Psychonomic Bulletin & Review
2010, 17 (4), 499-503
doi:10.3758/PBR.17.4.499

Sample size bias in the estimation of means

Andrew R. Smith
University of Iowa, Iowa City, Iowa

and

Paul C. Price
California State University, Fresno, California

The present research concerns the hypothesis that intuitive estimates of the arithmetic mean of a sample of numbers tend to increase as a function of the sample size; that is, they reflect a systematic *sample size bias*. A similar bias has been observed when people judge the average member of a group of people on an inferred quantity (e.g., a disease risk; see Price, 2001; Price, Smith, & Lench, 2006). Until now, however, it has been unclear whether it would be observed when the stimuli were numbers, in which case the quantity need not be inferred, and "average" can be precisely defined as the arithmetic mean. In two experiments, participants estimated the arithmetic mean of 12 samples of numbers. In the first experiment, samples of from 5 to 20 numbers were presented simultaneously and participants quickly estimated their mean. In the second experiment, the numbers in each sample were presented sequentially. The results of both experiments confirmed the existence of a systematic sample size bias.

People must often make judgments about the average or typical member of a group on a single quantitative dimension. For example, a teacher might be asked by his students what the class average was on an exam. Or a survey respondent might be asked to report the average number of times she engages in a given behavior (e.g., consumes an alcoholic drink) per day, week, or year. Although previous research has found that such *central tendency judgments* tend to be accurate—a conclusion that we do not dispute—we hypothesize that they also reflect a systematic *sample size bias*. That is, they tend to increase as a function of the sample size, so that larger groups are judged to have greater central tendencies than smaller groups are. Furthermore, we believe that this is a fairly general bias that has implications for understanding a variety of judgment phenomena and also basic processes involved in quantitative reasoning and judgment.

Early researchers who studied central tendency judgments conceptualized people as "intuitive statisticians" (Peterson & Beach, 1967) and found them to be quite accurate when estimating the arithmetic mean of a sample of numbers (Anderson, 1964; Beach & Swenson, 1966; Levin, 1975; Spencer, 1961, 1963). For example, Spencer's (1963) participants estimated the mean of several sets of either 10 or 20 numbers that varied in terms of their variance and skewness. His overall finding was that "mean errors were remarkably low for all conditions" (p. 256). Beach and Swenson conducted a similar study with similar results, leading them to conclude that "the most important result of this experiment is the high degree of accuracy evidenced in [participants'] estimates" (p. 162).

Recently, however, we have found evidence of a systematic sample size bias in people's central tendency judgments.[1] For example, Price (2001) showed participants descriptions of several fictional employees in terms of their risk factors for having a heart attack and asked the participants to judge the heart attack risk of the average employee. He found that the risk of the average employee was judged to be higher as the company size increased from 5 to 10 employees, then again as the company size increased from 10 to 15 employees. In an extensive set of follow-up studies, Price, Smith, and Lench (2006) found a similar sample size bias when the stimulus people were presented in photographs and participants judged the likelihood that the average group member would experience a wide variety of negative, positive, and even neutral events. In their final study, Price et al. observed the sample size bias when the stimuli were identical stick figures and participants estimated their average height. This result is important, because it casts doubt on two plausible explanations of the sample size bias. One is that people misunderstand their task to be that of judging the likelihood that at least one person in the group will experience the event in question. No such misunderstanding is possible for height judgments. The second is that people attend primarily to extreme (e.g., riskier, taller) individuals, or weight them more heavily in their judgments. Because the groups in this study consisted of identical stick figures, however, there were no extreme individuals.

A. R. Smith, andrew-r-smith@uiowa.edu

Price et al. (2006) suggested that the sample size bias occurs because people automatically encode the sample size and integrate it with their central tendency judgments. Such a general explanation suggests that the sample size bias should be a very general phenomenon. For example, it should be observed when people make intuitive estimates of the mean of a set of numbers, as in the intuitive statistics research. In fact, some of those data are consistent with this hypothesis. Levin (1975, Experiment 3) showed participants groups of numbers that were said to represent the percent price increase for randomly selected items in a store, and their task was to estimate the mean percent price increase. The estimated means for samples of 8, 16, 32, and 64 items were 45.7, 46.2, 48.1, and 49.7, respectively. Although there appears to have been a sample size bias, the stimuli were not constructed, nor the data analyzed, to test this hypothesis specifically.

