The Intuitive Conceptualization and Perception of Variance

Elizaveta Konovalova[1,2] and Thorsten Pachur[2]

[1]Warwick Business School, University of Warwick, Coventry, United Kingdom

[2]Center for Adaptive Rationality,Max Planck Institute for Human Development, Berlin, Germany

Author Note

Elizaveta Konovalova, ORCID: 0000-0002-3299-6450, Warwick Business School, University of Warwick, Coventry, CV4 7AL, United Kingdom; Thorsten Pachur, ORCID: 0000-0001-6391-4107, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany.

The authors declare no conflicts of interest in preparing this article.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Upon publication, data and codes can be found at https://osf.io/7h8pd/.

Correspondence concerning this article should be addressed to Elizaveta Konovalova, Warwick Business School, University of Warwick, Coventry, CV4 7AL, United Kingdom. Email: elizaveta.konovalova@wbs.ac.uk.

Abstract

Statistical concepts (e.g., mean, variance, correlation) offer powerful ways to characterize the structure of the environment. To what extent do statistical concepts also play a role for people assessing the environment? Previous work on the mind as "intuitive statistician" has mainly focused on the judgment of means and correlations (Peterson & Beach, 1967). Much less is known about how people conceptualize and judge variance. In a survey and three experimental studies, we explored people's intuitive understanding of variance as a concept and investigated the factors affecting people's judgments of variance. The survey findings showed that most people hold concepts of variance that they can articulate; these concepts, however, reflected not only statistical variance (i.e., deviations from the average) but also the pairwise distance between stimuli, their range, and their variety. The experimental studies revealed that although people's judgments of variance are sensitive to the statistical variance of stimuli, variety and range also play an important role. The results can inform psychological models of judgments of variance.

Keywords: variability, variance, judgment, intuitive statistics

The Intuitive Conceptualization and Perception of Variance

1. Introduction

Making correct inferences about properties of the environment—such as the central tendency and variability of a variable, or the co-occurrence of events—is key for making good decisions. Consider a foraging animal. To decide which patch of land to visit, the animal needs to make a prediction about where it will find sufficient food. Let us assume that patch A gives just enough food most of the time (the central tendency is low and the variability is low). In contrast, patch B usually gives a lot of food but sometimes none at all (the central tendency is high and the variability is high). In this case, the prediction would be more certain for A than for B, but the predicted value would be higher for B than for A. To ensure it finds enough food to survive, the animal needs to balance the central tendency and the variability of different food patches. Similar concerns apply to a broker picking stocks.

Statistical concepts such as mean, variance, and correlation offer a powerful characterization of the environment. An influential approach in the judgment and decision making literature has proposed that people are "intuitive statisticians"—they can judge the statistical properties of the world reasonably accurately and their responses to the environment can be well described based on statistical concepts (e.g., Peterson & Beach, 1967). In research on the mind as an intuitive statistician, perceptions of sample means (for a review, see Peterson & Beach, 1967; Pollard, 1984) and correlations between variables (Kareev, 2004) have received much attention; however, little is known about how people conceptualize and perceive variance of stimuli in the environment, and whether their perception of variance is in line with the statistical notion. This is an important gap, as several influential models of decision making implicitly assume that people are sensitive to statistical variance (Markowitz, 1952; Weber, Shafir, & Blais, 2004).

From the scant amount of work on variance perception, Peterson and Beach (1967) concluded in their seminal overview that people's judgments largely align with statistical variance and that statistical variance could, therefore, serve as a good first approximation for describing people's judgments of variability in the environment. However, the judgments tend to be

systematically lower than statistical variance (see also Kareev, 2004), suggesting that people may not rely on statistical variance directly but on other, qualitatively different aspects of the sample when judging variance—either with or without considering statistical variance.[1] But what are the aspects feeding into people's concept of variance?

Our first goal was to explore people's intuitive understanding of "variance". Do they hold an explicit concept of "variance", and if so, to what extent does this concept converge with or deviate from statistical variance? Even if people (who have not been trained in statistics) may be unlikely to spontaneously come up with the statistical definition of variance (i.e., the expected squared difference of a variable from its mean), they might still produce a concept that reflects the principle underlying statistical variance, namely that of deviation of individual values in a sample from the average value. But people may also have qualitatively different concepts, that reflect other possible principles guiding conceptualizations of variance (whose quantitative measures, however, might still be correlated with statistical variance). Based on semantic analyses of people's descriptions of their concept of "variance" as well as their associations with the term as collected in a survey, we established that most people hold a concept of variance that they can articulate, even if they are not trained in statistics. In fact, about a fifth of the participants articulated a concept closely related to statistical variance (namely, deviation from the mean); this was more likely for people with more education. Other participants articulated a concept that drew on other aspects, such as range, variety (the number of unique elements in a sample), and pairwise distance (the difference between pairs of values in a sample)—concepts that are based on principles that deviate from that of statistical variance.

