UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Adults Are Sensitive to Variance When Making Likelihood Judgments

Permalink
https://escholarship.org/uc/item/57f8z3sw

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33)

ISSN
1069-7977

Authors
Chesney, Dana L.
Obrecht, Natalie A.

Publication Date
2011

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Adults Are Sensitive to Variance When Making Likelihood Judgments

Dana L. Chesney (Dana.Chesney.3@nd.edu)
University of Notre Dame, Department of Psychology, 118 Haggar Hall
Notre Dame, IN 46556

Natalie A. Obrecht (obrechtn@wpunj.edu)
William Paterson University, Department of Psychology, 300 Pompton Road
Wayne, NJ 07470 USA

Abstract

People have shown sensitivity to variance in studies in which variance has been provided separately from other statistical information, but not in other studies in which variance must be derived from raw data. However, such studies typically test people's sensitivity to variance via *probability* judgments: participants are asked to make judgments based on how confident they are that sample means are representative of a population. In this study, we instead investigate whether people are able to use variability when making *likelihood* judgments: participants determined from which of two possible populations a sample was more likely to have been drawn. Choices were influenced by variance, even when controlling for sample size, base rate, and the absolute difference between sample means and population μs.

Keywords: statistical inference; judgment; decision making; probability; normative behavior; variance; categorization

In this day and age, statistical information is widely available to any person with an internet connection. However, although people do have some statistical intuitions (e.g. Obrecht, Chapman, & Gelman, 2006), they do not make use of statistical factors - such as mean, sample size and variance - in a precisely normative fashion (Kahneman & Tversky, 1972). It remains in question how people's statistical intuitions affect their judgments. In this paper, we offer evidence that people are not only sensitive to the effect of variance on probability, but also that they can use variance information to judge the likelihood that a sample came from a particular population rather than another.

Both Obrecht, Chapman, and Gelman (2007) and Masnick and Morris (2008) examined whether people can use within group variability in statistical datasets when making judgments. In Obrecht et al.'s study, participants considered rating data given to pairs of hypothetical products. Participants saw not only the raw rating data but also were told the means and standard deviations of the data sets. Based on this information, participants judged their confidence that the product with the higher mean rating was really better than the product with the lower mean rating. Within group variability was manipulated across comparison pairs. Normatively, one should be more confident in a difference when within group variability is low, compared to when it is high. Indeed, they found that when the within group variability of the product ratings was low, participants were slightly, but significantly, more confident in a difference between groups, compared to when the variability was high. Masnick and Morris' study was similar in design to Obrecht et al.'s (2007). Child and adult participants compared pairs of datasets (e.g. a set of six throwing distances from two different players) so as to judge whether they differed from each other. However, in that study, means and standard deviations were not explicitly given. Masnick and Morris' participants failed to use within group variance in a normative fashion. In fact, their adult participants were sometimes actually *more* confident when within group variability was high.

Other studies have examined use of variability when such information comes from prior knowledge, rather than from sample data. Nisbett, Krantz, Jepson, and Kunda (1983) gave adult participants information about small samples that all shared some characteristic (i.e. three samples of an element are all conductive; three members of a tribe are all obese). Participants were willing to attribute the property found in the samples (e.g. conductivity, obesity) to a higher percentage of the general population from which the samples were drawn for cases referring to properties which have prototypically low variability (such as conductivity), rather than properties which are typically quite variable (such as people's weight). Jacobs and Narloch (2001) found that children also have the ability to use their prior knowledge of category variability to make inferences. Further, Obrecht, Chapmen, & Suárez, (2010) showed that people can combine statistical data with their prior knowledge about category variability to make reasonable inferences.

