韓國數學敎育學會誌 시리즈 E
<數學敎育 論文集>
제 16집, 2003. 8. 139-147.

J. Korea Soc. Math. Ed. Ser. E:
Communications of Mathematical Education
Vol. 16, Aug. 2003. 139-147.

A Case Study of an Activity Based Mathematical Education: A Kernel Density Estimation to Solve a Dilemma for a Missile Simulation

G. Daniel Kim(Oregon Univ.)

While the statistical concept "order statistics" has a great number of applications in our society ranging from industry to military analysis, it is not necessarily an easy concept to understand for many people. Adding some interesting simulation activities of this concept to the probability or statistics curriculum, however, can enhance the learning curve greatly. A hands-on and a graphic calculator based activities of a missile simulation were introduced by Kim(2003) in the context of order statistics. This article revisits the two activities in his paper and point out a dilemma that occurs from the violation of an assumption on two deviation parameters associated with the missile simulation. A third activity is introduced to resolve the dilemma in the terms of a kernel density estimation which is a nonparametric approach.

1. INTRODUCTION

The statistical concept "order statistics" is usually introduced in the senior or first year graduate level probability or statistics classes. It has abundant applications in our world ranging from industry to military analysis. During World War II, British and Americans tried to estimate the number of German tank production on the basis of the order statistics. Unfortunately, it is not a very easy concept to understand for many people. Adding some interesting simulation activities of this concept to the curriculum, however, can enhance the learning curve greatly. A hands-on and a graphic calculator based activities of a missile simulation were introduced by Kim(2003) in the context of order statistics. This article revisits the two activities in his paper and discuss what kind of dilemma can occur when an assumption on two deviation parameters associated with the missile simulation is violated.

Key words and phrases. missile simulation, normal distribution, order statistics, kernel density estimation.

A third activity is introduced to resolve the dilemma in the context of a kernel density estimation which is a nonparametric approach.

Suppose that an enemy target was detected in a combat zone, and a missile was fired to a target point, say (μ_1, μ_2). Obviously, the missile may or may not hit the target. So let us consider two variables X_1 and X_2 that identify the location where the missile hits, where X_1 denotes the east-west coordinate of the location, and X_2 the corresponding north-south coordinate. Since the missile may not hit the target exactly, X_1 and X_2 should be considered as random variables. We make assumptions on X_1 and X_2 as follows: X_1 and X_2 are independent, and have $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions, respectively. Until we reach the Activity III, we will assume that σ_1 and σ_2 are equal to a common value σ.

One would be interested in the amount of the error distance between the target and where the missile hit, which is given by

$$Y = \sqrt{(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2}.$$

The probability distribution of Y is stated in Lemma 1 and its proof can be found in Kim(2003).

Lemma 1. Suppose that X_1 and X_2 are independent random variables subject to normal probability distributions with centers μ_1 and μ_2, respectively, and the common standard deviation σ. If $Y = \sqrt{(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2}$, the probability density function of Y is

$$f(y) = \frac{y}{\sigma^2} e^{-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2}, \quad y \geq 0.$$

In modern air attack by field artillery, missile, or bomb attack, multiple number of shells are fired at time. For example, a 155 mm or 105 mm howitzer company fires 4 to 6 shells at a time. Some multiple rocket launching systems such as M270-MLRS fire up to 12 rockets with a range up to 18 miles in less than a minute. The assessment of the effectiveness of those multiple launching attack systems either by howitzer or by missiles is an important issue with regard to the precision of the attack system.

Suppose that a multiple number, say n, of missiles are fired at a target. Let $Y_1, Y_2, \cdots, Y_n$ represent the error distances made by the missiles. Let $Y_{(1)} = \min\{Y_1, Y_2, \cdots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, Y_2, \cdots, Y_n\}$. Here the minimum order statistic $Y_{(1)}$ represents the shortest error distance to the target, and the maximum order statistic $Y_{(n)}$ the largest error distance to the target. While $Y_{(n)}$ is attained by the hit with the lowest accuracy among the n missiles, $Y_{(1)}$ is attained by the hit with highest accuracy. In Lemma 2, the probability

distributions of the minimum and the maximum order statistics are stated without proofs, which can be found in Kim(2003).