Our second goal was to test whether the concepts we identified in the survey are indeed implicated in people's judgments of variance and to compare the influence of these factors with that of other factors previously shown to impact judged variance. One such factor is the stimulus magnitude. Research on number cognition and in risky choice has found that people perceive

---

[1] Note that Weber et al. (2004) found that people's risky decisions were more consistent with the coefficient of variation than with statistical variance. The coefficient of variation is, however, based directly on statistical variance.

differences between larger numbers as smaller than if the same differences are between smaller numbers (Lathrop, 1967; Weber et al., 2004). As a consequence, the larger the absolute (mean) magnitude of the values, the smaller a given level of variability appears to be—leading to a negative correlation between the mean value of the stimuli and perceived variability. Another factor is the sample size—the number of observations to which a variance judgment refers. In research on the perception of group variability, a higher sample size of interactions with in-group members led to judgments of the in-group as being more variable (Linville, Fischer, & Salovey, 1989). Similar findings were reported for the perception of numbers and objects (Kareev, 2004; Konovalova & Le Mens, 2018).

Although previous research has identified several factors that may play a role in judgments of variance, these factors have never been tested together within the same analysis. This limits previous conclusions because, for many types of distributions, range, sample size, and statistical variance are highly correlated. For instance, sets of stimuli covering a larger range of values usually also have larger variance than sets with a smaller range. The effects of these factors on judgments of variance may thus be confounded.

In the following, we first report a survey (Study 1) in which we used semantic network analysis to investigate people's intuitive concepts of variance. Specifically, we analyzed word co-occurrences in people's free descriptions of and associations with the term "variance." We then outline three experimental studies in which we investigated the effects of several factors that potentially impact people's intuitive judgments of variance. Using both an approach with tightly controlled stimulus properties (Studies 2a and 2b) and an approach with naturally varying stimulus properties (Study 3), we examined the role of concepts identified in Study 1 as well as factors known from the literature. Finally, we discuss the implications of the studies for models of judged variance.

2. Study 1: Intuitive Conceptualizations of Variance

In some previous studies on judgments of variance, participants were provided with a definition of variance, such as "the degree to which numbers cluster about their mean" (Beach & Scopp, 1968, p. 114). Given that not everybody is familiar with statistics, can people relate to the term "variance" when not explicitly provided with a definition? And if so, what concepts of variance they hold?

Study 1 had two main objectives. The first was to establish to what extent "variance" is a meaningful concept for people. The second was to explore the content of their intuitive concept(s) of variance. To that end, we asked respondents to define "variance" and to generate words that they associate with the term. Additionally, we collected information about respondents' level of education and other demographic indicators. We wanted to test whether people with more education and statistics training are more likely to have a concept of variance that is consistent with statistical variance.

To map how people conceptualize variance, we derived a co-occurrence network from their responses to represent the underlying semantic network. In a co-occurrence network, the words serve as the nodes and the size of the node is determined by the number of people who used that word in their response. The edges of the network represent the co-occurrence of pairs of words, which are weighted by the proportion of respondents who used both words.
Co-occurrence networks are commonly used in cognitive science (Baronchelli, Ferrer-i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013) to explore word meanings (Bullinaria & Levy, 2007) and in computer science to identify the topics in a given text (Jia, Carson, Wang, & Yu, 2018). They have, for example, been employed to study emerging topics in scientific publications (Mane & Börner, 2004).

Here we apply co-occurrence network analysis to words obtained in a definition task and words from a free association task (see Section 2.1 for details). For the definition task, the structure of the network relies on the associations between different words emerging from proximity in natural language (Lund & Burgess, 1996). In essence, the network shows which

