DEVELOPMENT OF INTEGRATED QUANTITY JUDGMENTS: MEANS OF DISTRIBUTIONS APPEAR MORE DIFFERENT THAN THEY ARE

Franca Agnoli[1], Gianmarco Altoè[2], and Tatiana Marci[2]
[1]University of Padova, Italy
[2]University of Cagliari, Italy
franca.agnoli@unipd.it

Inferential statistics are used to decide whether two or more samples are from the same or different distributions, a decision that is generally difficult to make by visual inspection of sample frequency distributions. We investigate how children (8, 10, and 12 year old) and adults compare two sets of five vertical bars, similar in appearance to histograms, to determine which set represents a greater quantity or whether they were equal quantities. Our bars are conceptually simpler than histograms because only bar lengths matter, not their position. Participants of all ages correctly identified about 75% of the sets with greater quantity. Judgment accuracy was, however, strongly affected by age when sets contained equal quantities, increasing from 13% correct for 8 year olds to 61% for adults. Recognizing equality is difficult when integrating across multiple bars, and difficulty increases with variability. Implications for comparing statistical distributions are discussed.

STATISTICAL REASONING AND THE EFFECT OF VARIABILITY

Variability (i.e., dispersion of observed data) has a central role in statistics and in quantitative decisions in everyday life. Scientists must often decide whether two collections of measurements are samples from the same distribution or are from different distributions. A definitive answer to this problem is rarely possible because of sampling variability, and consequently scientists employ statistical inferential methods that yield probabilistic solutions. The mechanics of these statistical methods are relatively easy to learn and are commonly taught to undergraduate students in the social sciences, but the concepts that underlie statistical reasoning are difficult to learn (Garfield, 2003; Garfield & Ben-Zvi, 2008; Safran, 2010). Even students who complete a university statistics course and successfully demonstrate their mastery of these techniques may have little understanding of statistical reasoning (Garfield, delMas, & Chance, 2007).

Research suggests that adult students fail to understand statistical reasoning because it clashes with their informal understanding (or misunderstandings) of variability (e.g., Noss, Pozzi, & Hoyles, 1999; Makar & Confrey, 2004). delMas and Liu (2005; 2007) found that undergraduate students who had studied distributions, central tendency, and variability in a statistics class could not judge which of two histograms represented a distribution with a larger standard deviation. Their participants judged that a histogram showing a uniform distribution had a lower standard deviation than a histogram with a large bar in the middle and smaller bars on either side. They confused the variability of bar heights with the variability of the data represented by those bars. As this study demonstrates, histograms are conceptually challenging. Proper interpretation of a histogram requires integrating information about the positions and lengths of all its bars, and comparing two histograms adds to the challenge.

DEVELOPMENT OF EFFECTS OF VARIABILITY ON QUANTITY JUDGMENTS

Altoè and Agnoli (2013) simplified the task of histogram comparison by eliminating the positional value of the bars. They simply asked participants to compare the total quantities represented by the bars without regard to their positions. This task is conceptually much easier than histogram comparison and can be performed by both children and adults. Altoè and Agnoli used this task to study the development of effects of bar length variability on quantity judgments. Children's understanding of and reasoning about variability have been investigated at different ages and in diverse contexts within the field of statistics education (Garfield & Ben-Zvi, 2008), but there has been little systematic study of the development of statistical reasoning because the tasks performed by children and adults have generally been very different.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Altoè and Agnoli (2013) investigated developmental changes in the effect of variability on quantity judgments. Participants (4-year-olds, 5-year-olds, 6-year-olds, 8-year-olds, 12-year-olds, and university students) compared the quantities in two sets of five vertical bars similar in appearance to histograms, as shown in Figure 1. Children were told that these were bars of chocolate, and were asked, "Which side has more chocolate? This side, that side, or are they the same?"

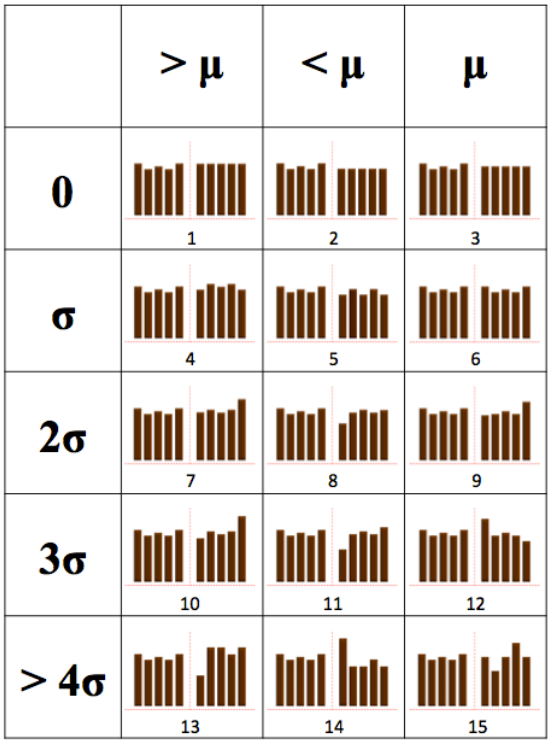

Figure 1. The 15 stimuli from Altoè and Agnoli (2013). The comparison set was larger (column 1), smaller (column 2) or equal (column 3).