Lemma 2. Suppose $Y_1, Y_2, \cdots, Y_n$ are defined as Y in Lemma 1. Then, the probability density functions of $Y_{(1)}$ and $Y_{(n)}$ are given by

(a) $f_{Y_{(1)}}(y) = \frac{ny}{\sigma^2} e^{-\frac{n}{2}(\frac{y}{\sigma})^2}, \quad y \geq 0.$

(b) $f_{Y_{(n)}}(y) = \frac{ny}{\sigma^2} \left[1 - e^{-\frac{1}{2}(\frac{y}{\sigma})^2}\right]^{n-1} e^{-\frac{1}{2}(\frac{y}{\sigma})^2}, \quad y \geq 0.$

The population standard deviation σ of the two random variables X_1 and X_2 is unknown in nature due to various unexpected natural, technical, and human factors. It will have to be estimated based on a given sample set of outcomes. The following pooled standard deviation s_p was used to estimate σ:

$$s_p = \sqrt{\frac{(n-1)s_1^2 + (n-1)s_2^2}{2n-2}}$$

where n represents the number of missile fired, and s_1^2 and s_2^2 are the sample variances of X_1 and X_2, respectively. Now we utilize Lemma 2 to consider two missile simulation activities.

2. Simulation Activities

Three simulation activities are conducted with a group of 12 students who were taking an advanced mathematical statistics course from the author during the spring term of 2002. In Simulation I, samples of X_1 and X_2 are generated through a dart play, and in Simulation II, graphing calculators are used to generate samples of X_1 and X_2. Simulation III is devoted to the case where σ_1 and σ_2 are significantly different. A nonparametric solution is suggested and discussed.

SIMULATION I: A DART PLAY

The twelve students in the class were grouped in three, and each group was provided with a dart and a dart board. One student in each group was chosen to play the role of a shooter, and the other two students played the roles of observer and recorder. The shooter who stood 1.2 meters away from the dartboard threw the dart to the target point six times. For purposes of convenience we assumed that the target was located at the origin and hence both μ_1 and μ_2 are assumed to be equal to zero. The observer measured the horizontal coordinate X_1 and the vertical coordinate X_2 of the dart. The recorder gathered

the observed values of X_1 and X_2 in a data collection sheet such as Table 1. After six shots, students switched the roles so that everyone in each group could play all the three different roles. After the data collection was completed, each student calculated the following two quantities:

1) The error distance of each shot:

$$Y = \sqrt{(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2} = \sqrt{X_1^2 + X_2^2}$$

2) The pooled sample standard deviation s_p:

$$s_p = \sqrt{\frac{(n-1)s_1^2 + (n-1)s_2^2}{2n-2}} = \sqrt{\frac{5s_1^2 + 5s_2^2}{10}}$$

The pooled standard deviations s_p by the twelve students were averaged out, which yielded 4.82 cm. We assumed that the quantity 4.82 cm represented the common population standard deviation σ of X_1 and X_2 given the situation. In subsequent we assume $\sigma = 4.82$ cm.

Table 1. Data Collection (The sample for shooter #1)

Shot number	X_1	X_2	$Y = \sqrt{X_1^2 + X_2^2}$
1	5.27	-3.79	6.49
2	-9.45	2.01	9.66
3	3.64	1.46	3.92
4	2.11	-1.34	2.50
5	-3.74	-3.43	5.07
6	7.72	1.18	7.81

Table 1 shows the observed values of X_1 and X_2 by the shooter #1, and in this case, the minimum value $y_{(1)}$ is obtained on the fourth shot and the maximum value $y_{(6)}$ on the second shot. Table 2 put together all the observed values of the minimum order statistic $y_{(1)}$ and the maximum order statistic $y_{(6)}$ by the 12 students. The sample means and the sample standard deviations are calculated.

