inadequate by omitting statements about the questions, the students have to generate answers by self-explanations. Chi et al. [1994] emphasized that improvement in learning was achieved when students were merely prompted to self-explanation. However, the results in this experiment suggest that the specific training (for example, leaning vector representations) is needed in order to generate good self-explanations. For example, if students learned statistics without vector representations, they could not solve problem 1.

When and why vector representations are useful.

It should be emphasized that vector representations are often useful for understanding textbooks and problem solving. We can say that the use of pictorial representations is as valuable for university-level mathematics courses as it is for elementary school mathematics. However, it should be also emphasized that vector representations are not always useful.

This representation was useful for understanding the proof about the range of the correlation coefficient and solving problem 1. It may be easy to draw appropriate diagrams for these tasks, and subjects are not required special abilities except for moving or manipulating diagrams geometrically. Larkin and Simon[1987] proposed that diagrammatic reasoning has an advantage because perceptual reasoning is easy.

In contrast to this case, subjects were required to have additional abilities for problem 2 and 3, for example, the abilities to calculate and to draw a difficult diagram. Therefore, we can conclude that the usefulness of vector representations is due to easy geometric manipulation. It stands to reason that the problems become difficult if additional abilities are needed. However the important thing is that this study clarifies where the students' limitation is. Note that the range of usefulness of vector representations is not very wide.

Generally speaking, it is difficult for students to draw an appropriate diagram [e.g., Anzai 1991]. We can not answer the questions what abilities are needed or what training is useful in order to draw appropriate diagrams. Further research is needed about the additional abilities.

References

[Anzai 1991] Anzai, Y. (1991). Leaning and use of representations for physics expertise. In K.A. Ericsson & J. Smith(Eds.) *Toward a general theory of expertise*, Cambridge: Cambridge University Press.

[Chi et al. 1989] Chi, M.T.H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, **13**, 145–182.

[Chi et al. 1994] Chi, M.T.H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, **18**, 439–477.

[Eaton 1983] Eaton, M. L. (1983). *Multivariate Analysis: A Vector Space Approach*, John Wiley & Sons.

[Ferguson-Hessler & de Jong 1990] Ferguson-Hessler, M.G.M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition & Instruction*, **7**, 41 –54.

[Larkin & Simon 1987] Larkin, J., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, **11**, 65–99.

[Takeuchi et al. 1982] Takeuchi, K., Yanai, H., & Mukherjee, B. N. (1982). *The Foundation of Multivariate Analysis.* New Delhi: Wiley Eastern.

[Tanaka & Wakimoto 1983] Tanaka, Y., & Wakimoto, K. (1983). *Methods of Multivariate Statistical Analysis.* Tokyo: Gendai Sugaku Sya. (in Japanese)