The primary goal of the present study, therefore, was to systematically test the hypothesis that there is a sample size bias in people's intuitive estimates of the mean of a set of numbers. It is quite possible that we will not observe the sample size bias for this task. Recall that in the work of Price and colleagues (Price, 2001; Price et al., 2006), each stimulus individual's standing on the quantitative dimension of interest had to be estimated or inferred, and the concept of central tendency was generally ill-defined. It may be only under these conditions that the sample size is integrated with central tendency judgments. By contrast, the standing of a numerical stimulus on the dimension of number is straightforward—no inference is necessary—and the concept of central tendency can be defined precisely. Under these conditions, it is possible that participants do not integrate sample size with their central tendency judgments.

EXPERIMENT 1

Method

Participants. Fifty-two participants from the University of Iowa and 85 participants from California State University, Fresno, participated in this study as partial fulfillment of a course requirement.[2]

Stimuli. Each participant estimated the mean of 12 samples of numbers that varied in terms of both their sample size (5, 10, 15, or 20) and mean (20, 30, or 40). To create these samples, we began with a sample of five numbers (9.4, 15.1, 17.2, 26.5, and 31.8) that had a mean of 20. To create additional samples of five numbers with means of 30 and 40, we added 10 and 20 to each of the original five numbers. Then to create samples of 10, 15, and 20, we repeated the numbers in each sample of five either two, three, or four times. Thus, we varied the size and mean of the samples without augmenting the variability or range. To ensure that there was nothing peculiar about the sample of five numbers that we started with, we created three more stimulus sets of 12 samples, based on slightly different initial samples of five numbers.

Design and Procedure. All the instructions and stimuli were presented using a personal computer. Participants were instructed that they would be estimating the arithmetic mean of several samples of numbers. To ensure that participants understood what we meant by arithmetic mean, we provided an example in the instructions and then had them complete four simple problems in which they mentally computed the mean of a small set of numbers (e.g., 5, 10, and 15). Additional instructions indicated that they would have to make intuitive estimates for samples that would be presented too quickly to make a precise calculation. Participants then completed one practice estimation trial under a time limit before proceeding to the 12 regular estimation trials.

Participants were randomly assigned to see one of the four stimulus sets of 12 samples. The samples were presented in a new random order for each participant. Each sample was displayed in the center of the screen in a grid with 5 rows and 4 columns. Samples with fewer than 20 numbers filled the grid starting from the leftmost column. The participants responded by typing their estimates using the number pad of the keyboard. Once they typed in their estimates, they pressed the enter key to see the next sample.

The participants were randomly assigned to one of two time limit conditions. In the constant-time-per-sample condition, they had 5,000 msec per sample to make their estimates. This held the time per sample constant, but allowed the time per number to vary. In the constant-time-per-number condition, they had 500 msec per number in the sample (e.g., the sample of 5 numbers was displayed for 2,500 msec). This held the time per number constant but allowed the time per sample to vary. On each trial in both time limit conditions, a vertical bar appeared to the right of the number grid. When the trial began, the bar started becoming shorter, completely disappearing when the allotted time was up.

Results and Discussion

Three participants were dropped from the analyses because their responses indicated they either misunderstood their task or were not attempting to provide accurate estimates. We conducted a repeated measures ANOVA with sample size and objective mean as within-subjects factors and stimulus set and timing condition as between-subjects factors. Focusing on the linear contrasts, this analysis confirmed that there was a main effect of sample size [$F(1,126) = 6.33$, $p = .013$, $\eta_p^2 = .05$]. As Figure 1 shows, participants' estimates tended to increase as the size of the sample increased. The participants were sensitive to the objective mean of the samples, as evidenced by the significant main effect of objective mean [$F(1,126) = 61.35$, $p < .001$, $\eta_p^2 = .33$]. There was no main effect of stimulus set [$F(3,126) = 0.30$, $p = .82$, $\eta_p^2 = .007$]. There was a marginally significant main effect of timing condition [$F(1,126) = 2.87$, $p = .09$, $\eta_p^2 = .02$], indicating that participants in the constant-time-per-number condition gave higher estimates than participants in the constant-