In many of the studies discussed, participants were asked to use statistical data to rate their confidence in an answer already given (Masnick and Morris, 2008; Obrecht, et al. 2007; Obrecht, et al. 2010) or to make inferences about the general population from sample data (Nisbett et al., 1983; Jacobs & Narloch, 2001; Obrecht, under review). Such contexts supply a great deal of information to the participants. For instance, participants are told which populations the samples came from. Additionally, Masnick and Morris (2008), Obrecht, et al. (2007), and Obrecht, et al. (2010) all informed participants that samples came from different populations. Further, which population is more likely to be "better" (better player, better product, etc.) was often already implied by the difference between the sample means and by the phrasing of the questions participants were asked.

It might also be noted that in the work described above, participants only appeared sensitive to variability when such information was provided separately from the raw data, whether given explicitly in the tasks (Obrecht et al., 2007, Obrecht et al. 2010) or implied via prior category knowledge (Jacobs & Narloch, 2001; Nisbett et al., 1983, Obrecht et al. 2010). When Masnick and Morris (2008) did not provide variance information explicitly, but rather left it to be derived from raw data, their participants did not respond at all normatively to within group variability. Similarly, Kahneman and Tversky (1972) found that people failed to reason normatively about variance in *dichotomous* data (e.g. male vs. female), for which variability is not an independent parameter, but rather is a function of sample size and the percentage of the group sharing a particular feature. Participants in their study typically claimed that large and small hospitals would have about the same number of days in a year in which more than 60% of the children born were male. Normatively, a small hospital will have more days when more than 60% of children are born male, given a dichotomous population (male vs. female) with a μ of 50% (about half of children born are male). However, using variance information in Kahneman and Tversky's task was not a simple proposition: One must combine the information that the μ of male births is 50% and that the number of births was higher at the larger than the smaller hospital and also note that means of smaller samples typically depart farther from population μs than means of larger samples before one can use variance to determine the relative likelihood of a sample with a particular mean coming from either population. It is likely that this complexity contributed to people's non-normative behavior (e.g. Evans & Dusior, 1977).

Here, we further explored the question of whether people can make use of variance when determining which of two populations they believe a sample was more likely to have been drawn from. Like Kahneman and Tversky (1972), we tested whether people's choices took variance - as implied by sample size and the percentage of a group sharing a dichotomous feature - into account. However, unlike Kahneman and Tversky, we held sample size constant and instead manipulated variability via population μs and sample means. Also, we asked participants to consider the likelihood of a *particular* sample having come from one population or another, rather than to reason about a *range* of possible samples.

Method

The purpose of this study was to see whether people can use variance information when determining whether a sample was more likely drawn from one population or another. As has been done for a number of prior studies (Kahneman & Tversky, 1972; Nisbett et al., 1983; Obrecht, under review), we chose to use a dichotomous feature as the basis of comparison. This allowed the variance of a sample to be determined solely from sample size and the proportion of the sample exhibiting that dichotomous feature:

$$\sigma^2 = np(1 - p)$$

Participants were given information about two different populations of trees (i.e. 2% of Aoco trees have white flowers, 18% of Boco trees have white flowers) and a sample (i.e. 10% of trees in a grove of 100 trees have white flowers). They were asked to indicate which population the sample was more likely to have come from, and how sure they were of their choice.

Participants

The participants in this study were 266 undergraduate students at the University of Notre Dame participating for course credit. Of these, 40 were excluded for failure to complete the task. An additional 5 were excluded from the analysis for failure to complete the task within a reasonable time period (taking either less than five minutes or more than two hours).

Design

This study used a within subjects design. Every participant was asked 48 pairs of questions. Every set of questions involved comparing a sample mean to a pair of population μs. In each pair of population μs, one was more central (closer to 50%) and one was more extreme (farther from 50%). Sample means always fell between the population μs. We manipulated:

a) **Centrality: Whether the population μs were Central or Extreme** When comparing conditions where the absolute difference between the population μs is the same, the relative likelihood of samples coming from the population with the more central μ as opposed to the more extreme μ was greater when the population μs and sample means were from a more Extreme range (e.g. 18% vs. 2%) rather than from a more Central range (e.g. 48% vs. 32%). This is because populations from the Central range have inherently higher variance ($\sigma^2 = np(1 - p)$). This manipulation allowed us to vary the relative likelihood of a sample being produced by either population independently of the difference between the sample means and the population μs**.**

b) **Spread: Whether the spacing between population μs was Narrow or Wide** Population μs either differed from each other by 16 percentage points in the Narrow condition (e.g. 2% vs. 18%) or 29 percentage points in the Wide condition (e.g. 2% vs. 31%).

