COMPARISON OF DATA SETS AS A PRECURSOR TO INFERENTIAL STATISTICS

[1]PARK, Min-Sun, [1]PARK, Mimi, [1]KO, Eun-Sung, and [2]LEE, Kyeong-Hwa
[1]Graduate School of Seoul National University, Korea,
[2]Seoul National University, Korea
Minsun26@gmail.com

ABSTRACT

Comparing two data sets can be a powerful tool in light of its use toward a consideration of inferential statistics. Both informal and formal statistical reasoning are developed when comparing data sets, which has implications for researchers who investigate ways to help students transfer from informal to formal reasoning. In this paper, we examined students' reasoning to identify how they treat data value, center, spread, and sample, which are important factors in inferential statistics. Students' understanding of data value, center, and spread were appropriate, but that of sample was not. From the results, we suggest instructional ideas for a task which can connect descriptive and inferential statistics.

INTRODUCTION AND BACKGROUND

Students learn formal inferential statistics through coursework, which deals with probability distributions, hypothesis testing, and p-value. However, there is little or no research that shows how teachers or students who begin as statistical novices reach a well-developed understanding of formal inference (Rubin, Hammerman, & Konold, 2006, p.1). As an alternative to formal inference, researchers introduced the term "informal inferential reasoning" to emphasize to students the importance of connection (Zieffler, Garfield, DelMas, & Reading, 2008; Makar & Rubin, 2009). However, there is still a lack of research about the complexity of informal inferential reasoning and the methods for connecting descriptive and inferential statistics. Connections between types of statistics continue to elude many teachers and students (Pfannkuch, 2006, p.27). Thus, it is important to identify the factors that can lead students' informal inferential reasoning to inferential statistics.

Previous studies showed that comparison of data sets helped students engage in reasoning about distributions (Konold & Higgins, 2002), which means that data set comparison is related to descriptive statistics. On the other hand, the comparison of data sets is also related to inferential statistics. The ability to compare leads directly to statistically valid decisions and inferences about situations involving data (Ciancetta, 2007, p.15). In addition, the comparison of two data sets can be a powerful tool towards a consideration of inferential statistics (Makar & Confrey, 2004).

Researchers have argued that students should begin to develop informal ideas of inference at early grades, to be better able to later learn and reason about formal methods of statistical inference (Zieffler et al., 2008, pp.45-46). Zieffler et al. (2008, p.44) defined informal inferential reasoning as the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples. Makar and Rubin (2009, p.85) defined it as a reasoned but informal process of creating or testing generalizations from data. Rubin et al. (2006, pp. 1-2) suggested that informal inferential reasoning involves the following related ideas: Properties of aggregates, types of variability, sample size, controlling for bias, and tendency. Pfannkuch (2006), who examined the factors including summary, signal, spread, and sampling that teachers use when they compare boxplots, believed that it is helpful to connect informal inferential reasoning to formal inference.

In this paper, we identified aspects of students' reasoning from data set comparison situations that can serve as a basis for transitioning from informal inferential reasoning to inferential statistics. By examining the aspects, we suggest instructional ideas for connecting descriptive and inferential statistics.

METHOD

PARTICIPANTS

The participants were three groups of seventy students in grades 7, 8, and 9. They had learned neither formal inferential statistics nor the comparison of data sets. In the Korean statistics and probability curriculum, grades 7 to 9 include mean, range, the number of cases, and basic probability (figure 1). Since students in grades 7 to 9 can compare data sets using their basic statistical knowledge, they are appropriate participants.

	Grade	Area	Contents
	3-4	Statistics	Bar graph, graph of broken line
	5-6	Statistics	Stem-leaf chart, pictograph, pie graph, average
7th graders		Probability	The number of cases, probability
8th graders	7	Statistics	Frequency table, histogram, relative frequency, absolute frequency
9th graders	8	Probability	The number of cases, probability
	9	Statistics	Mean, median, mode, variance, standard deviation

Figure 1. Statistics and probability curriculum in Korea

TASK

The task consists of two questions taken from previous studies on the comparison of data sets (figure 2, Shaughnessy, Ciancetta, Best, & Canada, 2004; Ciancetta, 2007). All of the data sets were from sample data, not populations. The data sets in question 1 have the same mean and different ranges. The data sets in question 2 have a similar mean, but are of different sizes. Both questions involve realistic context, movie theater wait times and ambulance response times.

Task 1.