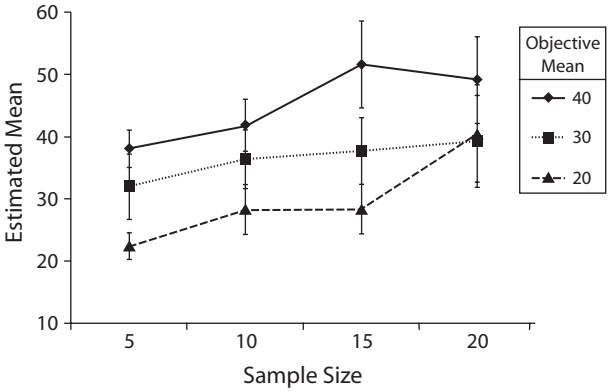

Figure 1. Mean estimates for all participants in Experiment 1 as a function of sample size and objective mean. Error bars represent ±1 standard error.

time-per-sample condition. There were no significant interactions.

Overall, the accuracy of the participants' estimates was fairly poor. The mean absolute deviation (MAD) for all judgments was 17.60 ($SD = 44.76$). This raises the concern that the sample size bias might be driven by a relatively small number of participants who gave particularly inaccurate estimates. To rule out this explanation, we first identified extreme responses that were smaller than 10 or larger than 100. Nearly all stimulus numbers were within this range, so mean estimates falling outside of it were assumed to be the result of typing errors, a misunderstanding of the task, or a lack of motivation to provide accurate responses. In all, 210 of the 1,644 estimates (12.77%) were classified as extreme. We then focused a second analysis on those participants who did not give any extreme responses ($n = 83$). Not surprisingly, the MAD of the remaining participants was much lower and less variable ($M = 4.84$, $SD = 2.21$; see Figure 2). Among these more accurate participants, there was still a significant effect of both sample size [$F(1,75) = 10.26$, $p = .002$, $\eta_p^2 = .12$] and objective mean [$F(1,75) = 773.12$, $p < .001$, $\eta_p^2 = .91$]. There were no other significant main effects or interactions.

A possible explanation for the sample size bias is that people attend to, and base their estimates on, a subset of the largest numbers in the sample. For example, the mean of the five largest numbers in each sample does, in fact, increase as the sample size increases. Not only would this produce a sample size bias, it should also result in a tendency toward overestimation. Although there was a tendency toward overestimation when participants who gave extreme responses remained in the analysis, the tendency reversed when these participants were eliminated. Among the most accurate participants, therefore, there was both a sample size bias and a tendency toward underestimation; the mean signed deviation of their estimates was significantly less than zero ($M = -1.73$, $SD = 2.88$) [$t(82) = 5.47$, $p < .001$]. This argues against the idea that the sample size bias occurs because people focus on the largest numbers.

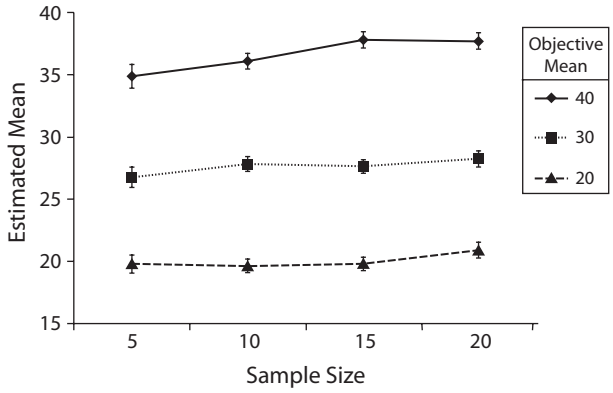

Figure 2. Mean estimates for accurate responders in Experiment 1 as a function of sample size and objective mean. Error bars represent ±1 standard error.