c) **Parity: Whether the population μs were Low or High** Populations with μs equally distant from 50% (e.g. 25% & 75%, 10% & 90%) are also equally variable. Thus when constructing questions sets, the values of population μs and sample means were reflected under (Low) and over (High) 50%. For example, if one question referred to populations μs of 2% and 18%, with a sample mean of

Table 1: Population μs and sample means used in constructing stimuli, with their absolute and relative probabilities

Population Centrality, Spread, and Parity	Central μ (SD)*	Extreme μ (SD)*	Sample %s (Location)	*P* of sample %s given Central μ	*P* of sample %s given Extreme μ	Ratio of *P*s of sample %s ($P_{Central}/P_{Extreme}$)
Extreme/Narrow/Low	18%	2%	8% (Extreme)	2.4×10^{-3}	7.4×10^{-4}	3.3
	(14.76)	(1.96)	10% (Mid)	1.1×10^{-2}	2.9×10^{-5}	3.8×10^{2}
			12% (Central)	3.2×10^{-2}	7.3×10^{-7}	4.4×10^{4}
Extreme/Wide/Low	31%	2%	16% (Extreme)	2.8×10^{-4}	1.6×10^{-10}	1.8×10^{6}
	(21.39)	(1.96)	20% (Mid)	4.6×10^{-3}	1.1×10^{-14}	4.1×10^{11}
			24% (Central)	2.8×10^{-2}	2.9×10^{-19}	9.7×10^{16}
Central/Narrow/Low	48%	32%	38% (Extreme)	1.1×10^{-2}	3.7×10^{-2}	2.9×10^{-1}
	(24.96)	(21.76)	40% (Mid)	2.2×10^{-2}	2.0×10^{-2}	1.1
			42% (Central)	3.9×10^{-2}	9.0×10^{-3}	4.4
Central/Wide/Low	51%	22%	36% (Extreme)	8.7×10^{-4}	5.2×10^{-4}	1.7
	(24.99)	(17.16)	40% (Mid)	7.1×10^{-3}	2.3×10^{-5}	3.1×10^{2}
			44% (Central)	3.0×10^{-2}	5.2×10^{-7}	5.2×10^{4}
Extreme/Narrow/High	82%	98%	92% (Extreme)	2.4×10^{-3}	7.4×10^{-4}	3.3
	(14.76)	(1.96)	90% (Mid)	1.1×10^{-2}	2.9×10^{-5}	3.8×10^{2}
			88% (Central)	3.2×10^{-2}	7.3×10^{-7}	4.4×10^{4}
Extreme/Wide/High	69%	98%	84% (Extreme)	2.8×10^{-4}	1.6×10^{-10}	1.8×10^{6}
	(21.39)	(1.96)	80% (Mid)	4.6×10^{-3}	1.1×10^{-14}	4.1×10^{11}
			76% (Central)	2.8×10^{-2}	2.9×10^{-19}	9.7×10^{16}
Central/Narrow/High	52%	68%	62% (Extreme)	1.1×10^{-2}	3.7×10^{-2}	2.9×10^{-1}
	(24.96)	(21.76)	60% (Mid)	2.2×10^{-2}	2.0×10^{-2}	1.1
			58% (Central)	3.9×10^{-2}	9.0×10^{-3}	4.4
Central/Wide/High	49%	78%	64% (Extreme)	8.7×10^{-4}	5.2×10^{-4}	1.7
	(24.99)	(17.16)	60% (Mid)	7.1×10^{-3}	2.3×10^{-5}	3.1×10^{2}
			56% (Central)	3.0×10^{-2}	5.2×10^{-7}	5.2×10^{4}

*Standard deviation of the sampling distribution from a dichotomous population with the given μ where N = 100.
Note: *P* refers to probability of drawing a sample (N=100) with a given % from a population with a given μ: *P*(sample % | μ).