Table 2. Samples of $Y_{(1)}$ and $Y_{(6)}$ of size 12, respectively

Shooter #	1	2	3	4	5	6	...	7	8	9	10	11	12	Mean	SD
$Y_{(1)}$	2.5	4.30	1.80	1.44	3.96	5.13	...	3.06	1.57	3.05	0.81	3.55	3.21	2.63	1.49
$Y_{(6)}$	9.66	13.83	5.09	10.9	19.05	20.40	...	12.01	10.68	6.46	17.86	18.10	16.16	13.35	5.02

To learn what are expected for $Y_{(1)}$ and $Y_{(6)}$, the four theoretical quantities $E(Y_{(1)})$, $E(Y_{(6)})$, $\sigma_{Y_{(1)}}$, and $\sigma_{Y_{(6)}}$ were calculated. Students used their TI-89/83 calculators to evaluate the quantities based on the probability density function $f_{Y_{(1)}}(y)$ and $f_{Y_{(6)}}(y)$ given in Lemma 2. The results are displayed in Table 3 with their corresponding statistics side by side. Figure 1 illustrates three graphs: (a) the probability density function of Y, (b) the adjusted frequency histogram of the observed $y_{(1)}$s in Table 2 and the probability density function $f_{Y_{(1)}}(y)$, and (c) the adjusted frequency histogram of the observed $y_{(6)}$s in Table 2 and the probability density function $f_{Y_{(6)}}(y)$. The probability density curve of Y in (a) illustrates the probability distribution of the error distance Y if a single missile was fired. The continuous curves in (b) and (c) show the probability distributions of $Y_{(1)}$ and $Y_{(6)}$ if six missiles were fired. A TI-89 graphing calculator was used for the graphs with window dimensions Xmin=0, Xmax=22, and the class width=0.5.

Both Table 3 and Figure 1 (c) show that the theoretical expectation of the maximum order statistic $Y_{(6)}$ and the empirical observation of the order statistic are not matching very well. Two probable reasons pointed out by the students are: (1) the expected standard deviation of the maximum order statistic, $\sigma_{Y_{(6)}} = 2.55$, is not so small, and (2) the sample size of $Y_{(6)}$ was just twelve, which students interpreted as a small sample size.

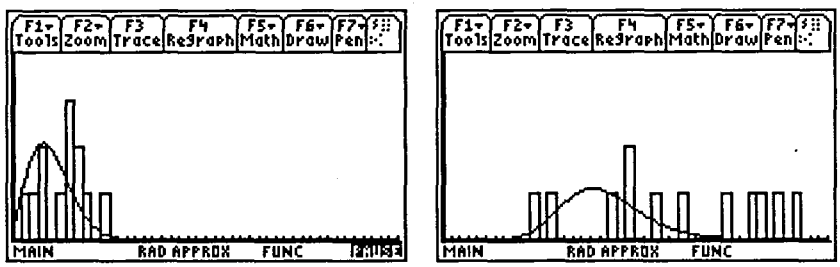


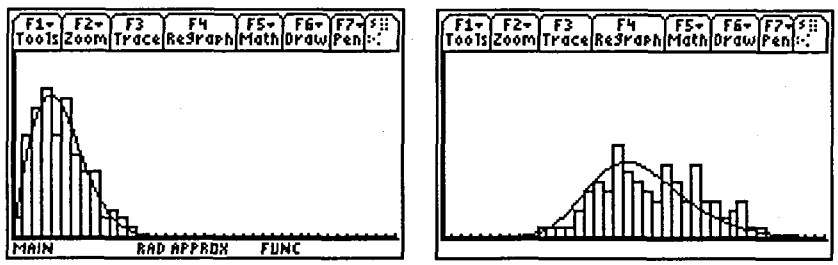
Figure 1. (a) The adjusted histogram of the $y_{(1)}$ in Table 2 and the pdf of $Y_{(1)}$, (b) The adjusted histogram of the $y_{(6)}$ in Table 2 and the pdf of $Y_{(6)}$

Table 3. The parameters and the statistics of $Y_{(1)}$ and $Y_{(6)}$

Mean	Sample mean	SD	Sample SD
$E(Y_{(1)}) = 2.466$	$\bar{Y}_{(1)} = 2.63$	$\sigma_{Y_{(1)}} = 1.29$	$s_{Y_{(1)}} = 1.49$
$E(Y_{(6)}) = 10.36$	$\bar{Y}_{(6)} = 13.35$	$\sigma_{Y_{(6)}} = 2.55$	$s_{Y_{(6)}} = 5.02$