These results revealed a sample size bias for intuitive estimates of the arithmetic mean of groups of numbers. This effect cannot be attributed to the variability of the numbers, a small number of participants who gave inaccurate responses, or a tendency to focus on a subset of the largest numbers in each sample. It could be attributed, however, to the envelope area of the numbers—the area of the smallest polygon that contains the numbers—because larger samples had larger envelope areas.

EXPERIMENT 2

In Experiment 2, we addressed the issue of envelope area as a confounding variable in a way that emphasizes the generality of the sample size bias. Specifically, we displayed the numbers in each sample sequentially rather than simultaneously.

Method

Participants. One hundred twenty undergraduate psychology students at California State University, Fresno, participated as partial fulfillment of a course requirement.

Design and Procedure. Aside from presenting the numbers sequentially rather than simultaneously, the design and procedures were essentially the same as in Experiment 1. After reading instructions about their task, the participants completed the same simple computational problems and practice estimation trial used in Experiment 1. They then proceeded to the 12 regular estimation trials.

Each estimation trial began with a white plus sign presented in the middle of a blue background. This served as a fixation point. When the participant pressed the enter key, the numbers in the sample appeared at the fixation point, with a brief intertrial interval (ITI). After all the numbers in the sample were presented, the participant was prompted to enter his or her estimate of the mean, at which point the next trial began.

Participants were randomly assigned to one of two timing conditions. In the constant-time-per-number condition, each number was presented for 1,000 msec, with an ITI of 500 msec. This held the time per number constant. In the constant-time-per-sample condition, each number was presented for 1,000 msec but the ITI was varied. This held the time it took to present a sample constant at 22.5 sec.

Results

We submitted the estimates of all participants to a repeated measures ANOVA, as described in Experiment 1. Again, there was a significant linear effect of sample size [$F(1,112) = 10.58$, $p = .002$, $\eta_p^2 = .09$]. Participants' estimates tended to increase as the size of the sample increased (see Figure 3). There was also a significant effect of objective mean [$F(1,112) = 51.59$, $p < .001$, $\eta_p^2 = .32$]. There were no other significant main effects or interactions. As in Experiment 1, the MAD was quite high ($M = 18.08$, $SD = 47.13$), because the full data set included several participants who made estimates lower than 10 and greater than 100. After excluding 34 participants who made at least one extreme estimate, the MAD of the remaining participants was much lower and less variable ($M = 5.09$, $SD = 4.00$; see Figure 4). Among these more accurate participants, there was still a significant effect of both objective mean [$F(1,78) = 1,034.12$, $p < .001$, $\eta_p^2 = .93$] and sample size [$F(1,78) = 6.78$, $p = .01$, $\eta_p^2 = .08$]. There were no other significant main effects or interactions.

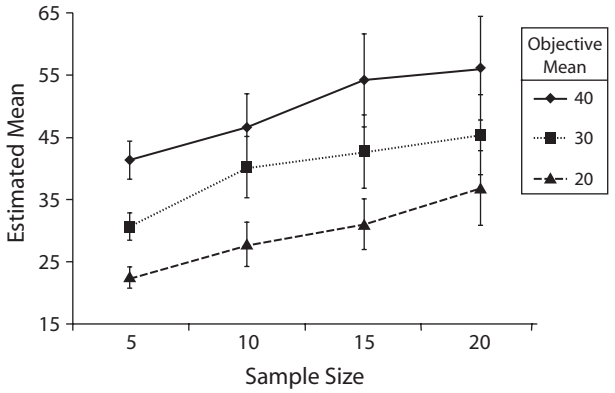

Figure 3. Mean estimates for all participants in Experiment 2 as a function of sample size and objective mean. Error bars represent ±1 standard error.

As in Experiment 1, there was a marginally significant tendency to underestimate the sample mean among the more accurate participants; the mean signed deviation of mean estimates was −0.88 (SD = 4.36) [$t(85)$ = 1.88, p = .06]. Therefore, it is unlikely that the sample size bias occurs because people focus on the largest numbers in each sample, since this would result in a tendency toward overestimation.