10% (a low population parity question), another question referred to population μs of 98% and 82% with a sample mean of 90%. This allowed us to balance whether the μ of the population that the sample was more likely to have been drawn from was greater or less than the mean of the sample (see Table 1.) High Parity conditions might be thought of as negative parity versions of Low Parity conditions: "98% of Doco mango trees *have* white flowers" is logically equivalent to "2% of Doco mango trees *do not have* white flowers."

Control: Sample % locations Three different sample percentages were presented with each of the eight pairs of μs: one closer to the extreme μ (Extreme), one closer to the central μ (Central), and one half way between the other two sample means (Mid). (See Table 1.)

Additional controls For half of the trials the population with the more central μ was described first, while for the other half the population with the more extreme μ was described first. This yielded 6 questions sets per pair of μs (see Table 1). The question sets were presented in random order. Sample size and base rate were also controlled. Participants were always told that there were 100 trees in the grove and that groves of either population occurred with equal frequency. Additionally, the relative and absolute likelihoods of μs in the Wide & Central conditions were matched as closely as possible using μs described by whole number percentages to the relative and absolute likelihoods of μs in the Narrow & Extreme conditions (see Table 1). This allowed us to manipulate both variance and the absolute difference between population μs while controlling the relative and absolute probability of a sample being produced by either population.

Procedure

This study was conducted online. Participants signed up via a university system, and followed a link to a web page that included the following text:

> "In this study you will be given information about different types of trees. For example, Ukon cherry trees tend to have yellow blossoms. In contrast, Kanzan cherry trees tend to have pink blossoms. Suppose you see a grove where someone planted either all Ukon or all Kanzan trees. If you did not know which kind of tree was planted, you could use the color of the blossoms in the grove to make an inference. For example, if the blossoms were mostly yellow, you might guess that Ukon, rather than Kanzan, trees were planted.
>
> In this study you will be asked to make inferences

about which of two types of trees seems more likely to have been planted in a grove based on the percent of blossoms that are a certain color."

After viewing this text, participants followed a link to a survey made up of 48 pairs of questions, presented in random order, constructed using the sets of population μs and sample percentages described in Table 1. Each pair of questions was preceded by an information block, like the one below:

"Mango trees can have either white or yellow flowers.
2% of Aoco mango trees have white flowers.
18% of Boco mango trees have white flowers.
There are equal numbers of Aoco and Boco groves.
You see a grove of 100 mango trees. This grove consists of either all Aoco trees or all Boco trees. You see that 8% of the trees have white flowers."

Participants were then asked to indicate which kind of grove this was more likely to be. For example, "Is this more likely to be an Aoco mango grove or Boco mango grove?" They are also asked to rate on a scale of 1-7 how sure they were of their answer, where 1 means "no idea" and 7 indicated one was "completely sure."

After completing these 48 sets of questions, participants completed a 10 question multiple choice numeracy evaluation similar to that used by Obrecht et al. (2007) that required conversions between percentages, proportions and frequencies. They were also asked their math and verbal SAT scores, as well as what math and/or statistics classes they had taken or were currently taking.

Analysis and Results

For dichotomous features, the variance of a population ($\sigma^2 = np(1 - p)$) becomes smaller as the proportion of the population exhibiting that feature becomes increasingly distant from 50%. There is 0 variance in populations where the percent of the population exhibiting a feature is 0% or 100% and maximal variance in populations where that percentage is 50%. As a result, it is more likely that, for example, a population for which 82% of trees have white flowers would produce a sample of 100 trees where 90% have white flowers, than that a population for which 98% of trees have white flowers would do so. Similarly, it is more likely that a population for which 18% of trees have white flowers would produce a sample of 100 trees where 10% have white flowers, than that a population for which 2% of trees have white flowers would do so. Thus, in our analysis, we coded participants' responses by whether or not they indicated that the grove was more likely to have come from the population with the more central population μ (close to 50%), as this allowed us to most directly compare participants' behavior across Low and High Parity conditions.