A class of 21 students investigated the wait-times at two popular movie theater chains: A Movie Theater and B Movie Theater. Each student attended two movies, a different movie in each theater. The class's results are shown below. The wait-time for a movie is the difference between the advertised start time and the actual start time for the movie.

A movie theater

5.0	12.0	13.0	5.5	9.5	13.0	5.5
11.5	8.0	8.5	14.0	13.0	8.5	7.0
8.5	12.5	13.5	11.5	9.0	10.0	11.0

(mean=10 min)

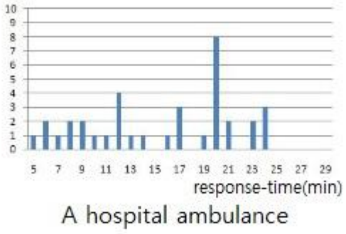
B movie theater

11.5	11.0	9.0	10.5	8.5	11.0	9.0
10.5	9.5	8.5	10.0	11.5	10.5	8.5
9.0	11.0	11.0	9.5	10.0	9.0	11.0

(mean=10 min)

Task 2.

The school board had to make a decision about which of two ambulances service companies to call when emergencies arise at their school. The two ambulance companies near the school are A hospital ambulance and B hospital ambulance. The school board members obtained the 36 most recent response times for A hospital ambulance and the 74 most recent response times for B hospital ambulance. These response times are shown below.



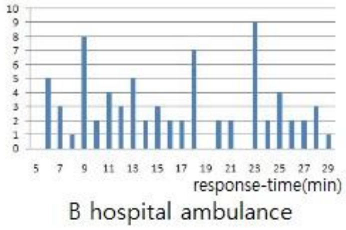


Figure 2. Comparison tasks

DATA COLLECTION AND ANALYSIS

Students were given various questions related to comparing data sets. The seventh-graders solved only the movie theater question, which was relatively easier, while the eighth- and ninth-graders solved both of the questions. After the students solved the questions, they discussed their selections. The data collected consisted of the questionnaire, audio files, and video files. For the data analysis, all class audio and video files were transcribed. From the transcriptions of students' discussions, data were analyzed on four factors: data values, center, spread, and sample. The researchers focused on how students used the four factors to compare the data sets.

RESULTS AND DISCUSSION

DATA VALUES

In inferential statistics, the object of comparison is the sample, not the population, and the size of each data set is usually different. Thus, it is necessary to recognize each of the data values in the data sets as an aggregate, rather than an individual point (Yang & Chang, 2010).

01 S1: If we enlarge A hospital ambulance's data 2 times, then they look similar.
02 Teacher: What do you mean by 'enlarge 2 times'?
03 S1: Since the number of A's data is about half that of B's, we should compare them after expanding A's graph to twice its size.

In question 2, S1 recognized that he should make the data sets the same size using proportions before making the comparison, because the data sets were different sizes (line 03). It means that S1 used multiplicative reasoning, considering the proportion.

Students could use multiplicative reasoning when the sizes of the data sets were unequal. Their natural use of multiplicative reasoning according to the sizes of the data sets shows that they can make the connection between data values and the whole data set. Also, they saw the data values as an aggregate by using proportions. To foster the view of considering data sets as an aggregate, it would be helpful to give students data sets of different sizes. A problem involving data sets of different size sets is sometimes a higher-level task; however, it allows students to see the data values in general.

CENTER

In inferential statistics, distribution center is determined according to the mean and variance. When students check the significance of the difference between the means, they need to consider sampling variability. Thus, it is necessary to consider intervals as a comparison standard in inferential statistics.

When students solved question 1, there were two changes of the comparison standard. The first one was the change from the specific point to the mean.

04 Teacher: How about A movie theater?
05 S2: The number of long wait-times is bigger than the number of short ones.
06 S3: Which is included in the set of long wait-times? What is the standard? I think 14 min. could be included in the set of short wait-times.
07 S4: The standard is not clear.
08 S2: The standard is 7 min.
09 S4: How did you arrive at that conclusion?
10 S2: Just because. Since 14 min. is the longest, 7 min. is just a half as long.
11 S4: Then there is also the shortest one.
12 S3: 5 min. is the shortest. How about 10 min.? 10 is 5 times 2.
13 S2: Oh, that's right.
14 Teacher: How can we decide on the long wait-times and short wait-times in graph A?
15 S3: If we use "compared with the mean," then we can.

