SAMPLE SIZE WEIGHTING IN PROBABILISTIC INFERENCE

by

NATALIE ANN LINDEMANN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Gretchen B. Chapman & Rochel Gelman

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2010

ABSTRACT OF THE DISSERTATION

SAMPLE SIZE WEIGHTING IN PROBABILISTIC INFERENCE

By NATALIE ANN LINDEMANN

Dissertation Directors:

Rochel Gelman and Gretchen B. Chapman

How do people evaluate data on the basis of sample size? Normatively, sample size is an important factor that one should consider when making judgments and inferences from sample data. Previous research is mixed regarding whether or not laypeople are sensitive to sample size. However, in this paper I show that laypeople attend to sample size, but that their sensitivity decreases as sample sizes become larger. This curvilinear functional form is found for both high and low numerate subjects across two different judgment tasks. However, high numerate subjects consistently show greater sensitivity to sample sizes than lower numerates, although they still underweight sample size relative to normative standards. Low numerate subjects' sensitivity to sample size may be increased by providing raw data and instructions that sample size matters.

ACKNOWLEDGEMENT AND DEDICATION

I would like to thank and acknowledge my two advisors, Gretchen Chapman and Rochel Gelman for their guidance and support over the past five years. They have selflessly given their time to discuss research ideas, edit manuscripts, suggest data analyses, and provide career counseling. I hope that I can be as supportive to my future students as they have been to me.

I would also like to acknowledge my fellow graduate students in the Chapman and Gelman labs. They have been an incredible source of encouragement. Also, I thank Manish Singh, Jacob Feldman, and Lance Rips for serving on my committees. Their thoughtful questions and comments have shaped how I think about my research. Additionally, I am grateful to the staff within the Psychology Department and Center for Cognitive Science for their assistance during my time at Rutgers.

Finally, I would like to thank my husband Christopher for his support. He willingly moved halfway across the United States with me so that I could come to Rutgers for graduate school. Also, he has accepted the hassles that come with being married to a graduate student; these include me working odd work hours, being away for conferences, and introducing a lot of uncertainty into our lives in terms of where we will end up living. This dissertation is dedicated to Chris.

TABLE OF CONTENTS

SAMPLE SIZE WEIGHTING IN PROBABILISTIC INFERENCE

INTRODUCTION

Psychologists have long been interested in whether or not laypeople have intuitions about the role that sample size should play in their judgments and decisions. Everyday, people make inferences on the basis of information. Normatively one should be more confident that a sample provides a good estimate of a population parameter when that sample is based on a larger, rather than a smaller, number of observations. For example, one might feel more confident purchasing a car that has been highly recommended by 1000, compared to only 10, people.

The idea that larger samples are better than smaller ones is referred to as the law of large numbers (Bernoulli, 1713). Although the law the large numbers is highly familiar to most psychologists, some research has shown that laypeople's judgments do not reflect this concept (Kahneman & Tversky, 1973, Pitz, 1967). Indeed, Tversky and Kahneman even propose that humans have a "Belief in the law of small numbers" (1971), referring to the idea that laypeople think that even small samples should be highly representative of the population from which they are drawn.

In contrast to the heuristics and biases research (Tversky & Kahneman, 1974), a number of studies have shown that humans *do* consider sample size when making inferences (Irwin, Smith, & Mayfield, 1956, Kaufmann & Betsch, 2009, Nisbett, Krantz, Jepson, & Kunda, 1983, Obrecht, Chapman, & Suárez, 2010). Some previous failures to demonstrate sample size intuition can be at least partially accounted for by problem complexity (Evans & Dusior, 1977) and distribution type (i.e. sampling vs. frequency distributions, Sedlmeier & Gigerenzer, 1997, Sedlmeier, 1998).

The majority of the studies that have examined lay intuitions about sample size have focused on whether subjects use or fail to use the law of large numbers. However, in this paper I will go beyond this dichotomy and argue that 1) sample size intuitions follow a curvilinear functional form, 2) individual differences in numeracy and the magnitude of the numbers being considered affect sample size sensitivity (the steepness of the curvilinear slope), and that 3) providing raw data and instructions can improve use of sample size, especially for subjects who are lower in numerical ability.

OVERVIEW OF PAST RESEARCH

About 50 years ago, the view of psychologists was fairly optimistic regarding lay use of statistical factors, such as sample size. Irwin, Smith, and Mayfield (1956) showed subjects samples of cards drawn from a large deck that was said to have been shuffled so to be random (but it was not). Each card displayed a number that was either positive or negative in value; they were displayed sequentially to subjects. Participants judged whether they thought that the average value of the entire deck of cards was greater or less than zero, and how confident they were in their judgment. Subjects were more confident in their judgments when they were shown 20, rather than 10, samples of cards from the deck. This sensitivity to sample size was replicated in a second experiment in which subjects considered cards from two separate decks, and judged which deck had the higher mean. Again, Irwin et al. showed that subjects were sensitive to sample size in the normative direction, as well as to the other statistical factors that were manipulated.

Kahneman and Tversky (1972) paint a quite different view of lay intuitions of sample size. They showed that laypeople fail to incorporate sample size into their representations of sampling distributions. For example, they asked subjects which of two

hospitals will have more days in a year in which over 60% of babies born are male. Subjects read that in a smaller hospital about 15 babies are born each day, while in a larger hospital about 45 babies are born each day. The majority of subjects said that the two hospitals will have about the same number of days in which more than 60% of births are male. The normatively correct answer is that the smaller hospital can expect to have more days than the larger hospital where the percentage of male births diverges from the expected value of 50%. Evan and Dusior (1977) and Sedlmeier (1998), however, show that more subjects answer correctly once the problem is reworded to be simpler. For example, a larger percentage of subjects are able to recognize that on a single day, the small hospital is more likely than the larger to have a male birth rate of 60%.

A number of studies have examined how humans use sample size to make inferences on the basis of data. Nisbett et al. (1983) gave adult subjects sample data and asked them to make inferences about their population characteristics. The sample size of the data was manipulated between subjects to be 1, 3, or 20. For example, one group of subjects read about a sample of 3 Barrotos people, each of whom was obese. From this sample, subjects inferred the percent of the whole population that they thought shared the sample characteristic (e.g. what percent of all Barrotos people are obese). Subjects' inferences depended on both the sample size and the implied variability of the category. When inferring the percent of the Barrotos population who are obese, subjects were relatively conservative in their estimates when they had read about 1 or 3 sample individuals being obese, but gave higher percent estimates after reading about a sample of 20 individuals, all of whom were obese. When subjects read about samples of ludium, a fictitious element, that conducts electricity, they readily inferred that almost the whole

population also conducts electricity, even when only given samples of size 1 or 3. This makes sense because we know that different samples of the same element kind should share exactly the same properties, that is, they should not vary among one another, while body weight varies within a group. From this work, it appears that laypeople are not only are sensitive to sample size, but also jointly able to incorporate their variability knowledge into their inferences (also see Obrecht et al, 2010).

Following the Nisbett et al. (1983) finding, developmental psychologists have examined how humans use sample size to make inferences at different ages. Jacobs and Narloch (2001) asked children and adult subjects to make inferences about populations on the basis of sample data. Like the Nisbett et. al. study, their sample data were always identical in regard to the characteristic of interest. For example, in one condition subjects were told that in a sample of 3 children at a school, all 3 were wearing green shirts. From this information subjects inferred what percent of the whole population, i.e. the school, shared this characteristic, i.e. wearing a green shirt (personal communication with Narloch, April 24, 2009). They varied between subjects whether participants were given data about samples of size 1, 3, or 30, and also whether the domain in which the data were described implied high verses low variability. Regardless of their age, when the sample data described outcomes in the low variability domain of biological characteristics (e.g. number of eyes that an animal kind has), subjects gave low population estimates based on samples of size 1, but high estimates for samples of size 3 or 30. However, when variability was assumed to be high in the behavioral domain, as in the tee-shirt example, subjects treated samples of 1 and 3 the same, but gave higher percentage estimates for samples of size 30. From this it seems that when variability is

very low, one only needs a few examples to draw a strong generalization to the population, but a sample of size 1 is insufficient. When variability is high, larger samples, in this case, of size 30, are better.

In a related developmental study, Masnick and Morris (2008) asked adult and children subjects to compare pairs of datasets to decide if they differ. The sample data used by Masnick and Morris had variability, that is, the data within a sample were not identical as they were in the previous Nisbett et al. (1983) and Jacobs and Narloch (2001) studies. In their task, the data given to subjects were said to come from two balls that were thrown or hit some distance. Based on the distances for Ball A vs. Ball B, subjects drew a conclusion regarding whether there was a difference between the two balls (or the person throwing them), and how confident they were in that choice. Sample size was manipulated to be 1, 2, 4, or 6 within a dataset pairing. For example, in a given comparison, a subject might compare the 4 recorded distances that Baseball A was thrown, and the 4 recorded distances that Baseball B was thrown. Adult subjects' confidence in a difference between the two balls' distances went up sharply as sample size increased; however, sample size had only a small effect on 6th graders' confidence, and 3rd graders showed a nearly significant effect of sample size in the counter-normative direction. In a second experiment in which data were presented sequentially, adult subjects again showed a significant increase in confidence as sample size increased, but the two groups of children did not. Despite that this comparison task used by Masnick and Morris was probably more difficult (because of the data variability and pairwise comparisons) than that used by Nisbett et al. or Jacobs and Narloch (2001), adult subjects still showed sensitivity to sample size information.

Obrecht et al. (2007) also asked subjects to make inferences on the basis of paired datasets. In each comparison pair, subjects viewed consumer rating data from two fictitious products and judged whether the product with the higher mean rating was actually better than the product with the lower mean rating. Obrecht et al. varied sample sizes across comparison pairs to be either 10 or 37, and also varied the difference between group means, and the within-product variability to have two levels. They selected the two levels of the three statistical factors (sample size, mean difference, and standard deviation) so that they were equated in terms of their effect on statistical power. That is, if subjects were statisticians performing a between subjects *t*-test on the sample data given for two products, their probability of finding a difference, given that it existed, would be equally affected by a change in sample size, mean difference, or standard deviation level. Obrecht et al. found that subjects' confidence ratings in a difference between products were affected by all three statistical factors. However, participants primarily focused on the difference between product means, and gave much less attention to the sample size or within group variability, despite the fact that these factors should have equally affected their confidence.

Obrecht et al. (2010) used a similar comparison task to Obrecht et al. (2007) in which subjects compared groups on the basis of sample data and decided whether their populations differed. However, they used samples size levels of 2 and 10 instead of 10 and 37. In these studies, the effect of sample size was much larger, actually larger than the effect of mean difference. One possible explanation of this difference in effect size is that laypeople's sensitivity to sample size is dependent on the specific numerical values provided. This suggests that laypeople weigh sample size in a negatively accelerating

functional form where differences between smaller values (2 vs. 10) are perceived to be greater than those between of larger values (10 vs. 37). Another explanation is that sample size sensitivity could be related to the ratio of the values being considered (e.g. 1 vs. 5 compared to 1 to 3.7). Notably, the Nisbett et al. (1983) and Masnick and Morris (2008) studies both employed some sample sizes of less than 10. Thus, lay people may be sensitive to differences among small sample sizes but less sensitive to differences among large sample sizes. I further discuss this idea in the Hypotheses section below.

Other indirect evidence for how people value sample size comes from work in the decisions from experience literature. The majority of this work has been done using choices between gambles where subjects experience data from two populations and then choose from which population they would like to draw an outcome. For example, one might choose between a deck of cards with a 100% chance of winning $3 or a deck with an 80% chance of winning $4 and a 20% chance of winning nothing. In order to discover the payout structure of each deck, subjects are allowed to sample, one by one, cards from the two decks until they feel comfortable that they can make a choice between them.

These experiments have been used to compare how subjects choose between gambles when they either learn about the payout structure by sampling from the population (the experience condition), verses when they are simply told what the payouts and probabilities are (the description condition). Subjects choose differently depending on which method of obtaining information is employed (Hertwig, Barron, Weber, & Erev, 2004, Hau, Pleskac, Kiefer, and Hertig, 2008, Gottlieb, Weiss, Chapman, 2007).

However, another interesting aspect of these studies is how many cards subjects choose to sample in the experience condition. Participants are allowed to view as many

cards as they would like. How many samples do people feel they need in order to accurately compare two groups? Of course this will partly depend on how much variability subjects assume the card populations have. However, in these studies the only cost to sampling was the time it took to click a button to flip over a virtual card. Nevertheless, when people are free to sample data from two populations in order to infer which offers the better gamble, they choose relatively small samples. Hertwig et al.'s (2004) subjects only sampled a median of 15 items in total across two populations (e.g. a sample of 7 outcomes from one population and a sample of 8 from the other). Hau et al.'s (2008) subjects sampled on average 11 cards in total (e.g. 5 samples from one deck and 6 from the other); when given a greater payout scheme, subjects increased sampling to be about 33 cards in total. This suggests that people feel that the amount of data they will gain from subsequent samples rapidly diminishes such that further information is of little value. This is consistent with Tversky and Kahneman's (1971) point that people assume that samples are highly representative of their respective population.

HYPOTHESES

Research on sample size intuition has largely focused on whether or not people use sample size in the normative direction when making inferences. However, here I will take this inquiry further by describing the functional form of sample size weighting. This is an important advance because it will help to reconcile discrepant past findings while also providing insight regarding how, and to what extent, the presentation of statistical information affects judgment.