As can be seen in Table 1, samples were more likely to have come from populations with more central μs when μs were Extreme rather than Central, due to differences in variance. However, samples were more likely to have come from populations with more central μs when μ Spread was Wide rather than Narrow, due to the differences in means as well as the variance. Further, Parity has no effect on likelihood. Thus, effects of Centrality can be attributed to the normative influence of variance, effects of Spread can be attributed to the normative influence of both variance and mean difference, and effects of Parity are not normative.

We ran a 2 x 2 x 2 repeated measures ANOVA where the factors were a) Centrality, b) Spread, and c) Parity. Participants were asked 6 forced choice questions (3 sample % locations x 2 presentation orders) regarding each of the 8 pairs of population μs. Thus, they were given scores corresponding to the proportion of these 6 questions for which they responded that the sample was more likely to have been drawn from the population with the more central μ for each of these 8 population conditions.

Variance influenced likelihood judgments

We found that individuals were more likely to choose the population with the more central μ in the Extreme than the Central conditions ($F_{(1,205)} = 48.4$, $p < .0005$, $\eta_p^2 = .18$), indicating that indeed, variance influenced participants' likelihood judgments. Results are displayed in Figure 1. A non-parametric test further supports this conclusion. Each participant was asked 24 pairs of questions that only differed in centrality (Spread, Parity, sample % location, and presentation order being otherwise matched). Of the 221 participants that answered all the forced choice questions, 137 gave more central answers when μs were extreme, while 42 gave more central answers when μs were central, and 42 showed no difference. This is significant by a binomial test ($p<.0005$).

Non-normative influence of Parity

The more central μ was chosen more often in the Low than the High Parity conditions ($F_{(1,205)} = 159.106$, $p < .0005$, $\eta_p^2 = .42$) Participants were also more strongly affected by Centrality when Parity was Low ($F_{(1,220)} = 13.777$, $p < .0005$, $\eta_p^2 = .06$).

Mean difference influenced likelihood judgments more strongly than variance

The more central μ was chosen more often in the Wide than the Narrow Spread conditions ($F_{(1,220)} = 259.3$, $p < .0005$, $\eta_p^2 = .54$), an effect that can be attributed to a sensitivity to mean difference as well as variance. Results are displayed in Figure 1. The initial analysis was followed up with a test to see whether relative influence of a) variance and b) absolute differences between population μs and sample means on people's choices was normative. We ran a 2 (Narrow & Extreme vs. Wide & Central) x 2 (Parity) repeated measures ANOVA that compared data compiled from cases where the population μs were both Central and Wide to data compiled from cases where the population μs were both Extreme and Narrow. As previously mentioned, absolute and relative probabilities were closely matched in these conditions. Thus, normatively, no effect of the Narrow & Extreme vs. Wide & Central condition should be expected. However,

though in Narrow conditions sample percentages were, on average, equidistant from the more central and the more extreme population μs, in Wide conditions sample percentages were, on average, closer to the more central population μ. Thus, if the participants were being more strongly influenced by these mean differences than by variance, they would tend to give more central answers in the Wide & Central than the Narrow & Extreme conditions. Indeed, this effect was observed ($F_{(1,220)} = 79.8$, $p < .0005$ $\eta_p^2 = .27$). The effect of Parity also remained significant ($F_{(1,220)} = 132.9$, $p < .0005$, $\eta_p^2 = .38$). Further, the effect of Narrow & Extreme vs. Wide & Central populations was stronger in the High Parity conditions ($F_{(1,220)} = 10.0$, $p < .005$, $\eta_p^2 = .04$). These findings are in line with previous research (Obrecht et al., 2007; Obrecht, under review) indicating that differences in means have a stronger influence on people's decisions than differences in variance.