16 S2: There are more long wait-times than short wait-times when the mean is used as a standard.

S2 distinguished between short wait-times and long wait-times for the data sets of movie theater A (line 05). S3 asked a question about the standard for distinguishing between wait-times that are short and those that are long, and that the maximum value of 14 min. was still short for him (line 06). Then S2 recognized that 7 min. was half of the maximum value (lines 08 and 10). S4 also raised the issue that there was also the minimum value (line 11). After that, S2 recognized the problem with setting a specific point as the center (line 13). As an alternative solution, S3 suggested that the mean could be the center (line 15), and S2 agreed with this idea (line 16). While discussing, students recognized that it was not possible to get everyone's consensus on setting a specific point as the center, so it was better to set the mean as the center.

The second change was the change from the mean to the interval including the mean.

17 S5: If we exclude the mean, then there is a stable range. That is, the range extends from a little bit later and a little bit earlier (than the mean). I think this would be a more proper standard than the mean.
18 Teacher: You mean you want to change the standard?
19 S5: I think it is better to consider 1 or 2 min before and after the mean along with the mean rather than the mean alone. Taking the range as a standard, A has more long wait-times.

S5 knew that a little time earlier and later than the mean did not make a big difference, so he called it "a stable range" (line 17). He recognized that a range around the mean would not affect things much, so he thought that an interval was a better idea (line 19).

It is important to focus on the point that students came up with the idea of using a range including the mean, not just an overlapping range. To encourage students to consider an interval as the center, a task that includes data sets with the same mean but a very different variance could be helpful. When recognizing the difference, students can recognize the necessity of considering both mean and variance at the same time. They can set the interval as the range.

SPREAD

In inferential statistics, statistics like variance and standard deviation are used for hypothesis testing and comparing the variability of data sets in analysis of variance. The students used various types of spread when they compared the two data sets. When they solved question 1, the movie theater, we could find three kinds of spread: range, variability, and variance. The first one noticed was the use of the range.

20 S6: By using common sense, I can see that the possibility of watching the movie sooner at A than B is greater.
21 S7: Whatever. Don't regret it after you wait for 13 min.
22 S6: I hope you wait for 11.5 min while I wait for 5 min.
23 S7: That is the biggest value! Your biggest is 14. You can't say you will wait for 5 min. That is just by luck.

Students who used range as a measure of spread usually mentioned maximum and minimum values. S6 chose movie theater A by observing the minimum value 5 (line 22).

The next option considers variability as a measure of spread. Strictly speaking, variability is hard to see as a measure. However, students regarded the shape of each data set's spread as a degree of variability, so variability also could be a measure of spread.

24 Teacher: Why did you choose movie theater B?
25 S8: I think stability should be a greater consideration than the frequency.
26 S9: A movie theater is risky. Movie theater B starts from 8 min. to maximum of 11 or 12 min., so it is more stable.

S8 and S9 recognized that if the data set has small variability and is close to the mean, then it was more stable (lines 25 and 26). When considering spread, there is a difference between considering only the range and considering variability. Considering the range depends on the specific maximum and minimum values, while considering variability is recognizing the stability of the distribution.

Finally, some students viewed variance as a measure of spread. It means that they considered the distance of the values from the mean, as well as how many values are apart.

27 S7: We should not just consider the times and possibility. We don't know how much profit each value makes. If we divide the 3 intervals into before 8.5, from 8.5 to 11.5, and after 11.5, the profit of the first interval is 11.5. So by dividing by 5 which is the number of value, we get 2.3. And the profit of the third interval is 10.5, so by dividing by 7, it is about 1.5. So A's profit is bigger than that of B.

S7 recognized the need to not only compare the number of data values in the range, including the mean, but also find out how far each data value is from the range. For the range below 8.5, she added every data values' distance from 8.5 and got 11.5. By dividing it by 5, the number of data points, she got 2.3. In the same way, for the range above 11.5, she added every data values' distance from 11.5 and got 10.5. By dividing it by 7, she got 1.5. Since 2.3 is greater than 1.5, she concluded that movie theater A would be the better choice.

The change from range, through variability, to variance when students considered the spread can be connected to emphasizing variance and considering variability within and between data sets in inferential statistics. To guide students to have a proper view of spread, teachers can present tasks in which the data sets have the same mean but a significantly different variance, as we mentioned in the former section. To find the difference in variance, students would consider which data set is more stable or they would calculate some statistics.