Looking across previous work, it appears that sample size sensitivity might be greater when smaller sample size values are used (Masnick and Morris, 2008, Obrecht et

al, 2010), compared to when larger values are being compared (Obrecht et al., 2007). This pattern could suggest a nonlinear diminishing sensitivity function such that judgments made as a function of sample size may be well fit by a negatively accelerating curve (e.g. power function with an exponent between 0 and 1, or a logarithmic function), rather than a linear weighting function. In the current paper, the shape of this functional form will be examined across four experiments that employ two quite different tasks.

Furthermore, subjects' numeracy levels will be examined in relation to their sample size judgments. A number of studies have shown that individual differences in numeracy (Lipkus, Samsa, and Rimer, 2001) relate to peoples' decisions (e.g. Peters, Vastfjall, Slovic, Mertz, Mazzocco, & Dickert, 2006). Peters, Slovic, Vastfjall, and Martz (2008) argue that higher numerate subjects, compared to those lower in numeracy, have more precise mental numerical representations that relate to the numerical choices that they make. If higher numerate subjects do indeed have more precise numerical representations, then we can predict that they should be more sensitive to differences between sample sizes compared to less numerate subjects; their representations of numerical values should overlap less, making different sample sizes feel more different compared to those with less precise representations. Thus, high numerate subjects' sensitivity to sample size should be best fit by a function with a steeper slope, compared to subjects who score lower in numeracy. Also, if low numerate subjects lack a precise mapping between the magnitude of a numerical quantity and an Arabic numeral that it is represented by (Peters et al, 2008), then they should benefit from presentation formats that make clearer these quantities.

OVERVIEW OF THE CURRENT STUDIES

In Experiments 1 and 2 (using a paradigm similar to Obrecht et al., 2007 and 2010), subjects compared pairs of hypothetical consumer products and judged their confidence in a difference. For each product within a comparison pair they were provided average consumer ratings, the number of raters (sample size), and the within-group standard deviation of those ratings. Sample size was varied to have 10 levels. In Experiment 1, confidence ratings as a function of sample size showed the expected curvilinear functional form, and higher numerate subjects showed greater sensitivity than lower numerate subjects. Also, lower numerate subjects benefited from enhanced presentations that highlight sample size by showing individual rating data.

In Experiment 2 half of the subjects were explicitly told that sample size matters and should be incorporated into their judgments. The curvilinear weighting function pattern was replicated and the instruction increased the use of sample size for low numerate subjects.

When making inferences from samples to populations, the amount of information gained increases by the square root of N (sample size). This value ($N^{.5}$) was compared as a normative standard to subjects' judgments in Experiments 1 and 2.

In Experiments 3 and 4, subjects were given a different task. They viewed percentage data from multiple sources and were asked to combine these to make a judgment regarding the chances of an event occurring. Each percentage that was provided in a set was paired with a sample size. It is clear from subjects' combined estimates that they do not weight the percentage data using a weighted linear average as they should, but instead appear to weigh percentage data using a curvilinear weighting

function for sample size. Furthermore, the slope of the weighting function decreases when the magnitude of the sample sizes increases (Experiment 4).

EXPERIMENT 1

The purpose of Experiment 1 was to examine the weighting function of sample size in a task where participants made pairwise comparisons. Subjects compared pairs of products based on their rating data. They judged how confident they were that the product with the higher mean rating was actually better than its comparison. That is, subjects were asked to make an intuitive *t*-test inference, similar to the procedure in Obrecht et al. (2007).

METHOD

Undergraduate subjects taking introductory psychology at a large, diverse university (N=104) participated for course credit. In all subsequent experiments, subjects were drawn from similar undergraduate psychology subject pools. All materials were presented online.

Design

Within subjects factors

Subjects were asked to compare pairs of fictitious products on the basis of consumer ratings. For each product within a comparison, they were told how many consumers rated the product (i.e. the sample size), the mean rating of the product, and the standard deviation of the product's ratings. Sample size, difference between paired product means, and within-product standard deviation were varied within subjects. Sample size was manipulated to have 10 levels (i.e. 1, 2, 5, 8, 10, 13, 16, 20, 27, 37). The mean difference in a comparison pair was varied to be high or low; specifically, either

M1=8 and M2=7 (on a 1 to 15 rating scale) for a difference of 1 or M1=9 and M2=7 for a difference of 2. Within-product standard deviations were either high or low (i.e. SD=2.83 or SD=1.41). Figure 1 shows an example comparison pair where sample size was 8, mean difference was high (i.e. 2), and standard deviation was low (i.e. 1.41).

Within a comparison pair, the sample size of the two products was always equal (e.g. both products were rated by 10 consumers), as were the standard deviations (e.g. the SD of the ratings for Product A was 1.41 and the SD of the ratings given to Product B was also 1.41). Thus, these factors, sample size and standard deviation, were varied between the different pairs of products. In total, every subject compared 38 product pairs. This was the result of crossing the 9 levels of sample size that ranged from 2 to 37 with the 2 levels of mean difference and two 2 levels of standard deviation. When sample size equaled 1 only mean difference could be manipulated, not standard deviation. These combinations yielded a $9 \times 2 \times 2 + 2$ design.

Between subjects factors

Between subjects I manipulated data type and order (2×2). Data type refers to whether subjects were given just the three summary statistics (i.e. sample size, mean, and standard deviation) or the summary statistics and, additionally, the corresponding raw rating data. The top of Figure 1 above the dotted line shows what subjects viewed in the statistics-only condition. The full figure, excluding the dotted line, shows the information given to those in the statistics+data condition. The presentation order of the product pairs was varied so that about half of subjects viewed the 38 pairs in one order, while the other half received the information in an alternative order.

Introductory Materials

Subjects were told that they would be shown information summarizing the ratings that consumers gave to products. Each product could be given a rating from 1 to 15 where 15 was the best possible rating. They read that they would be given three pieces of information about each product: the number of people who rated it, the average consumer rating, and the standard deviation of those ratings. These correspond, respectively, to sample size, mean rating, and standard deviation. Participants were given brief explanations of each of the statistical concepts, along with simple examples using hypothetical products and their consumer rating data. Subjects were told that because six people rated Product Z, the number of raters would be six. They were also shown how the mean consumer rating for Product Z is calculated using the six ratings that the product had received. Finally, participants were shown how the ratings given to Product Z differed a lot and had a standard deviation of 3.9. In contrast, they saw how the ratings given to another fictitious product, Product Y, were identical and thus had a standard deviation of zero. These two examples were used to explain the idea of standard deviation where larger values indicate more disagreement among raters.

Next, subjects were shown how the three descriptive statistics could be displayed on number lines. The *Number of Raters* line ranged from 0 to 40 with labels of *no raters* and *many raters* at either end. The *Average Rating* line went from 1 to 15 with labels of *lowest rating* and *highest rating*, respectively. Finally, the *Standard Deviation of Ratings* line ranged 0 to 4; the end points were labeled *high rater agreement* and *low rater agreement* (see Figure 1). Subjects were given check questions about each of the statistical concepts. For example, they had to indicate which product had a standard

deviation value that indicated dissimilar opinions among consumers. They were allowed to proceed in the experiment only once they answered all 6 check questions correctly.

Experimental Materials

Following the introduction and check questions, subjects viewed the 38 pairs of products, each presented on a separate webpage. For each pair, they were asked to rate on a 9-point scale how confident they were that the product with the higher mean was really better than its comparison. The 9-point dependent measure was labeled from *Extremely Unconfident* to *Extremely Confident* at its endpoints.

Numeracy measure

A multiple choice version of the Lipkus, et al. (2001) numeracy scale was used to assess subjects' numerical literacy (see Appendix I). The scale was one of many individual difference measures provided in a prescreening battery given to all research participants the psychology subject pool. Subjects' numeracy scores were simply the number of questions correctly answered out of 10.

RESULTS

As expected, subjects' sensitivity to sample size showed a curvilinear pattern (see Figure 2) [1]. Power and logarithmic regression models were tested and both showed reasonable fits and significant effects of sample size (although the log likelihood associated with the log regression model did 51 times better than the power model). A linear model was tested, but unsurprisingly, failed to provide as good of a fit to the data

[1] To confirm that this pattern was not the result of some subjects showing normative sensitivity and others showing no sensitivity, I conducted a simple analysis to see what percent of subjects showed at least some use of sample size. I found each subjects' average confidence for each level of sample size, collapsing across mean difference and standard deviation. I then calculated the percentage of subjects whose confidence was higher for the 5 largest sample sizes, compared to the 5 lowest sample sizes. 77% of subjects fell into this category. The majority of subjects who did not show higher confidence with the larger sample sizes scored low on the numeracy measure.

when compared to the nonlinear models (e.g. the data were 175 times in favor of the power model over the linear model)[2].

Just the power model ($a + N^b$) will be further discussed because it offers a clear normative standard to which subjects' judgments can be compared. From a statistical perspective, confidence should increase by the square root of sample size. This translates into a power function with an exponent of .5 (i.e. $N^{.5}$).

Mixed model nonlinear regression (Proc nlmixed in SAS) was used to test how sample size and other factors affected subjects' confidence judgments. Sample size (10 levels, from 1 to 37), mean difference (1 vs. 2), and standard deviation (2.83 vs. 1.41) were within subjects factors. Numeracy (0 to 10), data type (statistics-only vs. statistics+data), and presentation order (order 1 vs. order 2) were between subjects variables. The levels of each factor, except sample size and numeracy, were coded as -.5 and .5. Actual sample size values were used, while numeracy scores were transformed into z-scores.

The nonlinear mixed regression model was used to find the best fitting power coefficient given subjects' data. To start, a simple model was used to test just the main effects of the sample size coefficient, mean difference, standard deviation, data type, numeracy, and order:

$$confidence = b_0 + N^{b_1} + b_2(md) + b_3(sd) + b_4(data\_type) + b_5(numeracy) + b_6(order)$$

---

[2] Power regression model: $a + N^b$, log regression model: $a + \ln(N)$, linear regression model: $a + bN$. The log likelihood of log model fit minus the log likelihood of the power model fit was 13841-13892=-51; lower log likelihoods indicate a better fit. The log likelihood of the power model fit minus the log likelihood of the linear model fit was 13892-14067=-175. All subsequent log likelihoods comparisons in this paper were calculated in the same fashion.

This model yielded a significant (different from zero) power coefficient for sample size (β = .31, t (103) = 48.59, p<.0001). Thus, subjects' confidence in a difference between two products increased as sample size became larger, but at a shallower rate than should be normatively expected, that is, the 95% C.I. of .30-.32 does not include .5.

Mean difference (β = .52, t (103) = 12.56, p<.0001), and standard deviation (β = .30, t (103) = 7.24, p<.0001) also affected confidence ratings in the normative direction. Subjects were more confident in a difference when the difference between product means was large, rather than small, and when the standard deviation of ratings was low, rather than high. No main effects of numeracy, data type, or order were found.

Next, the nonsignificant main effect of order was removed from model, and all possible interactions among sample size, data type and numeracy were entered. Sample size and numeracy interacted (β = .47, t (103) = 11.32, p<.0001). Subjects higher in numeracy showed greater sensitivity to sample size. Sample size and data type also interacted (β = .34, t (103) = 4.69, p<.0001) showing that participants were more sensitive to sample size when they were provided with raw data in addition to the summary statistics. However, these effects were qualified by a 3-way interaction among sample size, numeracy, and data type (β = -.20, t (103) = -2.58, p=.0114). Lower numerate subjects were more sensitive to sample size when provided with raw data compared to when only given summary statistics. For subjects higher in numeracy, the addition of raw data had little effect on sample size sensitivity. This interaction can be seen in Figure 2 and is further analyzed in the section below. There was no interaction between numeracy and data type.

Sample size weighting by group

Based on the regression results reported above, the data were divided into two groups according to numeracy (high versus low using a median split) and data type. For each group, I found the best fitting power function parameter collapsing across mean difference, standard deviation, and order (see Figure 2). For high numerate subjects in the statistics+data condition, the best predicting sample size weighting function was $N^{.38}$. For high numerates in the statistics-only condition the best predicting equation was $N^{.37}$. Low numerate subjects' data in the statistics+data condition was well fit by $N^{.27}$. Data from low numerates in the statistics-only condition was best described by $N^{.17}$. The standard errors of these exponents ranged from .010 to .022, meaning that no confidence interval included the normative value of .5.

Analysis excluding small sample sizes

In examining Figure 2, it appears that confidence ratings for sample sizes of 1 and 2 may drive the observed relationships among sample size, numeracy, and data type. To test whether these small sample sizes account for the relationship, I reanalyzed the data, but excluded confidence ratings for samples of size 1 and 2. I used the power regression model that included main effects and all possible interactions among sample size, numeracy, and data type. The previously found main effects of sample size, mean difference, and standard deviation remained. Also, the interaction between numeracy and sample size weighting was again found, but the interactions involving data type, sample size weighting, and numeracy did not remain significant. Thus, for subjects lower in numerical ability, presenting raw data appears to increase sample size sensitivity by

decreasing confidence for comparisons that are being made on the basis of very small samples.

DISCUSSION

Experiment 1 shows that laypeople weigh sample size in a nonlinear fashion and that people higher in numerical ability are more sensitive to differences in sample size than are lower numerate people. The curvature of the weighting function shows that confidence generally increases as sample size increases, but sensitivity to differences between sample sizes decrease as the magnitude of sample size goes up. This means that people use sample size, but they appear to become less sensitive as the values increase.

Interestingly, responses from subjects who scored higher on the numeracy scale were better fit by a steeper sample size sensitivity functions compared to those lower in numeracy. This is consistent with Peters et al.'s (2008) account that higher numerates have more precise numerical representations than do lower numerates. Also, subjects lower in numeracy were more sensitive to sample size when they were provided with raw rating data in addition to statistical summaries; this effect appears to be driven by lower confidence ratings for samples of size 1 and 2. That is, it seems that presenting raw data to lower numerate subjects highlights for them just how small samples of size 1 and 2 are. As a result, they give lower confidence ratings for these values compared to participants who were not given raw data presentations. Although raw data presentations improved lower numerate subjects' sensitivity to sample size, their sensitivity function was still shallower than the power coefficients for both groups of high numerate subjects.