GENERAL DISCUSSION

In two experiments, we observed our hypothesized effect of sample size on people's intuitive estimates of the means of samples of numbers. These experiments were designed to control several potential confounding variables, including the variability of the numbers and their spatial and temporal distribution. We were also able to demonstrate that the sample size bias occurs for people who make relatively accurate responses and that the bias cannot be explained by a tendency to focus on a subset of the largest numbers in each sample.

We believe that any plausible theory of the sample size bias must take into account the generality of the effect. We have observed it for judgments about heart attack risk based on written profiles (Price, 2001), risk and likelihood judgments based on group photographs (Price et al., 2006), estimates of the heights of identical stick figures (Price et al., 2006), and now for estimates of the means of samples of numbers presented both simultaneously and sequentially.

One explanation that might account for all of these results—a variation on the idea that people automatically integrate sample size into their central tendency judgments (Price et al., 2006)—is based on the concept of *magnitude priming* (Oppenheimer, LeBoeuf, & Brewer, 2008). The idea is that presenting people with a small or large quantity can activate a modality-independent representation of magnitude, which is then assimilated into other quantitative judgments. For example, Oppenheimer et al. showed that people who first drew long lines tended to provide higher estimates for the length of the Mississippi River than did people who first drew short lines. From this perspective, the sample size bias might occur because the sample size activates a representation of a larger or smaller magnitude, which is then assimilated into people's central tendency judgments.

Of course, this explanation requires that sample size activate a magnitude representation and that it do so regardless of whether the individual stimulus elements are presented simultaneously or sequentially. Although we have provided no direct evidence for this assumption here, we should note that event frequency does appear to be encoded fairly automatically (Hasher & Zacks, 1979). People make reasonably accurate frequency estimates for stimuli that were presented sequentially and for which they had no conscious intention to encode frequency information (Naveh-Benjamin & Jonides, 1986). Furthermore, research on numerosity perception generally supports the idea that people form representations independent of the sensory modality in which the stimuli are presented and whether or not the presentation of individual stimulus elements is simultaneous or sequential (Barth, Kanwisher, & Spelke, 2003). So it seems likely that, if drawing a long or short line can activate a corresponding magnitude representation, being presented with a small or large sample can do the same. Consistent with this idea is other research showing an effect of irrelevant numerosity information on quantitative judgments (Friedenberg & Limratana, 2005; Pelham, Sumarta, & Myaskovsky, 1994).

The magnitude priming explanation is precise enough to suggest several additional hypotheses about the sample size bias. One is that if sample size results in a modality-independent magnitude representation, it should affect other kinds of judgments and responses. Consider that Oppenheimer et al. (2008) showed that drawing a long or short line affected not only people's estimates of the length of the Mississippi River, but also their behavior in a word completion task. For example, those who drew a long line were more likely to complete the fragment *_all* to create the word *tall*. Would this effect be observed for people who had just been presented with samples of various sizes? A second hypothesis is that what matters is the

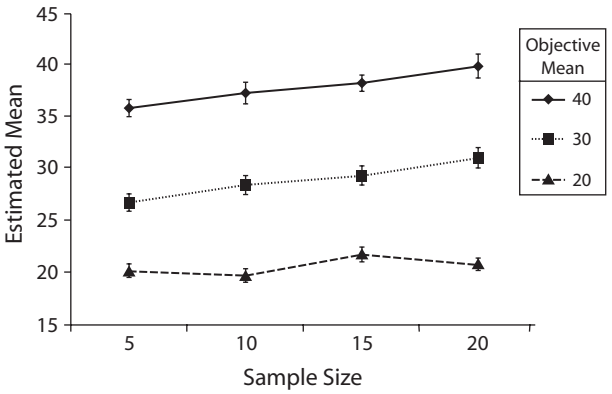

Figure 4. Mean estimates for accurate responders in Experiment 2 as a function of sample size and objective mean. Error bars represent ±1 standard error.

relative sample size rather than the absolute sample size. As Oppenheimer et al. pointed out, the fact that drawing a line on a piece of paper can affect estimates of the length of the Mississippi River suggests a unit-free representation of magnitude that must necessarily be context dependent. In our paradigm, a sample of 10 might elicit very different estimates when presented with samples of 6 and 8 than when presented with samples of 20 and 30.