SIMULATION II: A CALCULATOR BASED ACTIVITY

The missile simulation of Lemma 2 was well realized through the data generated by TI-89/83 calculators. Each student used the TI-89/83 built-in function ".randnorm()" or "randnorm()" to generate data sets of X_1 and X_2 of size 6. Then the distance formula, "min" and "max" functions were used to calculate the minimum $y_{(1)}$ and the maximum $y_{(6)}$. Each student repeated this procedure 100 times, instead of 12 times, as if there were 100 students in the classroom. The similar computation and graphing activities as Simulation I were repeated, and the results are presented in Table 4 and Figure 2. Like Figure 1, the graph (a) in Figure 2 shows the probability density curve of Y, (b) the adjusted frequency histogram of the 100 observed $y_{(1)}$s and the probability density function $f_{Y_{(1)}}(y)$, and (c) the adjusted frequency histogram of the 100 observed $y_{(6)}$s and the probability density function $f_{Y_{(6)}}(y)$. As both the numerical values in Table 4 and the graphical comparisons in (b) and (c) of Figure 2 demonstrate, the conclusions of Lemma 2 are very well realized by the calculator-generated data set. A TI-89 was used for the graphs in Figure 2 with the window dimensions Xmin=0, Xmax=20, and the class width=0.5.



(a) (b)

Figure 2. (a) The adjusted histogram of the sample size 100 and the pdf of $Y_{(1)}$, (b) The adjusted histogram of the sample size 100 and the pdf of $Y_{(6)}$

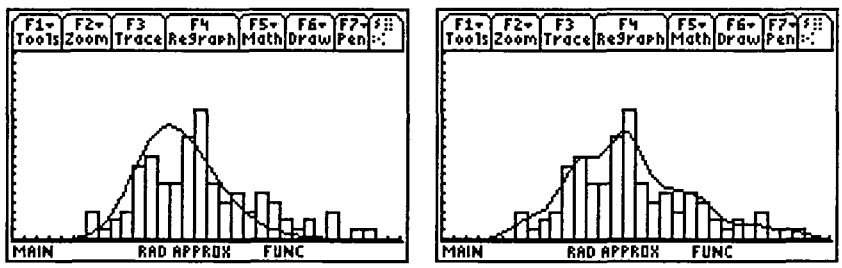
Table 4. Parameters and statistics of the simulated data of size 300

Mean	Sample mean	SD	Sample SD
$E(Y) = 2.466$	$\bar{Y} = 5.835$	$\sigma_{Y_{(1)}} = 1.290$	$s_Y = 4.542$
$E(Y_{(6)}) = 10.360$	$\bar{Y}_{(6)} = 10.241$	$\sigma_{Y_{(6)}} = 2.550$	$s_{Y_{(6)}} = 2.298$

SIMULATION III: A nonparametric model when σ_1 and σ_2 are not equal

The parametric models for the density functions of Y, $Y_{(1)}$, and $Y_{(6)}$ provided in Lemmas 1 and 2 are based on the assumptions that the horizontal coordinate X_1 and the vertical

coordinate X_2 are independent random variables subject to normal probability distributions with $\mu_1 = 0$ and $\mu_2 = 0$, respectively, and shares the common standard deviation σ. Through discussions of dart play activities, students pointed out that in practice there is no guarantee that σ is shared by two variables. More students thought that σ_2 may be larger than σ_1 in many circumstances where σ_1 and σ_2 are the standard deviations of X_1 and X_2, respectively. In order to see how the density functions of $Y_{(6)}$ plays out when the deviation parameters are significantly different, a data of size 100 of $Y_{(6)}$ is generated with $\sigma_1 = 1$ and σ_9. Figure 3(a) shows the histogram of the data and $f_{Y_{(6)}}(y)$ fit when $\sigma_1 = 1$ and $\sigma_2 = 9$. It appears that the density function $f_{Y_{(6)}}(y)$ seems to fit the histogram reasonably except the right-tail domain of the data. In fact, the parametric model significantly underestimates the density of the right-tail domain of the data. As it can be seen in Table 5, the mean from the parametric model is 13.77, which is a quite less value than the mean of the sample, 15.58.



(a) (b)

Figure 3. (a) The parametric fit $f_{Y_{(6)}}(y)$ of the simulated data of $Y_{(6)}$ when $\sigma_1 = 1$ and $\sigma_2 = 9$, (b) The kernel density fit of the same data of (a). The data is of size $n = 100$.