Additionally, it is interesting to note, as seen in Figure 2, that subjects do not readily use the lowest end of the 9-point scale. Even when they are only given samples

of size 1 participants' judgments still reflect some amount of confidence in a difference between groups.

Normatively, confidence should increase with the square root of sample size (power function with an exponent of .5). However, the best fitting power function exponents for the data ranged from .17 to .38 indicating that laypeople's weighting functions are shallower than they should normatively be. Can laypeople be pushed to weigh sample size in a normative fashion? An extreme way to test this is to directly inform subjects that sample size is an important factor that should affect their confidence.

EXPERIMENT 2

Experiment 2 was identical to Experiment 1, except for an additional between subjects manipulation. Approximately half of the subjects were given instructions that sample size should matter and that they should incorporate it into their judgments. These subjects were specifically told that they should be more confident "in a difference between two products when more people provided product ratings, as opposed to when fewer people provided ratings".

The goal here was to test whether subjects could normatively incorporate sample size when explicitly told that this factor should affect their judgments. This additional independent variable will be referred to as the use-N factor. The remaining subjects were not given these additional instructions; they serve both as a comparison to the those told to use sample size and also offer a replication of the parameters found in Experiment 1.

METHODS

Undergraduate subjects again compared 38 pairs of products and rated their confidence in a difference between each pair. As in Experiment 1, sample size (10 levels

ranging from 1 to 37), mean difference (1 vs. 2), and standard deviation (2.42 vs. 1.41) were varied within subjects (see Figure 1). Use-N, data type, and order were manipulated between subjects. Subjects answered the same numeracy questionnaire used in Experiment 1. They participated for partial fulfillment of course credit (N=159).

RESULTS

Again using a power model in a nonlinear mixed regression, I found a main effect of sample size (β = .39, t (158) = 92.28, p<.0001) such that subjects' data were fit by $N^{.39}$. Also, effects of mean difference (β = .51, t (158) = 13.40, p<.0001), and standard deviation (β = .31, t (158) = 8.18, p<.0001) were found, replicating the effects of Experiment 1. Subjects' confidence increased as sample size[3] and mean difference increased and as standard deviation decreased. Unlike Experiment 1, a main effect of data type was uncovered (β = .48, t = (158) 2.52, p=.0126) such that subjects were overall more confident in a difference between groups when they viewed both the statistical summaries and corresponding raw data, compared to when they only saw the statistical information on the number lines. There were no main effects of numeracy, order, or use-N.

The non-significant main effect of order was removed from the model and all possible interactions among sample size, numeracy, data type, and use-N were added. An interaction was found between sample size and use-N (β = .30, t (158) = 6.96, p<.0001). Subjects who were explicitly told that they should be more confident in a difference when sample sizes were larger did in fact appear more sensitive than those not given these instructions. Also, as in Experiment 1, an interaction was found between

[3] As in Experiment 1, I found each subjects' average confidence across the 5 largest sample sizes, and compared this to their mean confidence ratings when sample size was at the 5 lowest levels. 86% fit the pattern of more confidence for higher, as compared to lower, sample size values.

sample size and numeracy (β = .17, t = 7.98, p<.0001) such that higher numerate subjects were more sensitive to changes in sample size.

The interactions between sample size and data type (p=.57), and sample size, data type, and numeracy (p=.22) did not replicate. However, a three-way interaction was uncovered among sample size, data type and use-N (β = -.19, t = -2.18, p=.0311) such that the use-N instructions had a larger effect for subjects in the statistics-only group, compared to those who were also given raw data. The four-way interaction among sample size, numeracy, data type and use-N missed significance (β = -.14, t = -1.92, p=.0568). As regards the latter, inspection of the sample size coefficients presented in the next section reveal that subjects higher in numeracy weighted sample size similarly regardless of data type or use-N instructions. In contrast, lower numerates tended to be more sensitive to sample size when told that sample size should affect their confidence judgments.

Sample size weighting by group

As shown in Table 1, subjects were assigned to one of eight groups on the basis of data type, use-N, and numeracy (based on a median split to form two groups). Collapsing across all other factors (mean difference, standard deviation, and order), I fit the power function where $confidence = a + N^b$ (see Figure 3).

The best predicting coefficients for high numerate subjects ranged from .40 to .42 ($N^{.40}$ to $N^{.42}$), all with overlapping confidence intervals. Low numerate subjects in the use-N condition showed sample size sensitivity on par with high numerates (.44 and .41 in the statistics-only and statistics+data conditions, respectively). Low numerate subjects not given instructions to use sample size appeared less sensitive to this factor regardless

of data type ($N^{.29}$ and $N^{.31}$ with overlapping confidence intervals). All eight sample size coefficients are shown in Table 1; also, see Figure 3.

Analysis excluding small sample sizes

Again, as in Experiment 1, it appears that between group differences might be driven by differences between how samples of size 1 and 2 are treated. I employed the regression analysis used above that looked at main effects and also interactions among sample size, use-N, data type, and numeracy. The three-way interaction among sample size, use-N and data type disappears without these judgments. However, sample size weighting was still significantly affected by numeracy, data type, and use-N instructions.

DISCUSSION

Experiment 2 demonstrates that laypeople higher in numeracy are more sensitive to sample size than are lower numerate individuals. However, people lower in numeracy show a sample size weighting function that is similar to high numerates when they are told that sample size matters and should affect their judgments. Regardless of instructions or presentation format, high numerate subjects' confidence judgments were well fit by a power model with a sample size exponent of about $N^{.41}$, while lower numerates judgments were best fit by a power function of $N^{.30}$ except when told to use sample size ($N^{.42}$). These results show that the effects of low numeracy can be counteracted with appropriate instructions. Presenting raw data along with statistical summaries boosted subjects' general confidence in a difference, but did not significantly increase their weighting of sample size as in Experiment 1.

Overall, the power coefficients found here are higher than those found in Experiment 1, but are still significantly below the normative value of .5.

PRODUCT COMPARISON DISCUSSION

Experiments 1 and 2 show that laypeople use sample size when making inferences regarding whether two group differ. Their confidence increases as sample size goes up. However, sensitivity to sample size appears to be nonlinear such that subjects give less weight to changes in sample size as the magnitude of those numbers becomes larger. This power (or logarithmic) weighting function is consistent with past research findings in which subjects seem highly sensitive to sample size when examining relatively small values (Obrecht et al., 2010), but less sensitive when the numbers are larger (Obrecht et al., 2007).

These experiments show a consistent difference between sample size sensitivity as a function of numerical ability. Confidence judgments of higher numerate subjects change more as a function of sample size, compared to lower numerate subjects. This is consistent with the evidence that higher numerate subjects obtain more affective, precise feelings from numbers than do lower numerates (Peters et al., 2006) and also with the finding that high numerates' nonverbal numerical magnitude representations are more precise than are low numerates' (Peters et al., 2008).

From Experiment 1, it seems that low numerate subjects appear to benefit from seeing the datasets that correspond to sample size information. This suggests that for some individuals, the magnitude of a sample size may be better appreciated when viewed as a set whose size increases linearly with the number of items compared to when given as an Arabic numeral on a number line. Such presentation formats may make clear just how small samples of size 1 or 2 really are relative to larger values. Also, subjects lower in numeracy show greater sensitivity to sample size when given brief instructions that

pointed out its importance. That simple instructions have such an effect suggests that laypeople have intuitions about the importance sample size that is easily tapped into.

Interestingly, higher numerate subjects do not become more sensitive to sample size, that is, give judgments consistent with a larger power function exponent, under these conditions. Perhaps high numerate people already have a good representation of how large, for example, a sample size of 10 is, and so seeing the corresponding raw data does not provide further information. Also, the instructions to use sample size may seem redundant with high numerates' intuition that sample size matters. Perhaps high numerate subjects do not feel that they need to increase their weighting of this factor because it is something they already know to consider. Nevertheless, high numerates do not use sample size in a normative fashion. Their sensitivity to sample size is consistently shallower than the normative square root function.

I choose to model subjects' sample size sensitivity in terms of a power function so that exponents could be compared between groups, and also to the normative standard of $N^{.5}$. Experiments 1 and 2 show that laypeople's intuitions about sample size fall short of the square root of N standard. Exponents ranged from .17 to .44, all significantly below .5. However, it is nevertheless impressive that laypeople do have an intuition that sample size should matter and that they integrate it into the judgments.

Next, I switch to a different judgment task with a linear, rather than a square root, normative standard for weighting sample size. I test whether the curvilinear weighting of sample size shown in Experiments 1 and 2 is task dependent, or if a similar functional form will again be found.

OVERVIEW OF EXPERIMENTS 3 AND 4

From Experiments 1 and 2 it appears that laypeople attend to sample size. Their sensitively is well fit by a curvilinear function where the subjective difference between sample sizes decreases as it become larger. From a statistical perspective, this curvilinear shape makes sense because the power to find a difference between groups increases by the square root of sample size.

However, if given a task in which the normative action is to weigh sample size in a linear fashion, will people do so? Or, do laypeople apply a nonlinear sample size weighting function regardless of the task? In order to test the generality of the weighting function modeled in Experiments 1 and 2, I next employ a different sample size task with a linear normative function in Experiments 3 and 4.

If someone were trying to decide the chances of experiencing a side effect from a medication, she might collect information from multiple sources. For example, imagine that one source reports that out of 10 people he knows, 10% experienced side effects from the medication of interest. Perhaps a different person says that 20% of the 30 people she knows had side effects from the medication. When trying to judge the overall chances of an effect (e.g. of experiencing side effects), one should consider both the sample percentages obtained, and also their relative sample sizes. In this example the 20% figure should be given more weight because it comes from a larger sample.

A recent paper by Obrecht et al. (2009) showed that when combining percentages from multiple sources to estimate the likelihood of an event, people tend to provide estimates that look like they ignore sample size. Obrecht et al.'s data are consistent with an averaging model that excludes sample size (or a power model with an exponent near 0, i.e. N^0). Given the example above, a subject might be expected to report that the

chances of side effects are 15%, the value right in between 10% and 20%, despite that these two percents are based on differing amounts of data. Obrecht et al. put forth the encounter frequency hypothesis which states that people are sensitive to the frequency of encounters they have with information. Thus, when one receives multiple pieces of information, the relevant denominator, so to speak, is the sum of the number experiences, not the relative sample sizes of each encounter.

Normatively, data should be combined using a weighted average such that values are weighted in a linear fashion according to their corresponding sample sizes (e.g. $\frac{10\% \times 10 + 20\% \times 30}{10 + 20} = 17.5\%$). Obrecht et al. (2009, Experiment 3) used sample size values that ranged from 10 to 100,000 in their study; subjects largely ignored this factor. However, given the findings of Experiments 1 and 2, it could be that subjects are sensitive to sample size, but that this is difficult to detect with larger numerical values because of a curvilinear weighting function.

Next, Experiment 3 replicates the Obrecht et al. study, but uses much smaller sample sizes. In Experiment 4, sample size magnitude is directly manipulated. If subjects are unable to integrate sample size into their judgments in this paradigm regardless of sample size magnitude then the Obrecht et al. (2009) results should be replicated and the best fitting power model should have an exponent of about 0. However, if subjects do consider sample size, but apply a nonlinear weighting function as suggested in Experiments 1 and 2, then I should find evidence that judgments incorporate sample size when smaller magnitudes are used.

EXPERIMENT 3

The purpose of Experiment 3 was to test whether subjects use sample size information in a different experimental paradigm compared to the previous experiments. Subjects were asked to consider sets of percentages and their corresponding sample sizes in order to make judgments about the overall percent chance of an event occurring. Sample sizes ranging from 1 to 250 were used; these were much lower than the sample sizes used by Obrecht et al. (2009). If subjects are more sensitive to differences between sample sizes when they are lower in magnitude, then judgments should reflect some incorporation of sample size leading to a sample size power coefficient greater than zero.

METHOD

Undergraduate subjects (N=186) participated for course credit. All materials were presented online.

Materials

Subjects were given the following introduction.

> Imagine that you work at a very large nature preserve where many animals live. You are interested in learning more about the animals that you help. At a zoology conference you get a chance to talk to other nature preserve workers who have carefully recorded the chances of various outcomes.

Participants then were given six stories in which they read that an animal could have one of two possible outcomes for a given characteristic. For example, they read that leopards can have round or square spots. Within each story, subjects were given information from six different nature preserve caretakers about how many animals they had seen with the outcome of interest. These caretakers each reported a sample size and percentage (e.g. *One of the caretakers tells you that of the 5 leopards he has seen, 100%*

had round markings. Another nature preserve worker says that of 1 leopard she saw, 100% had round markings…). The percent provided was always possible; for example, when a person reported the outcome of a single event, only percents of 0 or 100% were given. Even though it is unusual to use percentage terminology with single cases, this language was used for consistency. The data from each individual nature preserve worker was given on its own webpage. After viewing reports from six different people, subjects were told that an animal (e.g. leopard) would soon be born on their nature preserve and were asked to estimate the chances that it would have the outcome of interest (e.g. round markings). Subjects gave both a percent estimate from 0 to 100 and a likelihood rating using a 9 point scale where a rating of 1 corresponded to *extremely unlikely* and a rating of 9 corresponded to *extremely likely*.