One set was held constant (the left-hand set in each box of Figure 1) with mean μ = 7.50 cm and variability σ = .36 cm. Mean bar length in the comparison set was 7.86 cm (column 1 in Figure 1), 7.14 cm (column 2) or 7.50 cm (column 3). As expected, Altoè and Agnoli found that performance increased monotonically with age, but the effect of variability was complicated. As expected, the performance of adult university students and the oldest children decreased as variability increased, but younger children (4- to 8-years-old) performed poorly when variability was very low or very high, achieving highest performance for intermediate variability.

The quantities were equal in a third of the stimulus sets (stimuli 3, 6, 9, 12, and 15 in Figure 1), and recognizing their equality was strikingly difficult for all ages. The mean percentage correct for these trials was only 12% for children averaging across all ages and 61% for adults. If a similar bias occurs with histograms, it could cause people to perceive differences that do not exist.

The current research extends these results in two ways. First, it includes 10-year-old children, who are expected to achieve performance between the 8- and 12-year-old children studied by Altoè and Agnoli (2013). Second, it explores the effects of two alternative modes of presentation of the bar sets and extends the research to older adults who are not engaged in the academic and educational worlds.

REPLICATION WITH 10-YEAR-OLD CHILDREN

We replicated Altoè and Agnoli (2013) in another age group (10-year-olds). The participants were 44 children from five schools in Sardinia, Italy. Children were tested individually at a computer. After completing three training trials, children viewed each of the 15 stimuli shown in Figure 1 in random order and responded (by pointing and speaking) which of the two sets was greater or whether they were equal in quantity. No feedback was given regarding the accuracy of the responses.

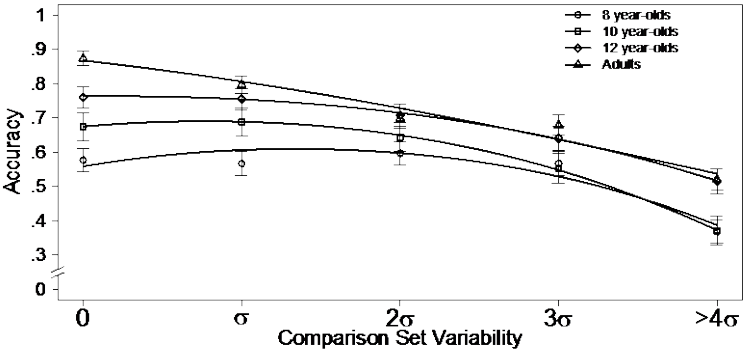

Figure 2. Proportion correct by age and stimulus variability with standard error bars. Lines represent estimated effects of the mixed-effects model (n = 257). Ages 8, 12, and adults are from Altoè and Agnoli (2013).

Figure 2 presents the proportion correct as a function of bar length variability for these 10-year-old children and the 8- and 12-year-old children and adults from Altoè and Agnoli. Because the data were repeated measurements of a categorical response, a logistic mixed-effects model was used with accuracy as a dependent variable. Fixed effects were age, standard deviation of bar lengths, and quantity in the comparison set (more, less, or same as the constant set). As expected, accuracy increased monotonically with age ($\chi^2(4) = 84.88$, $p < .001$). The 10-year-old children performed better than the 8-year-olds ($p = .041$) but not as well as the 12-year-olds ($p < .001$), who were not significantly less accurate than adults ($p = .056$).

The variability of bar lengths affected performance but the relationship was complex, with significant linear (χ^2 (1) = 167.40, $p < .001$) and quadratic (χ^2 (1) = 18.02, $p < .001$) terms. An increase in variability from zero to $\sigma = .36$ cm has little effect on children's performance, but when variability of bar lengths increases to more than 4σ, comparisons of the quantities represented by the two sets becomes very difficult. As Figure 2 shows, this quadratic effect is apparent in children but not in adults, resulting in significant interactions between both the linear term and age (χ^2 (3) = 8.19, $p = .042$) and the quadratic term and age (χ^2 (5) = 12.01, $p = .007$).

Accuracy was much lower when the two sets of bars contained equal quantities (χ^2 (2) = 485.60, $p < .001$), as shown in Table 1. Furthermore, the increase in performance with age appears to be largely due to the stimuli with equal quantities. When quantities were different, there was little difference in accuracy across the age range from 8-years-old to university students, but accuracy increased monotonically with age when quantities were equal, resulting in a significant quantity by age interaction (χ^2 (6) = 139.17, $p < .001$).