Individual differences

There is literature (e.g. Nisbett et al. 1983, Obrecht et al., 2007) suggesting that individual differences in statistical training and numerical knowledge influence how people make use of statistical information. Thus, we performed a subsequent 2 x 2 x 2 repeated measures ANCOVA including as a covariate whether or not participants had a perfect score on the numeracy evaluation (40% had a perfect score). Participants who had perfect scores on the numeracy scale were more likely to respond that a sample was drawn from the population with the more central μ; that is, these participants gave more normative responses compared to those who scored lower on the numeracy measure ($F_{(1,219)} = 4.4$, $p < .05$, $\eta_p^2 = .02$). Further, such participants were more strongly influenced by numerical factors that affected likelihood (interaction between Numeracy and Spread: $F_{(1,219)} = 8.3$, $p < .005$, $\eta_p^2 = .04$; interaction between Numeracy and Centrality: $F_{(1,219)} = 3.4$, $p < .07$, $\eta_p^2 = .015$, marginally significant), and less strongly influenced by factors that did not affect likelihood (interaction between Numeracy and Parity: $F_{(1,219)} = 7.6$, $p < .01$, $\eta_p^2 = .03$). It should be noted that effects of Centrality, Parity, and Spread, as well as the interaction between Centrality and Parity remained significant when having a perfect score on the numeracy evaluation was included as a covariate (all *p*s < .0005). Results are displayed in Figure 1.

These findings may be taken as evidence that, as suggested by Obrecht et al. (2007), more numerate individuals are more strongly influenced by numerical factors that affect probability than less numerate individuals. Individuals who scored perfectly on the numeracy evaluation had slightly higher mean SAT scores compared to those who made errors (727 (SD = 45) vs. 689 (SD = 70); $t = 4.25$, $p < .005$; 31 participants did not report SAT scores.) One point of note however is that a second ANCOVA found no effect of having taken a statistics course when this factor, rather than numeracy, was included as a covariate ($F_{(1,215)} = .01$, $p > .9$). It appears that statistical training did not boost performance on this task.

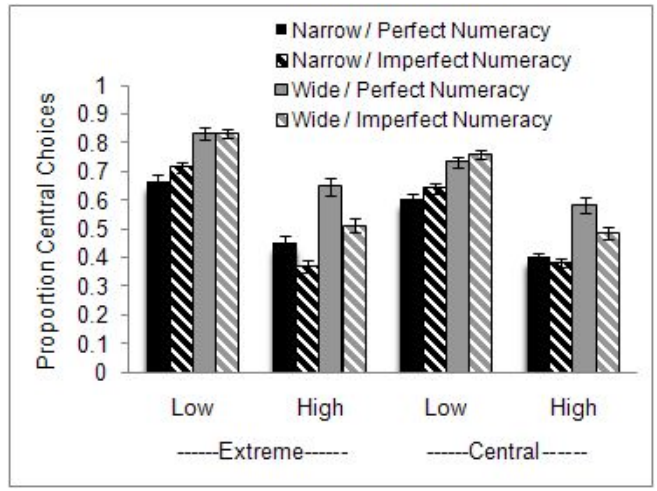

Figure 1: Mean proportion of central responses from participants with perfect and imperfect numeracy scores for questions referring to the eight different population pairs (see Table 1). Error bars represent standard error.

Discussion

These results indicate that participants' decisions were influenced by both normatively relevant and irrelevant numerical factors. As would be normatively correct, participants tended to indicate that the samples came from the population with the more central μ more often when the spacing between the populations was wide, rather then narrow, and when the populations were from a more extreme, rather than more central range. This offers evidence that a) people have some sensitivity to the likelihood of a sample being drawn from a given population and b) people can use variance when making such determinations. Although variance shifts people's behavior in the normative direction, one cannot call their behavior precisely normative. An ideal observer would have selected the central population in all of the extreme population range trials and 80% of the central population range trials. This was not the case (see Figure 1).