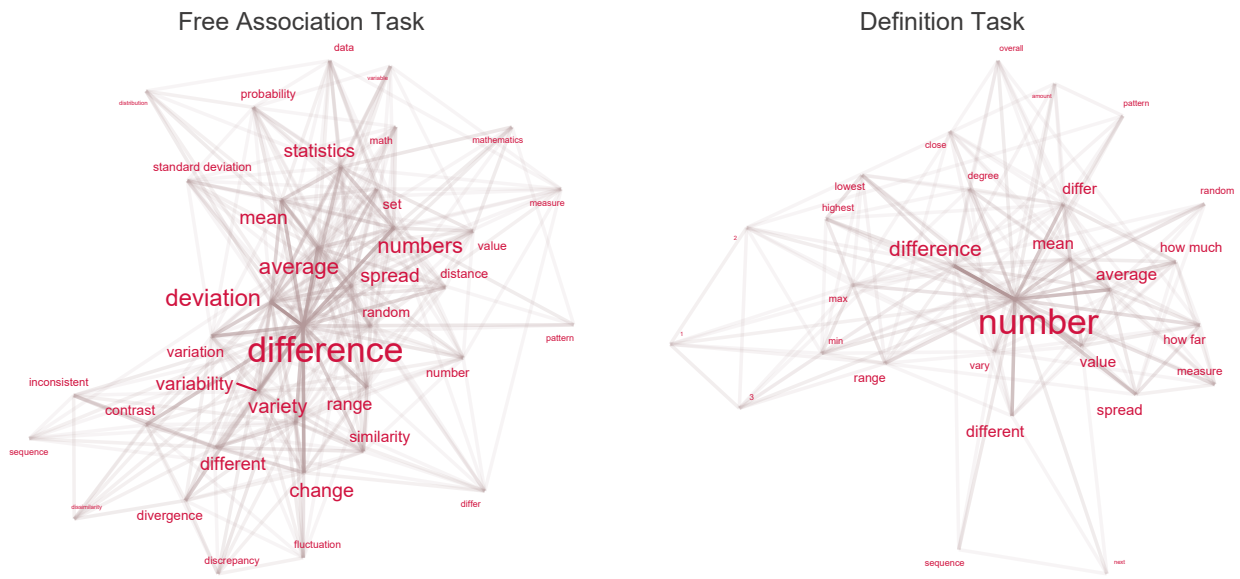

Figure 1. Semantic network based on word co-occurrence in the two tasks in Study 1. Nodes represent words mentioned by at least five respondents. The size of the nodes is proportional to the number of participants who mentioned the word. Edges represent the co-occurrence of pairs of words. The shade of the links is proportional to the frequency with which a given word pair occurred.

words are used together when people provide a definition of variance. For the free association task, the network relies on the proximity of the words in semantic memory. Research has shown that when asked to produce lists of items from memory (e.g., animals), people search their memory locally within a representation before transitioning to a global search (Hills, Todd, & Jones, 2015; Hills & Pachur, 2012). These findings suggest that words that are mentioned together in a free association task are also close to each other in the semantic space.

2.1 Materials and Methods

2.1.1 Participants. We recruited 202 (46% female) respondents on Prolific Academic; all respondents received a flat fee of £1 as compensation. The median age group was 25–34 years, the median level of education was an undergraduate degree, and the median level of statistical knowledge was the basics of statistics learned at school (i.e., secondary education).

Table 1
Most frequently mentioned words and percentage of respondents who mentioned them in the free association and definition tasks in Study 1.

	Free association task			Definition task	
Rank	Word	Mentioned by % of respondents	Rank	Word	Mentioned by % of respondents
1	difference	48%	1	number	81.6%
2	deviation	19.3%	2	difference	23.3%
3	average	16.3%	3	average	13.4%
4	numbers	15.8%	4	mean	12.9%
5	variety	12.9%	5	different	11.9%
6	change	12.3%	6	differ	10.9%
7	mean	11.9%	7	value	10.4%
7	spread	11.9%	8	spread	9.9%
8	statistics	10.9%	9	how much	7.9%
8	variability	10.9%	10	how far	6.9%
9	different	10.4%			
10	range	9.9%			

2.1.2 Procedure. After providing consent, respondents were informed that the study was concerned with the intuitive understanding of variance. In a definition task, they were presented with a set of numbers ("45, 55, 48, 53, 50, 51, 52, 55, 46, 51"), then asked to write in a text box what the "variance" of these numbers meant to them.[2] In a subsequent free association task, respondents were asked to provide a maximum of five words that they thought were "strongly associated with the term variance." Additionally, respondents were asked to indicate their age group by assigning themselves to one of the eight age ranges: from "17 or younger" to "75 or older," with the groups in between spanning 10 years each.[3] We also asked respondents to indicate their gender, their highest level of education, and their highest level of education in statistics, and we used the Berlin Numeracy test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) to assess their numeracy levels.

[2] The exact wording was: "The overall degree by which the numbers in this set differ from each other can be described in terms of variance. In the text box below, please describe in your own words what the variance of a set of numbers means to you."

[3] By collecting information about respondents' age using age groups, we followed a common method in demographic surveys that is used by organizations such as the Office of National Statistics in the United Kingdom

2.2 Results

2.2.1 Is variance a meaningful concept? Most people were able to provide a definition of variance. Only eight of the 202 (4%) respondents stated that they did not know or understand what variance is.