Design

Subjects considered six datasets; each consisted of six percentages and their corresponding sample size. For example, one of the six datasets was presented in the leopard story; within this story subjects received information from six people. The six sample sizes within each dataset were always the same (1, 2, 3, 5, 80, 250). The percentages within a dataset were varied within subjects to be either low, medium, or high in value. Within subjects, I manipulated how the percentages within a dataset were paired with the sample sizes. Either the larger percentages were paired with the larger sample sizes (large N-large percent pairing), or the smaller percentages were paired with the larger sample sizes (large N-small percent pairing); see Table 2. The two levels of percent-N pairing were crossed with the three levels of the percent range variable to create the six datasets in total. All subjects were given all six datasets.

Between subjects, the pairing of datasets and stories was manipulated to have two levels. Thus, for example, the leopard story was paired with two different datasets between subjects. Also between subjects, I varied the presentation order of the data within each dataset to have two levels. Finally, two possible story presentations orders were chosen, again between subjects.

Predictions

The main independent variable of interest in this study is percent-N pairing. According to the encounter frequency hypothesis (Obrecht et al., 2009), subjects should give larger percent estimates when the average of the percentages in a dataset is greater, regardless of their respective sample sizes. Thus, subjects' weighting of sample size should follow a power function with an exponent of 0 (N^0).

Normatively, if subjects weigh percentages according to sample size, they should give higher estimates when the weighted average of percents is higher such that their data should be fit by a sample size power function with 1 as the exponent (N^1). However, if subjects instead apply a curvilinear weighting of sample size, then their estimates will fall in between these two extreme predictions.

Below I use two methods to examine these hypotheses. First, I will apply the regression modeling used in Experiments 1 and 2 in order to find the best predicting power model parameter for sample size. If subjects ignore sample size, the best fitting sample size parameter will be close to 0. In contrast, the normative model predicts a coefficient of 1. Although the previously presented experiments are quite different, their parameters will be compared to those found in the current study. From Experiments 1 and 2, the overall average power coefficient was $N^{.32}$ for those subjects who were not

provided with raw data or instructions to use sample size[4]; if subjects here apply a curvilinear function, a similar power coefficient may be found. In order to allow the reader to compare these results to those reported in the Obrecht et al. (2009) paper in which an ANOVA was used, after presenting the modeling analyses, I will also present an ANOVA to test for main effects and interactions among the independent variables.

RESULTS

Regression modeling

Regression modeling was used with the percent estimate dependent measure. As in the previous experiments, a power regression model was used to determine the best fitting sample size coefficient for the data.

$$Percent\_estimate = \frac{\sum_{i}^{6} p_i (N_i^b)}{\sum_{i}^{6} (N_i^b)}$$

Where p_i = percent within a dataset, N_i^b = sample size within a dataset raised to some power *b*, and *i* counts across the items within a dataset. The nonlinear regression model assumes a weighted average form where, depending on the value of the power coefficient, sample size may be taken into account to a greater (b=1), lesser (b=0), or intermediate (0<b<1) extent. This model took subjects' repeated measures into account across each of the 6 datasets they considered. The power coefficient, *b*, was the only free parameter in the regression formula; the remaining variables, *p* and *N*, came from the percentages and sample sizes provided in the datasets given to subjects.

---

[4] $\frac{.30(58)+.35(35)}{58+35} = .32$

Overall, subjects' judgments were best fit by $N^{.28}$ (β = .28, t (185) = 15.14, p<.0001) indicating that they considered sample size to some extent. This value is surprisingly similar to the value of .32 observed in the relevant conditions from Experiments 1 and 2. This power model provided a better fit to the data than either the normative (N^1) or encounter frequency model (N^0). Comparing log likelihood fit statistics for the different models showed that the .28 exponent was 224 times more likely than the encounter frequency predicted exponent of 0, and 400 times more likely than the normative predicted exponent of 1[5].

I computed separate regression analyses for high verses low numerate subjects. The sample size coefficient for high numerates was $N^{.34}$ with a confidence interval of .29-.40; this was significantly greater than zero (β = .34, t (185) = 12.46, p<.0001). Low numerates' percents were fit by $N^{.22}$ with a confidence interval of 17 to .27 (β = .22, t (185) = 8.64, p<.0001); see Table 4. Figure 4 shows the predictions of the normative and encounter frequency models, along with the best fitting power model predictions and subjects actual average percent estimates for all six datasets.

ANOVA

---

[5] The log model was entered as $Percent\_estimate = \frac{\sum_i^6 p_i(\ln(N_i))}{\sum_i^6 (\ln(N_i))}$, while the linear model was entered as $Percent\_estimate = \frac{\sum_i^6 p_i(b(N_i))}{\sum_i^6 (b(N_i))}$. The log likelihoods associated with the models were subtracted in a pairwise fashion to test their relative fits to the data.

In order to allow for a comparison between the current experiment and that reported by Obrecht et al. (2009), I will next report a mixed-model ANOVA using the percent dependent measure. Analysis of the rating scale dependent measure will be excluded here and in Experiment 4 because it provided redundant results.

The percent measure showed an effect of percent-N pairing ($F(1,170)=12.92$, p=.0004, $MSE=212.21$). Subjects gave significantly higher chance estimates when considering the large N-small percent pairing datasets, compared to the large N-large percent datasets. This means that, overall subjects' data favored the encounter frequency model over the normative model. However, percent-N pairing interacted with percent range ($F(2,340)=28.96$, $p<.0001$, $MSE=142.74$) such that subjects' judgments were in the direction predicted by the encounter frequency hypothesis for the high and low range dataset, but in normative direction for the medium range dataset.

As expected, percent-N pairing interacted with numeracy ($F(1,170)=11.71$, $p=.0008$, $MSE=212.21$) such that lower numerate subjects gave percentages estimates that were closer in line with the encounter frequency hypothesis (N^0) compared to the normative model (N^1); in contrast, higher numerates' estimates did not significantly favor either model's predictions, but rather, estimates fell in between the values predicted by the two models.

Unsurprisingly, there was a main effect of percent range ($F(2,340)=1696.15$, $p<.0001$, $MSE=201.82$). This simply means that subjects gave percent estimates that were in line with the percentages provided in each dataset (i.e. gave lower percent estimates when given lower percents, gave higher percent estimates after viewing high

percents). There were also effects of little interest involving the counterbalancing factors.[6]

DISCUSSION

The results of Experiment 3 suggest that subjects treat sample size in a curvilinear fashion. The task employed here was quite different from that used in the first two experiments. Also, its normative standard requires weighting sample size linearly, rather than according to its square root. Therefore, it is surprising that subjects' judgments nevertheless suggest a similar power weighting function as in Experiments 1 and 2. Here subjects' judgments were well fit by a sample size power function with an exponent of about .28. This value is in the ballpark of the average exponent of .32 shown by subjects who were not given raw data or special instructions from the two earlier experiments.

Also, as in Experiments 1 and 2, high numerate subjects appear to consider sample size to a greater extent than low numerates. Their judgments reflect a power function with a steeper slope ($N^{.34}$) compared to lower numerate subjects ($N^{.22}$).

The ANOVA analysis showed, overall, that subjects' judgments significantly favored the encounter frequency hypothesis over the normative model. However, a comparison of model fits showed that subjects' data were still better fit by a model where sample size was weighted to a power. That is, subjects' responses were not described as well by a linear normative model (N^{1}), or by an encounter frequency model that ignores

[6] Percent estimates were affected by a main effect of data order ($F(1,170)=3.96$, $p=.0482$, $MSE=236.40$). Story order and data order interacted ($F(1,170)=7.40$, $p=.0072$, $MSE=236.40$). Story-data pairing interacted with numeracy ($F(1,170)=5.04$, $p=.0261$, $MSE=236.40$). Percent-N pairing interacted with story order ($F(1,170)=10.35$, $p=.0016$, $MSE=212.21$); this was qualified by interactions among percent-N pairing, percent range, story-data pairing and story order ($F(2,340)=3.73$, $p=.0251$, $MSE=142.74$), as well as percent-N pairing, percent range, story-data pairing and numeracy ($F(2,340)=3.35$, $p=.0362$, $MSE=142.74$). There were 3-way interactions among percent range, story-data pairing, and story order ($F(2,340)=4.57$, $p=.0110$, $MSE=203.82$), and also among percent range, story order and numeracy ($F(2,340)=6.71$, $p=.0014$, $MSE=203.82$).

sample size (N^0), as they were by a curvilinear weighting model with an intermediate power exponent for N.

The results of Experiment 3 differ from that found by Obrecht et al. (2009). The only obvious procedural difference between these two studies is that Obrecht et al. used larger sample size values. There are two independent aspects of sample size magnitude that could account for this difference.

First, a power function with an exponent of less than 1 implies that, for a given difference between two values, one will be more sensitive to the difference between two smaller sample sizes, compared to two larger samples sizes. For example, the difference between samples of size 6 and 2 will feel larger than the difference between 26 and 22, despite that both numbers pairs differ by 4. This is because, for example, with $N^{.3}$, $6^{.3} - 2^{.3} > 26^{.3} - 22^{.3}$. Thus, laypeople could apply a constant sample size weighting function (e.g. $N^{.3}$), but appear less sensitive to differences between larger numbers.

In contrast, an alternative account is that the coefficient of a power weighting function could change as a result of the sample size magnitude. Perhaps subjects are fairly sensitive to differences between samples sizes when they are under 100 (e.g. give judgments consistent with $N^{.4}$), but when asked to consider larger values, sensitivity drops (e.g. $N^{.2}$). Although the power coefficients found in Experiment 3 are larger than those found by Obrecht et al. (2009), it is not clear which or what combination of these two explanations account for this difference.

The second account can be tested by multiplying sample sizes by a constant such that the relationship among the values stays the same, but the magnitude increases. If subjects' percent estimates are better fit by a lower power coefficient when sample sizes

are larger, compared to when they are smaller, then I will have strong evidence that subjects' weighting of sample size depends on the range or magnitude of the values, not just the curvilinear decreasing sensitivity functional form. In Experiment 4, this hypothesis is tested.

EXPERIMENT 4

The goal of Experiment 4 was to test whether sample size sensitivity is affected by sample size magnitude. In Experiment 4, sample size was manipulated between subjects to be either low or high in magnitude. This was done by multiplying the smaller sample sizes by a factor of 10 so that the relationship among them stayed constant, but the magnitude increased. If subjects apply a general curvilinear weighting function of sample size, then their data should be best fit by similar power functions, regardless of the sample size magnitude. However, if subjects give less consideration to sample size when it is larger, then they should show lower power exponents in the high magnitude, compared to the low magnitude, condition.

METHOD

Undergraduate subjects (N=373) participated for course credit. All materials were presented online.

Materials

Subjects were given the same nature preserve introduction as in Experiment 3. For each of the six stories they read about, they judged the chances of an outcome by indicating the percent chance of the event and also by giving a rating on the 9-point scale. However, because both the percent chance and rating scale showed similar results, only the percent dependent measure will be further discussed.

Design

As in Experiment 1, subjects viewed 6 datasets. Each dataset provided 8 percentages that were paired with a sample size. Between subjects, participants were either given sample sizes that were relatively small (1, 1, 1, 1, 5, 25, 125, 625) or large (10, 10, 10, 10, 50, 250, 1250, 6250) in magnitude (see Table 3).

Within subjects, the range of the percentages within a dataset was manipulated to be either low (0 to 40%), medium (20 to 60%), or high (78 to 100%). These ranges only describe the percentages that were paired with samples of size 25, 125, 625 or 250, 1250, 6250, depending on condition. The percentage values were constrained to be possible given the sample size that each was paired with. Thus, samples of size 1 could only provide percentages of 0 or 100%. This also meant that the analogous samples of size 10 in the large sample size magnitude condition could also only be paired with either 0 or 100%. As samples of size 5 could only provide 6 possible percentages, percents paired with 5 or 50 sometimes fell outside of the low, medium, and high ranges listed above.

Also within subjects, I manipulated how percentages and sample sizes within datasets were paired together. Within a dataset, larger sample sizes were either paired with smaller percentages (large N-low percent paring) or they were paired with larger percentages (large N-high percent pairing). This means the small sample sizes were paired with larger percentages in the former case, while in the latter, small sample sizes were paired with smaller percents. This sample size-percentage pairing was crossed with the three percentage ranges to give the 6 datasets. Each of the 6 datasets contained the same 8 sample sizes, either low or high magnitude, shown above. To make this design

clearer, Table 3 shows the two high percent range datasets given to subjects in both sample size magnitude conditions.

The pairing of the 6 datasets within stories was manipulated between subjects to have two levels. Also between subjects, two presentations orders of the 6 datasets and two possible orderings of the 8 percentages within each dataset were used.

RESULTS

Regression Modeling

As in Experiment 3, subjects' percentage estimates were fit by a regression power model.

$$Percent\_estimate = \frac{\sum_{i}^{8} p_i (N_i^b)}{\sum_{i}^{8} (N_i^b)}$$

Where p_i = percent within a dataset, N_i^b = sample size within a dataset raised to some power *b*, and *i* counts across the items in the datasets given to each subject. As in the analysis for Experiment 3, this model was implemented so to account for repeated measures.

Overall, sample size was weighted by $N^{.30}$ (β = .30, t (372) = 20.75, p<.0001). However, as the between group analyses show, this coefficient differed across groups.

Between group analyses

In order to test for main effects of sample size magnitude and numeracy, two regression analyses were computed. Overall, the percent estimates of subjects given the low magnitude sample sizes were fit by a sample size power function of $N^{.34}$ (C.I. = .30-.38, β = .34, t (372) = 16.01, p<.0001). For subjects in the high magnitude condition, this

function was $N^{.26}$ (C.I. = .22-.30, β = .26, t (372) = 13.19, p<.0001). Figure 5 shows two hypothetical sample size power weighting functions with these respective slopes.