SAMPLE

Usually in inferential statistics, the object of inference is a sample, not a population. Thus, understanding the sample is essential. Students need to understand both variability and representativeness, and the characteristics according to the size of the sample (Saldanha & Thompson, 2002). In the given tasks, some students lacked an understanding of sample.

28 Teacher: Do you agree there is no difference between A's and B's wait-times?
29 S10: Yes. If we compare this situation with the lottery, while the possibility of winning the lottery does not change in the whole country, there is still a winner.

30 S11: These graphs represent only 21 cases among a far greater number of cases. So just considering these data's percentage would show a totally different result from the real result.
31 S12: These graphs could be wrong.

S10 said that he could not compare the difference by mentioning the invariability of the total probability (line 29). Even though he should have considered the probability of one event, he did not recognize the importance of variability. On the other hand, S11 and S12 considered variability excessively; they seemed to deny the representativeness of the sample by saying "a totally different result from the real result" and "could be wrong" (lines 30 and 31).

Meanwhile, there were two interpretations about the sample size effect.

32 S13: I think B is the better hospital, because more people go there. I think B is better.

33 Teacher: Some of you chose B. Why?
34 S14: B's size of data is much bigger.
35 Teacher: What does that mean?
36 S15: Accuracy.
37 Teacher: What is the advantage of having data of bigger size?

38 S16: It comes more often.
39 S14: It has many experiences. And the reliability is higher.

In line 32 and the dialogue from lines 33 to 39, students' opinions about the effect of the sample size are presented. S13 believed that since hospital B's data set was bigger, hospital B was better (line 32). It showed that S13 recognized the difference of the sizes of the data sets. However, he could not interpret the results in a statistical way and consider the results in a contextual way; his interpretation was an example of inference not based on data. On the other hand, S14 and S15 mentioned the words "accuracy" and "reliability" when interpreting the size of the data sets. They seemed to interpret the results in a statistical way (lines 36 and 39).

Most students were not good at understanding the variability and representativeness of the sample. Also, some students were unable to perform statistical thinking about the effect of the sample size, which is sampling variability. To support the understanding of the concept of sample, the task should include data sets which are samples. Students need to be given opportunities to explain the sample. In this way, they can understand the concept of a sample and see the relationship between the sample and the population.

CONCLUSION

In this paper, we identified factors that can connect both informal inferential reasoning and inferential statistics to the comparison activities of students in grades 7 to 9. When students compared the data sets, they could understand data value, center, and spread as taking steps toward inferential statistics. However, they lacked a strong understanding of samples.

We also identified the importance of tasks. To foster the view of data values as an aggregate, we need to provide students with a task which includes data sets of different sizes. To help students set an interval which includes the mean for the center, and consider variance and variability for the spread, we need to give them a task in which two data sets have the same mean but different variances. To help students understand the sample, we need to use data sets which show the relationship between the sample and the population. By considering these conditions for task data sets, students can see the data sets overall, understand both variability within data sets and between data sets and sampling variability, and infer more clearly about the population from the sample.

REFERENCES

Ciancetta, M. A. (2007). *Statistics Students Reasoning when Comparing Distributions of Data*. Unpublished Doctoral Dissertation, Portland State University, Portland.

Konold, C., & Higgins, T. (2002). Highlights of related research. In S. J. Russell, D. Schifter, & V. Bastable (Eds.), *Developing Mathematical Ideas: Working with Data* (pp.165-201). Parsippany, NJ: Seymour Publications.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp.353-373). Dordrecht: Kluwer Academic Publishers.

Makar, K., & Rubin, A. (2009). A framework for teaching about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistical Education Research Journal*, *5*(2), 27-45.

Rubin, A., Hammerman, J., & Konold. C. (2006). Exploring informal inference with interactive visualization software. In *Proceedings of the 7th International Conference on Teaching Statistics*. Salvador, Brazil.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Shaughnessy, J. M., Ciancetta, M., Best, K., & Canada, D. (2004). Students' attention to variability when comparing distribution. Paper presented at *the Research Presession of the 82nd Annual Meeting of the National Council of Teachers of Mathematics*, Philadelphia, PA.

Yang, K., & Chang, W. (2010). Strategies to make counter-examples while comparing two groups. In *Proceedings of the 8th International Conference on Teaching Statistics*. Ljubljana, Slovenia.

Zieffler, A., Garfield, J., DelMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58.