Table 1. Percentage correct (and standard error) for equal and different stimulus configurations by age

Age	Different	Equal
8	73.6 (1.7)	13.2 (1.9)
10	75.5 (2.1)	25.0 (2.9)
12	75.6 (1.7)	51.2 (2.8)
Adult	76.3 (1.5)	61.4 (2.4)

REPLICATION WITH ADULTS

The children who participated in this research (and in Altoè & Agnoli, 2013) all were tested one at a time, viewed the stimuli on a computer screen, and responded by pointing and speaking. The university students, in contrast, participated in a group, viewed the stimuli on sheets of paper, and responded on paper. Because these differences in methodology could contribute to the observed differences between adults and children, we asked adults to perform the task as it was performed by the children, and we added another representation using wooden blocks.

We replicated this research with a very different population of adults and compared two presentations of the stimuli. The participants were 59 adults from Sardinia, Italy with mean age of 43 years, and were recruited from a continuing-education program. None had attended a university and only 27 had completed high school. They participated one at a time, with 27 participants viewing the same stimuli as the children and the remaining 32 participants viewing sets of wooden blocks constructed to match the height, width, and configuration of the computer-based stimuli. They responded by speaking and pointing to the stimulus.

Table 2 presents the percentages correct averaged across stimulus modality as a function of stimulus variability and quantity (equal or different). The effects of stimulus modality, stimulus variability, and quantity (equal or different) were tested using logistic mixed-effects models, and there was no significant main effect or interaction with the stimulus modality.

Table 2. Percentage correct (and standard error) for equal and different stimuli by variability ($n = 59$)

Variability	Different	Equal
0	82.2 (3.5)	37.3 (6.3)
σ	71.2 (4.2)	54.2 (6.5)
2σ	75.4 (4.0)	28.8 (5.9)
3σ	81.4 (3.6)	32.2 (6.1)
$>4\sigma$	52.6 (4.6)	28.8 (5.9)
Mean	72.5 (1.8)	36.3 (2.8)

Both stimulus variability ($\chi^2(1) = 18.30$, $p < .001$) and stimulus quantity ($\chi^2(2) = 103.05$, $p < .001$) were significant main effects and their interaction was not significant. Performance decreased as variability increased, which again is largely due to a sharp decrease in accuracy when variability exceeds 4σ. Participants accurately recognized when the quantities in the two sets were equal only on 36% of trials, just half the performance they achieved when the quantities were different. Comparing with Table 1, we observe that when quantities were different the performance

of these adults was about the same as the older children and adults studied by Altoè and Agnoli (2013), but when quantities were equal their performance was more similar to 10- or 12-year-old children. Recognizing that quantities are the same is difficult.

CONCLUSION

Judging the relative sizes of two quantities in the presence of variability is a fundamental problem in all the sciences. In the sciences and in everyday life people may view data representations such as histograms and form opinions about magnitudes. Interpreting histograms is complicated because we must consider both the value and the quantity represented by each bar and integrate across all the bars to assess the values they represent. We sought to explore this problem in a simplified setting by decoupling bar quantity from bar value. The participants in our experiments do not need to consider the positions of the bars because positions are irrelevant in a judgment about quantity. Participants need only integrate the quantities represented by the bars.

These judgments are surprisingly difficult, and become more difficult as the variability in the lengths of the bars increases. Judgments are most difficult when the two sets of bars contain equal quantities. When the quantity is greater in one set than the other, older children and adults respond correctly about 70% of the time. But when the quantities are equal, accuracy increases strongly with age from 13% correct for 8-year-olds to 61% correct for university students. Difficulty recognizing quantitative equality may be dependent on education or may continue to evolve with age in adulthood, because our older, less educated adults responded correctly to only 36% of the equal stimuli. This strong bias for finding one set to be larger than the other was evident across all age groups.

Comparing histograms requires considering both quantity and position, and is often done to determine whether there is evidence that two samples come from the same distribution or different distributions. Our research finds a strong bias for perceiving quantity differences that do not exist when bar position is irrelevant. This bias could contribute to the perception of mean differences in histograms that do not exist, encouraging the belief that samples from the same distribution arise from distributions with different means. We hope to extend this research to histograms and explore ways to reduce this bias and its impact on decision making.

REFERENCES

Altoè, G., & Agnoli, F. (2013). The effect of stimulus variability on children's judgements of quantity. *Journal of Cognitive Psychology, 25*(6), 725-737.

delMas, R. C., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55–82.

delMas, R. C., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 87-116). Mahwah, NJ: Lawrence Erlbaum.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*, 23-28.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: connecting research and teaching practice*. Springer.

Garfield, J., delMas, R. C., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 117-147). Mahwah, NJ: Lawrence Erlbaum.

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), The challenges of developing statistical literacy, reasoning, and thinking (pp. 353-374). Dordrecht, Netherlands: Kluwer.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Statistics in practice. *Educational Studies in Mathematics, 40*, 25-51.

Safran, J. R. (2010). What is statistical learning, and what statistical learning is not. In S. E. Johnson (Ed.), *Neuroconstructivism. The new science of cognitive development* (pp. 180-194). New York, NY: Oxford University Press.