High numerates had higher coefficients of $N^{.37}$ (C.I. = .33-.42, β = .37, t (372) = 16.23, p<.0001) as compared to lower numerates $N^{.24}$ (C.I. = .21-.28, β = .24, t (372) = 12.88, p<.0001).

To further break this down, I computed the regression analyses for each group. High numerate subjects in the low sample size magnitude condition (β = .42, t (372) = 12.17, p<.0001) seemed to show greater sample size sensitivity compared to high numerates in high magnitude group (β = .33, t (372) = 10.61, p<.0001). The same pattern was observed between low numerates in the low magnitude sample size group (β = .28, t (372) = 10.17, p<.0001) compared to low numerates in high magnitude group (β = .21, t (372) = 8.01, p<.0001); see Table 4.

Thus, there are main effects of sample size magnitude and numeracy. People are less sensitive (show lower power coefficients) when sample size is larger, rather than smaller, in magnitude. Also, high numerates show greater appreciation of sample size compared to low numerates.

ANOVA

The apparent main effect of magnitude on sample size weighting was confirmed in a mixed model ANOVA in which there was a significant interaction between N-percent pairing and magnitude (F(1,340)=5.51, p=.0195, MSE=244.08). The N-percent pairing factor reflects whether subjects' percent judgments are significantly in the direction favoring the normative (N^1) or the encounter frequency (N^0) model predictions. This factor's interaction with N-magnitude shows that the group of subjects

who received higher magnitude sample sizes gave percent estimates that were closer to the encounter frequency hypothesis than those given low magnitude sample sizes. That is, higher magnitude sample sizes translate into lower power coefficients that are closer to N^0.

Percent-N pairing also interacted with numeracy ($F(1,340)=15.07$, $p=.0001$, $MSE=244.08$) such that lower numerate subjects' estimates were closer to the encounter frequency hypothesis (N^0) than higher numerates' percent estimates.

Main effects of N-percent pairing ($F(1,340)=14.28$, $p=.0002$, $MSE=244.08$) and percent range ($F(2,680)=2215.00$, $p<.0001$, $MSE=297.68$) were also found. Thus, subjects judgments overall favored the N^0 over the N^1 model, and their percent estimates reflected whether they viewed low, medium, or high percentages in a given dataset. Also, subjects' judgments were influenced by interactions between percent range and magnitude ($F(2,680)=10.31$, $p<.0001$, $MSE=297.68$) and percent range and numeracy ($F(2,680)=12.53$, $p<.0001$, $MSE=297.68$). That is, percent estimates were significantly in favor of the encounter frequency, over the normative, predictions for the high and low percent ranges, but neither model was favored for the medium percent range datasets. Also, higher numerate subjects showed a larger effect of percent range than lower numerate subjects. Counterbalancing effects of minor theoretical interest were uncovered[7].

[7] There were main effects of story-data pairing ($F(1,340)=4.16$, $p=.0423$, $MSE=302.64$) and story order ($F(1,340)=4.13$, $p=.0429$, $MSE=302.64$). Story order interacted with numeracy ($F(1,340)=5.92$, $p=.0155$, $MSE=302.64$). Percent-N pairing interacted with magnitude, story order, and numeracy ($F(1,340)=3.94$, $p=.0479$, $MSE=244.08$). Interaction were uncovered among percent range, magnitude, story-data pairing, and story order ($F(2,680)=3.21$, $p=.0426$, $MSE=297.68$), as well among percent range, magnitude, story-data pairing and numeracy ($F(2,680)=3.47$, $p=.0315$, $MSE=297.68$). Percent-N pairing interacted with percent range and story-data pairing ($F(2,680)=4.92$, $p=.0076$, $MSE=167.64$). This was qualified by an interaction among percent-N pairing, percent range, magnitude, story-data pairing, and story order

Modeling previous findings

Using the data from Obrecht et al.'s (2009, Experiment 3) study, I reanalyzed their results to find the best fitting power parameters. They gave subjects sample sizes that ranged from 10 up to 100,000. If the scale of sample sizes being compared affects the slope of the weighting function, then I should find a smaller power exponent for these data, compared to Experiments 3 and 4 in the current paper. In line with Obrecht et al.'s conclusion that subjects appear to nearly ignore sample size, the best predicting power exponent was .06 with a confidence interval of .03-.09 (t (98) = 4.30, p<.0001). When broken down by numeracy level, it appeared that low numerates gave sample size no weight (β = -.01, C.I. = -.05 to -.003, t (98) = -.44, p=.66), but those higher in numeracy still showed slight sample size sensitivity (β = .13, C.I. = .09 to.16, t (98) = 6.54, p<.0001).

DISCUSSION

The results of Experiment 4 show a clear effect of sample size magnitude. When asked to consider percentages paired with sample sizes that ranged from 1 to 250, subjects' percent estimates were fit by a power model of $N^{.34}$. When the same exact percentages were presented, but with corresponding sample sizes that were ten times larger (ranging from 10 to 2,500) subjects' percent estimates were fit by a $N^{.26}$ model. This suggests that laypeople do not treat sample sizes according to one function, but that their sensitivity is affected by the scale or magnitude of the values (see Figure 5).

(F(2,680)=5.62, p=.0038, MSE=167.64). Finally, percent-N pairing interacted with percent range, magnitude, story order and numeracy (F(2,680)=3.25, p=.0396, MSE=167.64).

Further bolstering this conclusion, a reanalysis of the Obrecht et al. (2009, Experiment 3) data shows that when subjects consider datasets that include sample sizes going up to 100,000, they show almost no sensitivity to this factor ($N^{.06}$).

Consistent with the previous experiments presented in this paper, Experiment 4 demonstrates a clear relationship between numerical ability and the extent to which subjects consider sample size into their judgments. Higher numerate subjects again appear to have a steeper sample size sensitivity function slope compared to lower numerates.

GENERAL DISCUSSON

These experiments demonstrate that laypeople have intuitions about how sample size should affect their judgments. When making inferences, people's confidence generally increases as sample size goes up. Also, when combining data, they give greater weight to percentages that describe more information, compared to percentages that summarize less data.

However, the experiments presented here show that weighting of sample size follows a nonlinear pattern that is consistent with a power or logarithmic function. A function with this curvilinear shape has the property of decreasing sensitivity; that is, sensitivity to a given difference will seem larger at the lower end of the scale, compared to a higher end of the scale. For example, a difference of 5 will feel larger when considering samples sizes of 3 and 8, compared to when considering values of 33 and 38. This property could account for some of the discrepancies found in the literature in which laypeople appear highly sensitive to sample size in some cases where the values under consideration are relatively small (Masnick & Morris, 2008, Jacobs & Narloch, 1999,

Nisbett et al, 1983, Obrecht et al., 2010), but not others where the values under consideration are larger (Obrecht et al., 2007, Obrecht et al. 2009).

However, Experiment 4 shows that an alternative explanation is possible for the discrepant findings regarding lay sensitivity to sample size. In this study, subjects considered datasets containing percentages and their corresponding sample sizes. They used this information to judge the chances of some event occurring. I manipulated sample size directly and found that, when sample sizes were lower in magnitude, subjects' percent estimates were consistent with a sample size power function with a higher coefficient, compared to when the sample sizes were multiplied by a constant to make them larger in magnitude. This means that some aspect of sample size magnitude affects people's sample size sensitivity curve; it's not simply that subjects have one general sensitivity curve that appears to have a different slope depending on which section of the curve is being measured, but rather, that the steepness also changes as a function of the values being considered. Sample size magnitude could plausibly affect the slope of the power function in a couple of ways. It could be that sensitivity decreases as the range of values increases; e.g. the range from 1 to 650 is smaller than the range from 10 to 6,500. Or, it is possible that sample size weighting is curvilinear up to some value, but then the function levels off for values beyond some threshold. For example, it could be that subjects are sensitive to differences between numbers under 1,000, but weight all numbers over this threshold equally, leading to an overall lower power function value. The current experiments do not allow for such a fine grain analysis, but nevertheless still show that sample size weighting is affected in part by the magnitude of the values under consideration.

Overall, it appears that both the nonlinear shape of subjects' sample size weighting functions, and also the magnitude of samples sizes, play a role in how laypeople attend to this factor (see Table 4).

Nonlinear Implications

Across four experiments in two different tasks, subjects' weighting of sample size was well fit by a power function with an exponent of about .3. How should this exponent be interpreted? It could indicate a systematic bias such that laypeople consistently transform sample sizes according to a curvilinear function. Alternatively, laypeople's representations of sample size could be linear with scalar variability. This would mean that, on average, their sense of sample size magnitudes map the actual values linearly, but are represented in an increasingly imprecise fashion as values become larger; this would lead to reduced discrimination among larger numbers. These different accounts have been debated in the nonverbal numerical magnitude literature in which some argue for compressed logarithmic representations of number (Siegler & Booth, 2005), while others contend that humans represent numerical magnitudes linearly, but with scalar variability proportional to the magnitude (Gallistel & Gelman, 2005). Both of these positions could account for the numeracy findings reported in the current paper; compared to higher numerate subjects, those lower in numeracy may treat sample size according to power function with a low exponent or they may have more scalar variability in their representations.

Memory Considerations

In Experiments 3 and 4, subjects viewed percentage and sample size data sequentially. In Experiment 4 I found that subjects were less sensitive to changes in

sample size when these values were higher, compared to lower, in magnitude. Memory could play a role in explaining this result. If subjects attempted to explicitly hold all of the data that was presented within a story in working memory, this task might have been more difficult in the large magnitude sample size condition. For example, if remembering the value 10 requires more working memory resources than remembering a sample size of 1, then it could be that the sample size magnitude effect is related to subjects' ability to hold all of the information in mind for consideration. That is, forgetting some sample size values might relate to lower sample size sensitivity. The process by which people combine information over time, be it implicitly or explicitly, remains for future exploration.

Ties to Previous Research

Consistent with the decisions from experience literature (e.g. Hertwig, et al., 2004, Gottlieb et al, 2007, Hau et al., 2008), subjects appear to have a sharply diminishing sensitivity to sample size values. When sampling from a population in order to gain an understanding of a payout structure, subjects choose to view a median of about 7 samples. Obviously, many factors will affect how large of a sample seems large enough, such as incentives to gain an accurate understanding of the population (Hau et al., 2008) and the expected population variability (e.g. Nisbett, et al, 1983). Nevertheless, the low number of cards selected by subjects in these studies suggests that laypeople feel satisfied with small sample sizes; although additional samples would be easy to obtain, they are apparently not worth the time it takes to collect them. Although it is highly speculative to posit, this suggests a sample size sensitivity function with a slope that quickly drops off.

Also of interest given the current results, Obrecht et al. (2009) put forth the encounter frequency hypothesis in which they state that laypeople are sensitive to the frequency of encounters that they have with information, rather than the sample size values that correspond to that data. Experiments 3 and 4 in the current paper replicated Obrecht et al.'s Experiment 3, but used smaller sample sizes. Experiments 3 and 4 show that subjects do give greater weight to percentages with higher, as compared to lower, sample sizes. However, this sample size sensitivity decreases off as the magnitude of the values increases. A reanalysis of the Obrecht et al. (2009) Experiment 3 data shows that when subjects are given percentages with corresponding samples as large as 100,000, their sensitivity drops down to be near zero. Thus, the results of this paper provide a refinement of the encounter frequency hypothesis; laypeople do use sample size, but more so when they are dealing with numbers that are relatively small (1 to 650) compared to when numbers are larger in size (10 to 6,500). Further work needs to be done to explore whether the relevant factor is the range, overall magnitude of the sample size values under consideration, or other factors that related to sample size magnitude.

Numeracy

All four experiments presented here show a consistent relationship between numerical literacy and sample size weighting. Subjects who score higher on our modified version of the Lipkus et al. (2001) numeracy scale show steeper sample size weighting functions. This effect may be even more pronounced in the general population because, presumably, the university student samples used in the current experiments provided a somewhat limited and optimistic range of numerical ability across adults.

Recently, Peters et al. (2008) showed evidence that individual differences exist in the precision of the people's nonverbal mental numbers, and importantly, that these differences relate to the numerical choices subjects make (also see Halberda, Mazzocco, and Feigenson, 2008 regarding the relationship between nonverbal magnitudes and math achievement). Peters et al. measured how quickly subjects respond when asked whether a quantity was larger or smaller than a target value. As expected, reaction times (RTs) were faster when subjects compared values that were further, rather than closer, from one another (e.g. 5 vs. 9 compared to 5 vs. 6). However, the size of this distance effect (Moyer and Landauer, 1967) differed between subjects; some individuals showed smaller differences between their near versus far RTs compared to others for whom this difference was larger.

Peters et al. infer from this that the former subjects (those with smaller differences in their near verses far RTs) have more precise nonverbal numerical representations than do the latter. They describe these high verses low precision representations as relating to mental logarithmic number lines with different bases. Log functions, regardless of their base, share the same ratio-dependent relationships and have the same functional shape. However, as the base of the log function decreases, discriminability between a given pair of values increases. Considering the Gallistel and Gelman (2005) position, the analogous account would be to posit that smaller distance effects relate to proportionally less scalar variability among individuals' representations compared to those individuals showing larger distance effects.

When choosing between receiving $10 now or $15 later, overall, Peters et al.'s (2008) subjects preferred the $15 later. However, preference for the $15, as rated on a

preference scale, was higher for subjects with more precise numerical representations, compared to those with less precise representations. Similar effects were shown in other tasks, which together make a strong case for the idea that nonverbal numerical magnitudes influence how people make higher level decisions and choices involving numbers. Peters et al. (2008) note that subjects' scores on a numeracy scale showed some relationship to mental magnitude precision.