Further, even though Parity manipulation, normatively, should not have affected participants' judgments, participants were less likely to choose the population with the more central μ when the population μs were in the High Parity condition. There are different possible interpretations for this effect. First, people have more difficulty performing categorical reasoning about information with negative parity (e.g. *not red*) than with positive parity (e.g. *red*) (see Feldman, 2000). Second, while the *absolute* differences between the sample means and population μs were matched across conditions with values from the High and Low Parities, the relative differences were not: 2 is 16 away from 18, and 82 is 16 away from 98, but 2/18 is not equal to 82/98. People's ability to discriminate between numerical magnitudes is based on the relative, not the absolute, difference between them (Gallistel & Gelman, 2005). It is possible that performance was less normative for High vs.

Low Parity trials because the smaller relative differences between values used in High Parity trials made it more difficult for participants to discriminate between them and, subsequently, the probabilities they conveyed.

Regardless of whether their use of variance information is precisely normative, these data show that people have some sensitivity to the probability that a sample with a particular mean might be drawn from a given population. This has implications for the interpretation of results from prior studies on normative use of statistical information. In a natural context, an observer is not in a position to assume that the samples they have information about are necessarily representative of the general population. Further, difference between means, standard deviations, and sample sizes can themselves convey information about the likelihood that sets were sampled from the same general population. Thus it is possible that some departures from "normative behavior" can be attributed, at least in part, to a sensitivity to the likelihood of the samples being randomly representative of the same general population.

Consider, for example, the results of Obrecht, Chapman, and Gelman (2009). In this study participants were asked to make judgments about whether a particular kind of radio would break. They were told that a study found that 30 out of 1000 radios tested (3%) broke within a year. Some participants were also given sets of reports from individuals who owned that kind of radio of whom 2 out of 4, 3 out of 4, 8 out of 16, or 12 out of 16 reported that the product broke. This study, like others before it (see Kahneman & Tversky, 1972), found that people did not use sample size in a normative fashion: participants gave more credence to the individual reports than they should have, given the much larger sample size of the radio study. In other words, participants did not weigh means by sample size, as the authors considered to be normatively correct. However, while it is typically considered normative to weight set means by sample size, this assumes that these data represent a random sampling of the population. Consider if, for example, the 1000 tested radios were made at a factory in Manhattan, but individuals' reports were from owners of radios made in Nebraska. Lacking further information about the proportion of Manhattan and Nebraskan made radios in the general population, it may be reasonable to simply average these sample means without regard to sample size. The savvy statistician may determine that the randomness of the sampling was in question just from looking at the numbers: With a population μ of 3%, there would be less than a 1% chance of even 2 out of 4 sampled radios breaking, and less than a .0000001% chance that 12 out of 16 would break. It would be quite legitimate for participants to conclude that the radios tested in the study were different then those that the customers were buying.

Individuals may similarly be able to use statistical information to determine how likely it is that samples are drawn randomly from the same population (in which case weighting means by sample size is normative) or instead discretely sampled unknown subpopulations (in which case it is reasonable to average sample means without regard to sample size). A prerequisite to this ability is that people be sensitive to the likelihood that a sample is representative of a particular population. Our results indicate that people are indeed sensitive to such likelihoods. This interpretation is supported by the results from Obrecht (under review). Obrecht found that participants are more likely to weight means by their sample sizes when the sample sizes are smaller, rather than larger: The probability of a population producing samples with divergent means is lower when the sample sizes are larger, thus averaging means without weighting by sample size would be more normative in the higher than the lower sample size conditions. We are currently conducting a series of studies to further determine if people's judgments about statistical information are influenced by likelihood in such a fashion.

Acknowledgments

We thank A. Champan, R. Gelman, P. Mathews, N. McNeil, and L. Peterson for their help and support.

References

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

Feldman, J. (2000) Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630-633.

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.

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.

Masnick, A. M. & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79,* 1032-1048.

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. (under review). Sample size weighting in probabilistic inference.

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.