This brings up a question of whether there is a better density model that fits the data in the entire domain when the two deviation parameters σ_1 and σ_2 are severely different. It appears that solving this dilemma is not necessarily a simple task if the proof of Lemma 1 is closely examined. In fact, the whole proof of Lemma 1 is on the basis of the equality of σ_1 and σ_2, and this was well understood and agreed by students as we made discussions on this issue. As a way to get around the theoretical difficulty, a nonparametric approach called the kernel density estimation model was suggested. The kernel density estimate is a data-driven empirical function that smoothes out the shapes of the histogram in a neighborhood of a

given y value according to a kernel function. It is formally defined as

$$\hat{f}_h(y) = \frac{1}{nh} \sum_{j=1}^{n} w\left(\frac{y - y_j}{h}\right).$$

Here, h is a bandwidth, $w(\cdot)$, called a kernel function, is a continuous and symmetric probability density function, and n is the size of data. For our classroom use, $h = 1.5$, the standard normal density function for the place of $w(\cdot)$, and $n = 100$ were applied. Figure 3(b) is the kernel density estimate of the histogram with Xmin=0, Xmax=32, and the class width=1. As it shows, the nonparametric estimate does a great job to fit the histogram in the entire domain of the data. Especially it fits well in the right domain of the data where the parametric density model failed to estimate its exact densities. Table 5 compares the numbers of the mean and the standard deviation of the data, the parametric model, and the kernel density estimate. It indicates that the mean and the standard deviations of the kernel density estimate model match with the simulated data a lot better than the ones of the parametric model do.

Table 5. The means and the standard deviations of the data, the parametric model, and the kernel density estimate model

	Data of size 100	Parametric Model	Kernel Estimate Model
Mean	15.580	13.770	15.575
SD	4.890	3.370	5.000

3. Discussion

The statistical term "order statistic" has a number of important applications in our society. Based on years of teaching experience the author perceived that average students see the subject "order statistics" as a rather difficult concept to digest. In this article three missile simulation activities with the perspectives of order statistics were considered. It was the author's perception that the missile simulation activities held the students' attention, and the concept of order statistics was delivered to the students successfully. Another beneficial effect of the activity based curriculum was that the class came up with a variety of fun and rich discussions. Some of them are summarized below:

1) Ladies beat Gentlemen?: The twelve students were composed of six male students and six female students. As Table 2 of Simulation 1 indicates, the best shot among the 12 shooters was the shooter #3 whose shots range from 1.80 cm to 5.09 cm.

Interestingly enough, the shooter was a female student, not a male student. While there is absolutely no intention to claim that female students shoots better in dart play than male students do, observing such a thing in a small classrooms added a lot of fun to the class.

2) Should σ be common to both X_1 and X_2?: It seems that for minor difference between σ_1 and σ_2, the parametric density model of the maximum order statistic $Y_{(6)}$ fits the histogram of the data in reasonable manner. However, it appears that the quality of the fit diminishes as the difference between the two deviation parameters gets larger. Especially, the parametric model underestimates the density in the right-tail domain of the data. This may cause serious underestimation of the risk of collateral damage by the worst fire in real battle field when multiple shots are fired. In Activity III, we studied that a nonparametric solution could be used to solve this dilemma.
3) Three dimensional target and Moving Target: The students saw the possibility that without much technical difficulty, Lemma 2 can be extended to a three dimensional target point (μ_1, μ_2, μ_3). They also understood that if the target was a moving object instead of a steady object, the target point has to be identified as a stochastic process, and simulating such an idea would be challenging.
4) Dart Play versus Calculator: As discussed previously, Lemma 2 was better realized by the calculator-generated data set than by the dart play-generated data set. Students pointed out, however, that the dart play provides more realistic situation because in battle field there exist a lot of factors that are uncertain and unpredictable. As Figure 1 (c) shows, in practice actual results can deviate from theory considerably, in their case being quite off the mark.

References

[1] Kim, G. 2003. Order Statistics and a Missile Simulation Activity. *PRIMUS*, Vol XIII, No. 2., 182–193.

MATHEMATICS DEPARTMENT, SOUTHERN OREGON UNIVERSITY, ASHLAND, OR 97520, USA.
E-mail address: kimd@sou.edu