If higher numeracy indicates greater precision in nonverbal numerical magnitudes, then perhaps higher numerate subjects show consistently steeper sample size weighting functions because they are better able to distinguish between different sample size values, compared to lower numerate subjects. Also, it could be that, regardless of numeracy level, once numbers reach a certain size (e.g. 1,000 or 10,000) that people do not have a strong sense of the how large these values are. For example, perhaps one has mapped, from the nonverbal numerical system to Arabic numerals, values from 0 to about 1,000. However, Arabic numerals greater than about 1,000 may not have an underlying mapping to an analog magnitude. Thus, numbers above this cut off may all feel similarly large. Just because we can count to 1,000 or even a million does not mean that we have an intuitive understanding of how large these values are. In the current media environment in which very large numbers, including sample size values, are described, it may be the case that people have little sense of what these values refer to. Although people higher in numerical ability have more precise representations than lower numerate individuals, they still show decreased sensitivity when considering larger sample size values.

Improving Sample Size Use

In terms of intervention effects, Experiment 1 shows that people lower in numerical literacy benefit from viewing the raw data that correspond to sample size values. If low numerate subjects have imprecise representations of how large sample size values are, their number sense may be enhanced by providing displays that increase in magnitude as a function of sample size. This finding parallels work showing that subjects benefit from pictograph displays of statistical information compared to standard numerical presentations involving Arabic numerals (Zikmund-Fisher, Ubel, Smith, Derry, McClure, Stark, et al., 2008); also, the benefits gained from a pictograph appear more pronounced for lower numerate individuals, in comparison to higher numerates (Hawley, Zikmund-Fisher, Ubel, Jankovic, Lucas, and Fagerlin, 2008).

However, this boost in sample size sensitivity as a function of raw data presentation was not replicated in Experiment 2. There, instead, raw data presentations generally boosted subjects' confidence ratings. Also, low numerate subjects appeared to benefit from a brief instruction stating the importance of considering sample size information; subjects lower in numeracy who received these instructions weighted sample size similarly to high numerates.

High numerate subjects' best fitting power coefficient for sample size did not change as a result of these manipulations. It may be that higher numerate subjects already have fairly accurate representations of how large a sample size is and that they know that it should matter to their judgments. Thus, presenting the corresponding raw data, or giving instructions to use sample size may seem redundant to them. However, despite that high numerate subjects showed a steeper sample size weighting function than

those lower in numeracy, their judgments still consistently fell short of, and were insensitive, to normative standards.

Normative vs. Descriptive Gap

Although it is clear that laypeople do integrate sample size into their judgments and inferences, they consistently fall short of normative standards. In Experiments 1 and 2 subjects' weighting of sample size was on average .34, below the normative standard of .5. In Experiments 3 and 4 the overall best fitting power coefficient was .29[8], even further below the normative standard of 1. It appears that subjects have little understanding of the normative standards that were relevant in these studies. Given that the sample sizes in Experiments 3 and 4 were larger than those used in Experiments 1 and 2, and that sample size magnitude affects the slope of the best fitting power function (as shown in Experiment 4), it may be reasonable to find lower power coefficients for the latter studies, despite the drastic differences between normative standards.

However, it important to note that although judgments fell short of the normative standards in these studies, laypeople nevertheless have the intuition that sample size matters. This is probably why the simple manipulations employed in Experiments 1 and 2 were able to improve low numerate subjects' weighting of sample size.

One particularly interesting aspect of these studies is that subjects appear to weigh sample size in a similar fashion across two completely different tasks with different normative standards. In the first two experiments, sample size was manipulated across pairs. Subjects judged how confident they were that there was a difference between two

---

[8] $\frac{.28(186)+.30(373)}{186+373}=.29$

products. Normatively, because this task is analogous to an intuitive *t*-test, their confidence in a difference should have increased with the square root of sample size.

In contrast, in the last two experiments, subjects were to combine percentage estimates regarding the chances of some event occurring. Each percent was given with a sample size. Based on this dataset, subjects' were to estimate the percent chance of the event of interest. Normatively, judgments should reflect a weighted average where percents are weighted linearly according to their sample sizes. Here, sample size values were not manipulated across trials like in the first two studies, but the same sample size values appears in each story.

Despite these major differences, subjects' percent chance estimates were consistent with the sample size power exponents found in Experiments 1 and 2; across all four experiments power coefficient values were always under .5 and tended to be around .3 or .4 in value. It seems that when people consider data on the basis of sample size, sample size values are evaluated in a task independent manner. Also, regardless of the task, sensitivity to sample size is well fit by a curvilinear function. Overall it appears that laypeople may generally weigh sample size in a nonlinear fashion regardless of the relevant normative standard.

Low Level Influences

Currently, perhaps the most dominate theoretical account for how people make judgments and decisions is the dual system view. It is thought that people have two reasoning systems, System 1 and System 2. System 1 is thought to be an automatic, affective, intuitive system, while System 2 is a slow, deliberative, and more rational system (Kahneman, 2002, also see Stanovich, 1999 for an in-depth review). Tversky and

Kahneman's (1971) work points to the idea that humans neglect sample size because they rely on System 1 heuristics, such as representativeness. The idea seems to be that if subjects would instead employ their deliberative systems (i.e. System 2s), then they would make more rational judgments.

In my view, focus on the dual system account somewhat skirts the issue of representation. This is because it explains non-normative behavior by pointing to System 1 processes. It seems that there is a tacit assumption that people perceive sample size values veridically, as if they always represent an exact quantity. Problems with reasoning and inferences are primarily attributed to how people use these numbers, not how they represent them. Surely, both issues of representation and information processing are important to study.

Thus, it may be useful to take a step back and ask how humans represent quantities. We may give subjects a number such as 1,000, but we don't necessarily know what their representation of that value is. Recent work by Halberta et al. (2008) and Peters et al. (2008) is exciting because it suggests that lower level numerical representations influence higher level reasoning and mathematics. Therefore, it is important to understand how numbers are represented at a basic level in order to inform higher level judgment research.

Related to this point, it is interesting that similar curvilinear functions were found across the four experiments presented in this paper. This suggests that laypeople may generally treat sample size in similar, task independent, fashion. Also, the shapes of the curves appear similar to Weber fractions that are usually associated with perceptual magnitudes where the discriminability of two stimuli is proportional to their ratio. This

could suggest that low level representations may affect how people make presumably higher level, explicit judgments on the basis of numerical data, such as sample size.

Assumptions and Limitations

The modeling work in this paper makes a number of assumptions about how subjects attend to and combine statistical information. In Experiments 1 and 2, from looking at the data, it was clear that subjects treated sample size in a nonlinear fashion. This was backed up statistically by comparing the fits of power and logarithmic models to linear models. However, in this paper I am not claiming that laypeople definitively treat sample size according to a power function; other nonlinear models, such as a logarithmic model, also model the data well. However, because the power model allowed for comparison to a normative standard, it was a reasonable functional form to use for the sake of comparison in this paper.

With the modeling work from Experiments 3 and 4, I assumed that subjects attended to all of the data that they were provided with and that they combined the information in a weighted average fashion. It could be that different subjects employ different strategies or heuristics for combining data. Also, primacy or recency effects could lead subjects to give greater weight to data presented first or last in a scenario. However, because sample sizes within datasets were presented in a pseudorandom fashion, and data ordering within datasets was varied between subjects, such effects cannot account for the modeling results found. I tested a number of different heuristic models not described in this paper and none appeared to clearly predict subjects' percent estimates.

Furthermore, although I describe subjects' judgments in relation to power and weighted average models, I am not claiming that these are processes by which people cognitively combine and weight information. I do not assume that subjects are consciously keeping track of data so to explicitly compute a weighted value. Instead, I am interested in describing functional form that is consistent with laypeople's intuitive judgments.

Conclusions

The studies presented here show that laypeople are sensitive to the law of large numbers. They give greater weight to data that describe larger, as compared to smaller, sample sizes. However, people's treatment of sample size is best modeled by a curvilinear, negatively accelerating weighting function with decreasing sensitivity to larger values. Given that large sample size values are common in the modern world (e.g. a medical study with 9,000 subjects) it is important to understand how lay sensitivity to sample size can be improved in the large number range.

Individual differences in numerical ability relate to sample size sensitivity. Individuals with higher numerical abilities consistently show greater sensitivity to sample size than do people lower numerical ability. However, lower numerate people's attention towards sample size is improved by simple interventions.

Although lay use of sample size is non-normative, the results presented here suggest that people have intuitions regarding the utility of sample size. These intuitions provide a basis upon which training interventions may build in order to improve how humans draw inferences from data.

References

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis,* Basel: Thurneysen Brothers.

Darke, P. R., Chaiken, S., Bohner, G., Einwiller, S., Erb, H., & Hazlewood, J. D. (1998). Accuracy motivation, consensus information, and the law of large numbers: Effects on attitude judgment in the absence of argumentation. *Personality and Social Psychology Bulletin, 24*, 1205-1215.

Evans, J. St. B. T., & Dusior, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgments. *Acta Psychologica, 41*, 129–137.

Gallistel, C. R., & Gelman, R. (2005). Mathematical cognition. In K. Holyoak & R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 559–588). Cambridge: Cambridge University Press.

Gottlieb, D., Weiss, T., & Chapman, G. B. (2007). Presentation of uncertainty information affects decision biases. *Psychological Science, 18*, 240-246.

Halberda, J., Mazzocco, M. M. M, & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455,* 665-668.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21*, 493-518.

Hawley, S. T., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Lucas, T., Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counseling, 73*, 448-455.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare effects in risky choice. *Psychological Science, 15*, 534-539.

Irwin, W. F., Smith, W. A. S., Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experiment Psychology, 51*, 261-268.

Jacobs, Jacobs, J. E., & Narloch, R. H. (2001). Children's use of sample size and variability to make social inferences. *Applied Developmental Psychology*, *22*, 311-331.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. Nobel Prize Lecture. Retrieved March 13, 2010, from http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Kaufmann, M., & Betsch, T. (2009). Origins of the sample-size effect in explicit evaluative judgment. *Experimental Psychology, 56*, 344-353.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37-44.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519–1520.

Nisbett, R. E., Krantz, D. H., Jepson, C. & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339-363.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t*-tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14,* 1147-1152.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition, 37,* 632-643.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning, 16,* 26-44.

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mozzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*, 406-413.

Peters, E., Slovic, P., Vastfjall, D., Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making, 3*, 619-325.

Pitz, G. F. (1967). Sample size, likelihood, and confidence in a decision. Psychonomic *Science, 8*, 257-258.

Sedlmeier, P. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10,* 33-51.

Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. Journal *of Behavioral Decision Making, 11*, 281-301.

Siegler, R. S., & Booth, J. L. (2005). Development of numerical estimation: A review. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 197-212). Boca Ratan, FL: CRC Press.

Stanovich, K. E. (1999). *Who is Rational?: Studies of Individual Differences in Reasoning.* Mahwah, New Jersey: Lawrence Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185,* 1124-1131.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Zikmund-Fisher, B. J., Ubel, P. A., Smith, D. M., Derry, H. A., McClure, J. B., Stark, A. T., Pitsch, R., Fagerlin, A. (2008). Communicating side effect risks in a tamoxifen prophylaxis decision aid: The debiasing influence of pictographs. *Patient Education and Counseling, 73*, 209-214.

Table 1. Sample size coefficients by numeracy level (high verses low), data type (statistics-only verses statistics+data), and use-N (told to use sample size verses no instructions) from Experiment 2.

Group/Condition	N^b (95% C.I.)
High Numeracy	
Statistics-only, no Use-N	.40 (.38-.43)
Statistics-only, Use-N	.42 (.40-.45)
Statistics+data, no Use-N	.41 (.39-.43)
Statistics+data, Use-N	.42 (.40-.44)
Low Numeracy	
Statistics-only, no Use-N	.29 (.25-.33)
Statistics-only, Use-N	.44 (.42-.46)
Statistics+data, no Use-N	.31 (.28-.34)
Statistics+data, Use-N	.41 (.38-.43)

Table 2. Low data percent range datasets used in Experiment 3. In the large N-small percent pairing sample sizes of 80 and 250 were paired with the smaller percentages than the sample sizes of 2, 3, or 5. In the large N-large percent pairing the opposite was true. Normatively one would judge the percentages according to their sample size. However, if sample size is completely ignored according to the encounter frequency model, then the opposite pattern would predicted across these datasets. Power model predictions are shown if subjects weigh sample size similarly to Experiments 1 and 2, where the average exponent was .32 for subjects not given raw data or instructions to use sample size.

Large N-small % pairing		Large N-large % pairing	
N	Percent	N	Percent
1	0%	1	0%
2	50%	2	0%
3	33%	3	0%
5	20%	5	0%
80	10%	80	11%
250	3%	250	24%
Normative, N^1	5%	Normative, N^1	20%
Encounter Frequency, N^0	19%	Encounter Frequency, N^0	6%
Predicted Power, $N^{.32}$	13%	Predicted Power, $N^{.32}$	12%

Table 3. High data percent range datasets used in Experiment 4. In the large N-small percent pairing sample sizes of 125 (or 1250 in the high magnitude condition) and 625 (or 6250) were paired with the smaller percentages than the sample sizes of 5 (50) or 25 (250). In the large N-large percent pairing the opposite was true. A normative judge would weigh each percentage in a dataset according to its sample size. In contrast, if subjects ignore sample size altogether, according the encounter frequency model, they would show the opposite pattern. Predictions are also shown if subjects weigh sample size according a power model with an exponent of .32 (the average exponent from Experiments 1 and 2 for subjects who were not given raw data or use-N instructions).