Regardless, the sample size bias appears to be quite general and has implications for understanding judgments in various contexts. We have already noted that the sample size bias can contribute to the magnitude of comparative optimism—people's tendency to judge themselves to be at lower risk than their peers for negative events (Price, 2001; Price et al., 2006). This is because judgments about oneself are judgments about a very small sample and judgments about one's peers are judgments about a very large sample. Similarly, it might contribute to high school and college students' tendency to overestimate the extent to which their peers use drugs and practice unsafe sex (e.g., Page, Hammermeister, & Scanlan, 2000). Again, in this research, participants are generally asked to make judgments about themselves (a small sample) and their typical or average peer (a large sample).

AUTHOR NOTE

Correspondence concerning this article should be addressed to A. R. Smith, Department of Psychology, University of Iowa, Iowa City, IA 52242 (e-mail: andrew-r-smith@uiowa.edu).

REFERENCES

ANDERSON, N. H. (1964). Test of a model for number-averaging behavior. *Psychonomic Science*, **1**, 191-192.

BARTH, H., KANWISHER, N., & SPELKE, E. (2003). The construction of large number representations in adults. *Cognition*, **86**, 201-221. doi:10.1016/S0010-0277(02)00178-6

BEACH, L. R., & SWENSON, R. G. (1966). Intuitive estimation of means. *Psychonomic Science*, **5**, 161-162.

FRIEDENBERG, J., & LIMRATANA, W. (2005). Hierarchical number estimation. *Psychological Research*, **69**, 211-220. doi:10.1007/s00426-003-0169-y

HASHER, L., & ZACKS, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, **108**, 356-388. doi:10.1037/0096-3445.108.3.356

LEVIN, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General*, **104**, 39-53. doi:10.1037/0096-3445.104.1.39

NAVEH-BENJAMIN, M., & JONIDES, J. (1986). On the automaticity of frequency coding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 378-386. doi:10.1037/0278-7393.12.3.378

OPPENHEIMER, D. M., LEBOEUF, R. A., & BREWER, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, **106**, 13-26. doi:10.1016/j.cognition.2006.12.008

PAGE, R. M., HAMMERMEISTER, J. J., & SCANLAN, A. (2000). Everybody's not doing it: Misperceptions of college students' sexual activity. *American Journal of Health Behavior*, **24**, 387-394.

PELHAM, B. W., SUMARTA, T. T., & MYASKOVSKY, L. (1994). The easy path from many to much: The numerosity heuristic. *Cognitive Psychology*, **26**, 103-133. doi:10.1006/cogp.1994.1004

PETERSON, C. R., & BEACH, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, **68**, 29-46.

PRICE, P. C. (2001). A group size effect on personal risk judgments: Implications for unrealistic optimism. *Memory & Cognition*, **29**, 578-586.

PRICE, P. C., SMITH, A. R., & LENCH, H. C. (2006). The effect of target group size on risk judgments and comparative optimism: The more, the riskier. *Journal of Personality & Social Psychology*, **90**, 382-398. doi:10.1037/0022-3514.90.3.382

SPENCER, J. (1961). Estimating averages. *Ergonomics*, **4**, 317-328.

SPENCER, J. (1963). A further study of estimating averages. *Ergonomics*, **6**, 255-265.

NOTES

1. Elsewhere, we have referred to this phenomenon as the *group size effect*. We now believe that sample size bias is more descriptive and emphasizes its generality.

2. Estimates given by participants from the University of Iowa and California State University, Fresno, did not differ in any substantial way. Most importantly, university affiliation did not interact with sample size.

(Manuscript received April 1, 2009;
revision accepted for publication January 12, 2010.)