Large N-small % pairing		Large N-large % pairing	
N	Percent	N	Percent
1 (10)	100%	1 (10)	100%
1 (10)	100%	1 (10)	100%
1 (10)	100%	1 (10)	100%
1 (10)	100%	1 (10)	0%
5 (50)	100%	5 (50)	80%
25 (250)	96%	25 (250)	80%
125 (1250)	87%	125 (1250)	85%
625 (6250)	78%	625 (6250)	98%
Normative, N^1	80%	Normative, N^1	95%
Encounter Frequency, N^0	95%	Encounter Frequency, N^0	80%
Predicted Power, $N^{.32}$	88%	Predicted Power, $N^{.32}$	87%

Table 4. Best fitting power exponents across all four experiments and Obrecht et al.'s (2009) previous data broken down by sample size range and numeracy.

Experiment	Sample Sizes	Between Ss Manipulations	Power Exponent N^b (95% C.I.)	
			High Numerate	Low Numerate
Exp 1	1, 2, 5, 8, 10, 13, 16, 20, 27, 37	Data type	.37 (.36-.39)	.22 (.20-.25)
Exp 2	1, 2, 5, 8, 10, 13, 16, 20, 27, 37	Data type and use N	.41 (.40-.43)	.37 (.35-.38)
Exp 3	1, 2, 3, 5, 80, 250	N/A	.34 (.29-.40)	.22 (.17-.27)
Exp 4a	1, 1, 1, 1, 5, 25, 125, 625	N/A	.42 (.35-.48)	.28 (.22-.33)
Exp 4b	10, 10, 10, 10, 50, 250, 1250, 6250	N/A	.33 (.27-.39)	.21 (.16-.26)
Obrecht et al. (2009) Exp 3	10, 100, 750, 1,000, 10,000, 100,000	N/A	.13 (.09-.17)	-.01 (-.05-.03)

Figure 1. An example comparison pair from Experiments 1 and 2. Here sample size was 8, mean difference was high (i.e. 9-7=2, rather than 8-7=1), and standard deviation was low (i.e. 1.41, rather than 2.84). Subjects in the statistics-only condition were given just the statistical summary information displayed on the three number lines (shown above the dotted line). Subjects in the statistics+data condition were given both the number line and raw data representations.

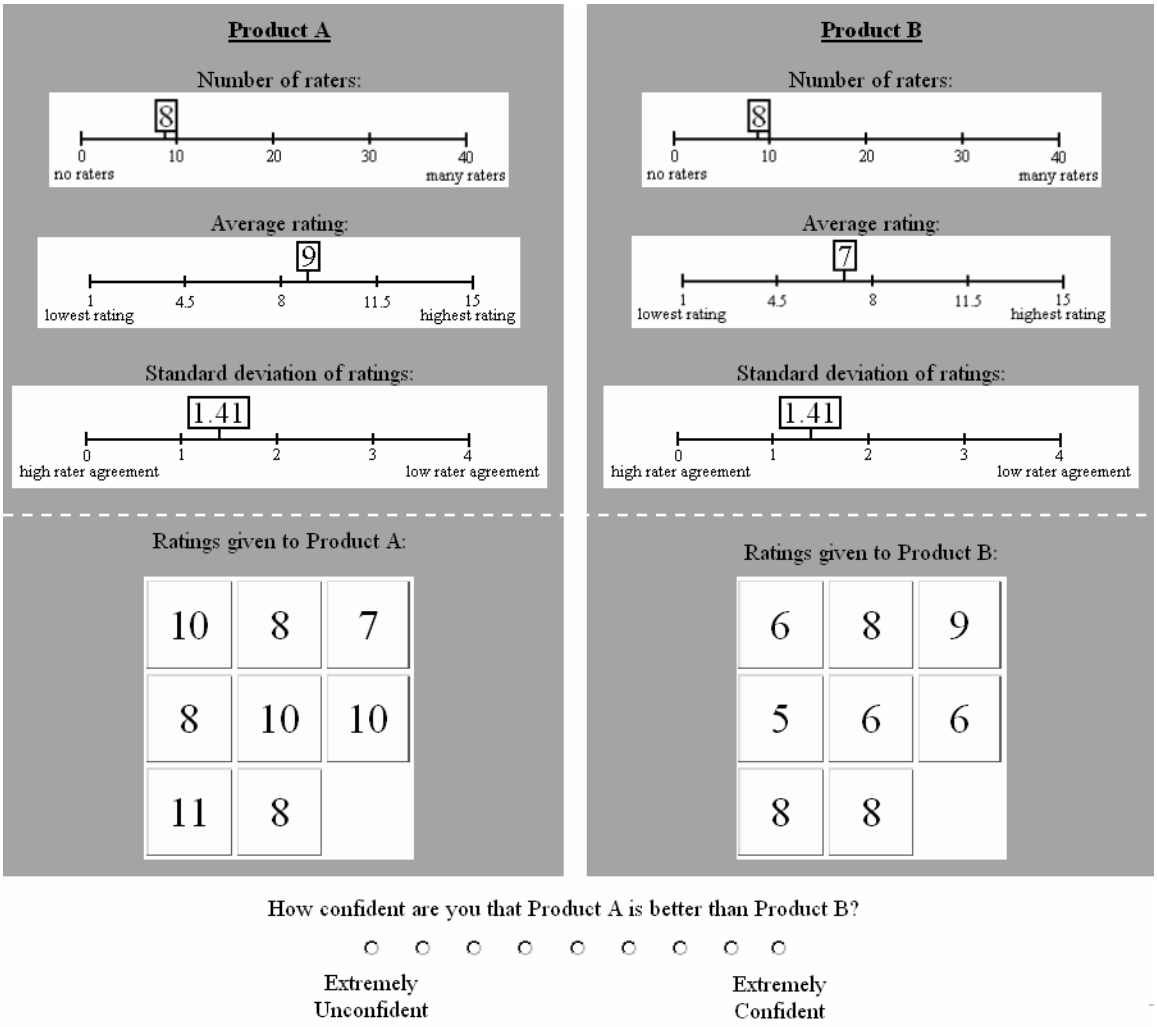
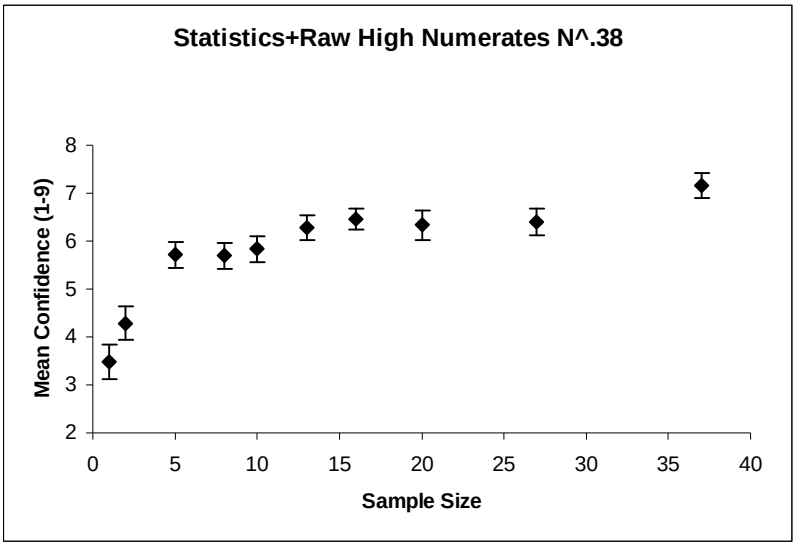
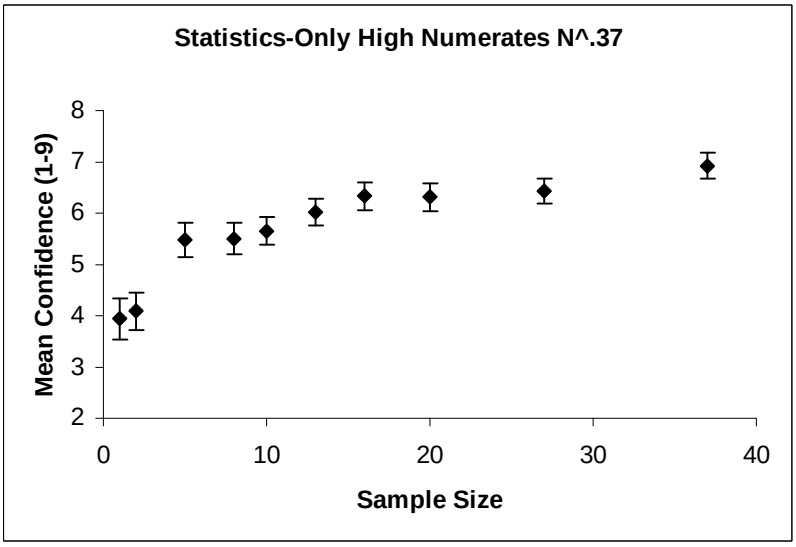
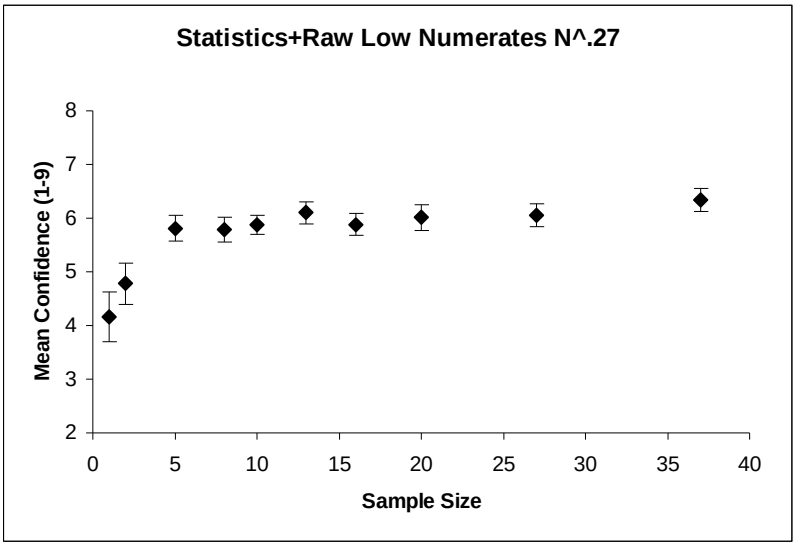


Figure 2. Subjects' confidence ratings in Experiment 1 as a function of sample size broken down by numeracy and data type. Higher numerates were more sensitive to sample size. Lower numerate subjects were more sensitive to sample size when they were given both statistical summaries and the corresponding raw data, rather than just the summary presentations.






Statistics+Raw Low Numerates N^.27
Mean Confidence (1-9)
8
7
6
5
4
3
2
0
5
10
15
20
25
30
35
40
Sample Size

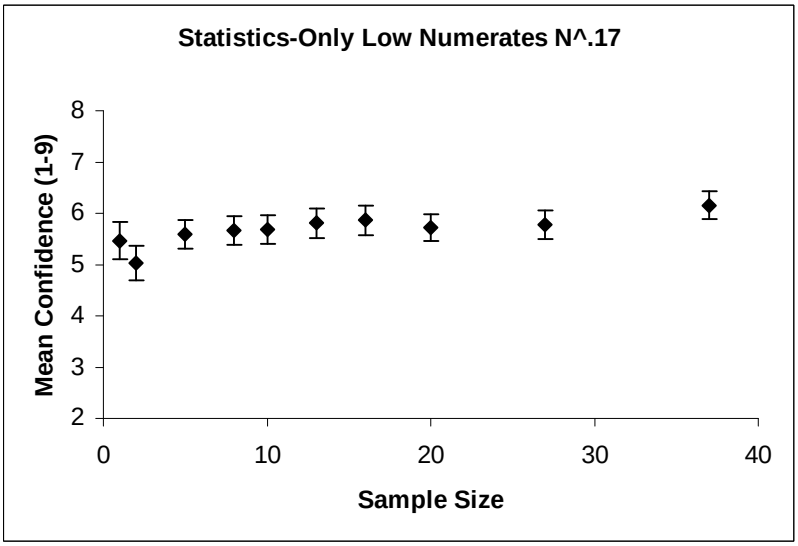
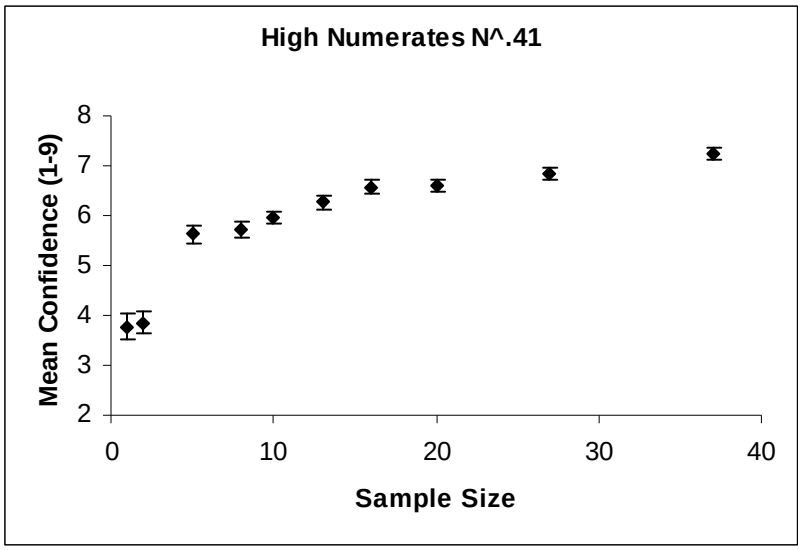
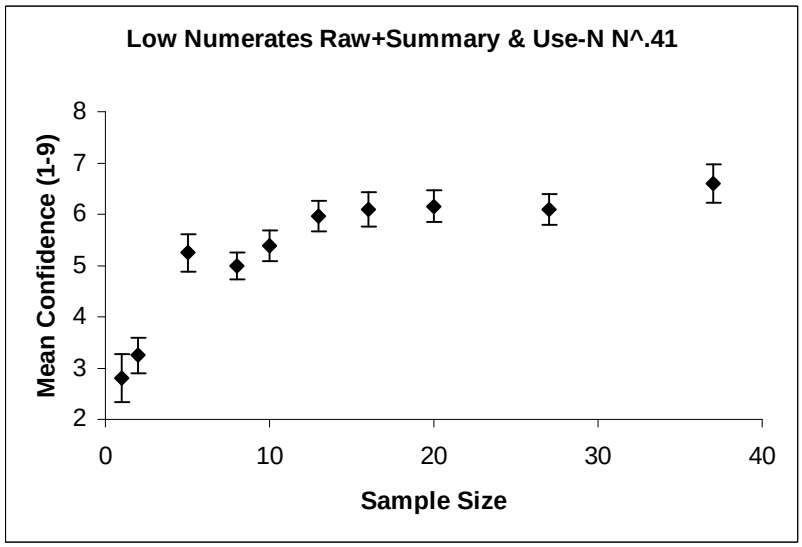

Statistics-Only Low Numerates N^.17
Mean Confidence (1-9)
8
7
6
5
4
3
2
0
10
20
30
40
Sample Size

Figure 3. Subjects' confidence as a function of sample size in Experiment 2. The first graph shows the high numerates' average ratings at each level of sample size collapsing across data type and use-N. The second graph shows low numerates' confidence ratings when provided with raw data and given instructions to use sample size. The final graph gives low numerates' ratings from both statistics-only conditions and the statistics+data condition in which subjects were not given the use-N instructions. Standard errors bars are in larger in graphs where fewer subjects' ratings are represented.





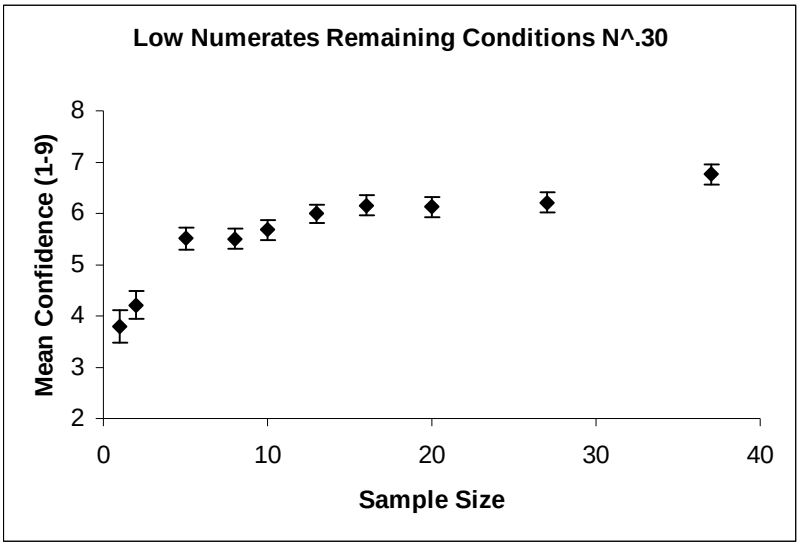
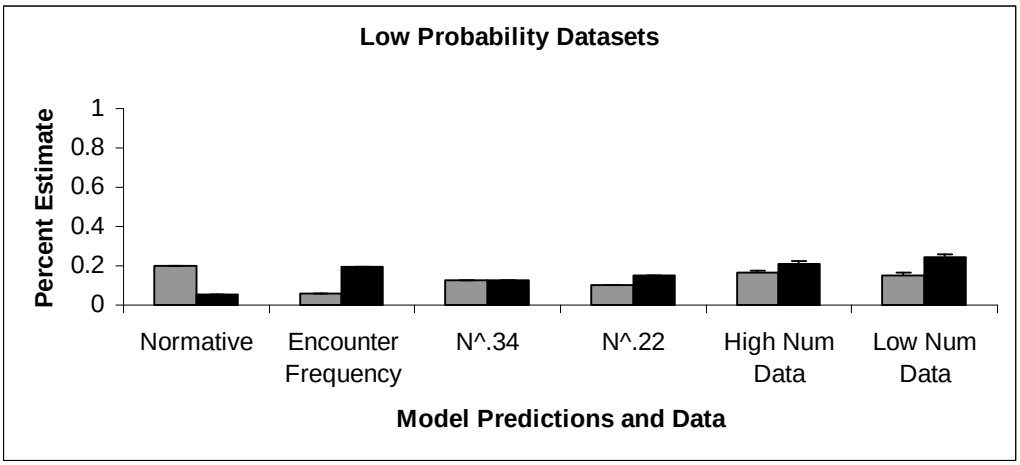
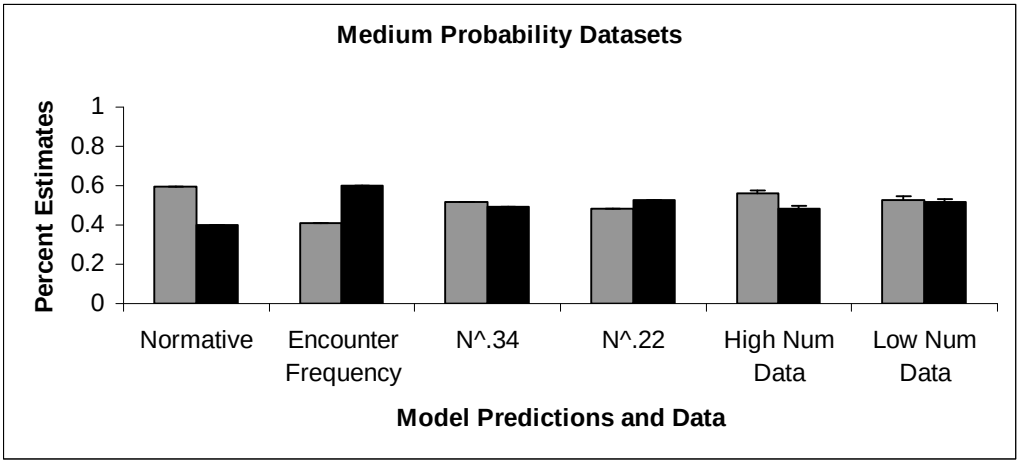

Low Numerates Remaining Conditions N^.30
Mean Confidence (1-9)
8
7
6
5
4
3
2
0
10
20
30
40
Sample Size

Figure 4. Gray and black bars represent the large N-large percent pairing, and the large N-small percent pairing conditions, respectively. Normative, encounter frequency, and power model predictions are shown in the first 4 pairs of bars in each graph. The last two pairs of bars show percent rating data from high and low numerate subjects in Experiment 3. Power model predictions are broken down for high and low numerate subjects. The first graph displays the predictions and data for the two low probability datasets. The second and third graphs display the predictions and subject data for the medium and high percent range datasets, respectively.




Medium Probability Datasets
Percent Estimates
1
0.8
0.6
0.4
0.2
0
Normative
Encounter Frequency
N^.34
N^.22
High Num Data
Low Num Data
Model Predictions and Data

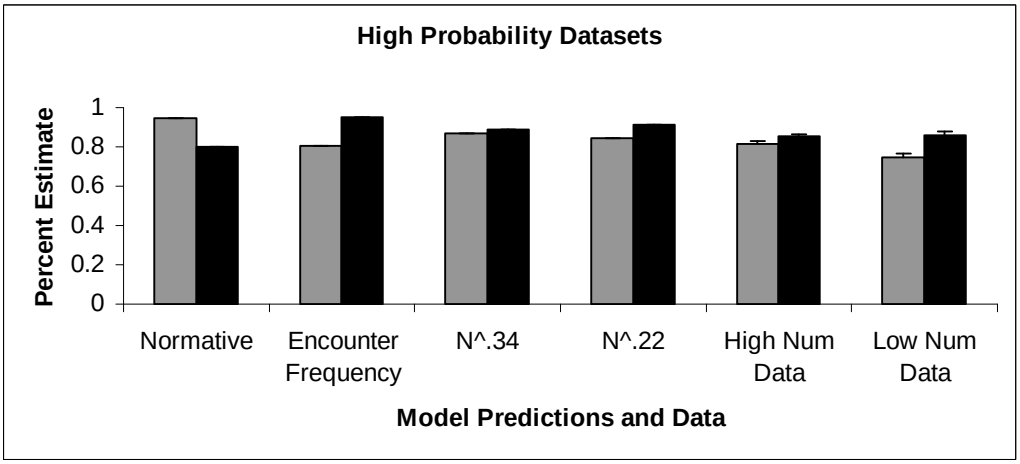
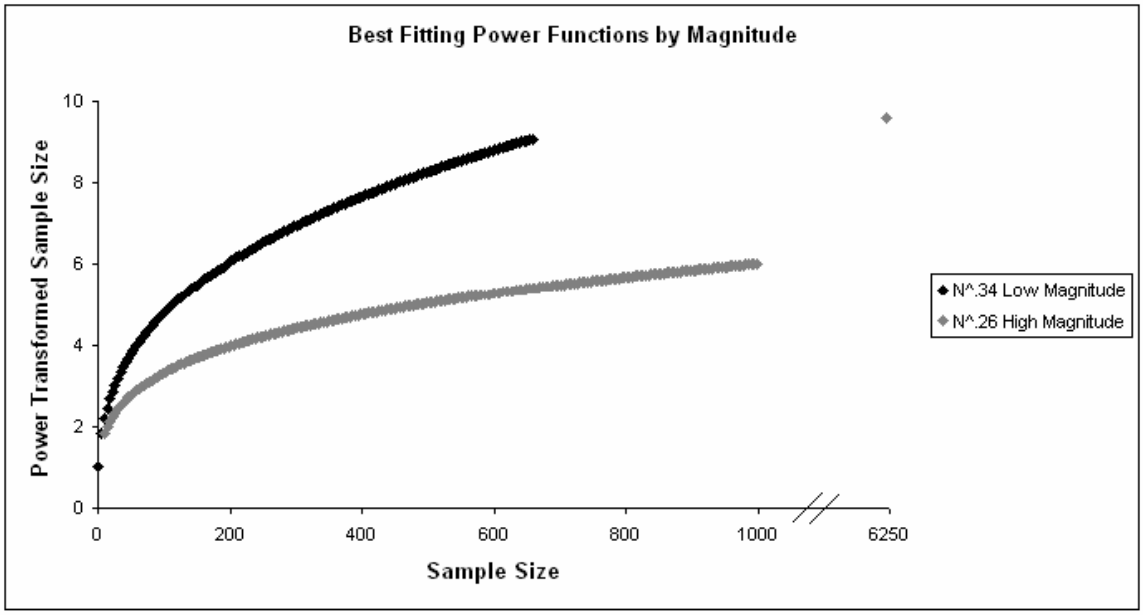

High Probability Datasets
Percent Estimate
1
0.8
0.6
0.4
0.2
0
Normative
Encounter Frequency
N^.34
N^.22
High Num Data
Low Num Data
Model Predictions and Data

Figure 5. Hypothetical power functions for subjects in the low verses high magnitude sample size conditions in Experiment 4. The low magnitude function is only shown up to 625 because that was the highest sample size that subjects considered in this group. Subjects in the higher magnitude group viewed sample sizes up to 6250; the x-axis is truncated to show this endpoint. The power function for the high magnitude condition starts at 10 because this was the lowest value shown to subjects in this condition.



Appendix: Numeracy scale

1. Imagine that we roll a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?

a) 500 b) 450 c) 200 d) 750

2. In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?

a) 100 b) 5 c) 1 d) 10

3. In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car?

a) .01% b) .001% c) .1% d) 1%

4. Which of the following numbers represents the biggest risk of getting a disease?

a) 1 in 100 b) 1 in 1000 c) 1 in 10000 d) 1 in 10

5. If Person A’s risk of getting a disease is 1% in ten years, and Person B’s risk is double that of A’s, what is B’s risk?

a) 20% in 10 years b) 2% in 10 years c) 1% in 1 year d) 2% in 5 years

6. If Person A's chance of getting a disease is 1 in 100 in ten years, and Person B's risk is double that of A, what is B's risk in ten years?

a) 2 in 50 b) 1 in 50 c) 2 in 200 d) 1 in 1000

7. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 100?

a) 1 b) 5 c) 100 d) 10

8. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?

a) 1000 b) 10 c) 100 d) 1

9. If the chance of getting a disease is 20 out of 100, this would be the same as having a ____% chance of getting the disease.

a) 20% b) 2% c) 5% d) 10%

10. The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?

a) 5 b) 2 c) 1 d) 50

CURRICULUM VITAE

Natalie Ann Lindemann

EDUCATION

2005	B.A., Psychology with a two-science minor in Biology and Chemistry, *magna cum laude*, departmental honors in Psychology, Oakland University, Rochester, MI
2007	M.S., Psychology, Rutgers: The State University of New Jersey, New Brunswick, NJ
2010	Ph.D., Cognitive Psychology and Certificate in Cognitive Science, Rutgers: The State University of New Jersey, New Brunswick, NJ

POSITIONS HELD

2005 – 2007	Fellow, The Graduate School – New Brunswick, Rutgers: The State University of New Jersey, New Brunswick, NJ
2007 – 2010	Fellow, National Science Foundation, Rutgers: The State University of New Jersey, New Brunswick, NJ

PUBLICATIONS

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t*-tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14,* 1147-1152.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition, 37,* 632-643.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning, 16,* 26-44.