2.2.2 What is the content of people's concepts of variance? In the free association task, respondents provided 956 words, of which 316 were unique.[4] In the definition task, responses had to have a minimum length of 20 characters. For the analysis, we removed the most common words in the English language ("stop words") as well as words that were used often by respondents because of the format of the question.[5] We also removed word repetitions because they cause both inflation in the number of word co-occurrences and loops in the network. This procedure resulted in a total of 1,002 words for the definition task, of which 369 were unique.

Figure 1 shows the resulting semantic networks in the two tasks based on the co-occurrence of words mentioned by at least five respondents. Table 1 lists the 10 most frequent words and Table 2 lists the 10 most frequent word pairs. Many of the most frequent words relate to deviation: "difference," "differ," "deviation," and "how far," as well as "change". These words seem to reflect two different concepts of variance. First, some respondents used them to express that variance relates to differences and distance between the numbers in a set. This understanding is indicated by word pairs "difference–number" and "differ–number" used in the definition task. Moreover, 7.4% of participants mentioned the related notion of "similarity," which also points to a distance (or a lack thereof), in the association task . We refer to this concept as *pairwise distance*.

---

and Pew Research Center in the United States.

[4] For the analysis, we equated "vary" and "varied" to "variability," "similar" and "similarities" to "similarity," and "dissimilar" to "dissimilarity."

[5] We used the package "tidytext" in R (Silge & Robinson, 2016) to remove stop words. Stop words are words that are frequently used in a language but do not carry the main meaning of the sentence. Examples include pronouns such as "I" and "they," the conjugated forms of verb "to be," and the conjunctions "and" and "or." Additionally, we removed the following words: "variance," "set [of numbers]," "one [from another]," "[number] can," "group [of numbers]," "[in this] case," "[in this] way," "data," "[variance] means," "two [numbers]." We replaced "10," "45," and "55" by "range," "min," and "max," respectively, because those numbers corresponded to the range, minimum, and maximum of the set of numbers provided in the task description. Finally, we replaced "values" by "value," "numbers" by "number," and "differs" by "differ."

Table 2
Most frequently mentioned word pairs and percentage of respondents who mentioned them in the free association and definition tasks in Study 1.

	Free association task			Definition task	
Rank	Word	Mentioned by % of respondents	Rank	Word	Mentioned by % of respondents
1	deviation–difference	10.4%	1	difference–number	20.3%
2	average–difference	7.4%	2	average–number	11.9%
2	change–difference	7.4%	3	different–number	11.4%
3	difference–numbers	6.9%	4	differ–number	9.9%
4	difference–mean	5.9%	5	mean–number	8.4%
4	difference–variation	5.9%	5	number–spread	8.4%
5	difference–variability	5.4%	6	number–value	7.9%
5	difference–variety	5.4%	7	how much–number	6.4%
6	contrast–difference	4.9%	8	how far–number	5.4%
6	difference–spread	4.9%	9	measure–number	4.4%
7	average–mean	4.4%	9	number–range	4.4%
7	difference–range	4.4%	9	number–vary	4.4%
7	difference–statistics	4.4%	10	difference–lowest	4%
8	deviation–mean	4%	10	highest–lowest	4%
8	difference–set	4%	10	degree–number	4%
8	difference–similarity	4%	10	max–number	4%
8	deviation–variation	4%	10	number–random	4%
9	average–deviation	3.4%	10	how far–spread	4%
9	difference–distance	3.4%	10	average–value	4%
9	deviation–spread	3.4%			
9	mean–statistics	3.4%			
9	numbers–statistics	3.4%			
9	deviation–variability	3.4%			
10	difference–divergence	2.3%			
10	different–inconsistent	2.3%			
10	range–spread	2.3%			
10	different–variety	2.3%			

Second, some respondents referred to the deviation of the numbers in a set from its central tendency. Here, they often used the words “average” and “mean,” which frequently co-occurred with “difference” and “deviation” (Table 2). These observations suggest that some respondents’ notion of variance relates to how the numbers in the sample spread around the mean or average value in the sample and thus is close to its statistical definition. We refer to this concept as *deviation from the mean*.

Respondents also mentioned words that point to two further, distinct concepts of variance. First, “variety” was one of the words most frequently associated with variance. This indicates that participants conceive of variance in terms of a collection of distinct numbers. The concept of

variety is widely used in research concerned with how individuals discriminate displays of different items (Young & Wasserman, 2001). Participants are shown two sets of icons (e.g., real-life objects or geometric figures) and asked to judge whether the displays are the same or different. Low variability in such "same–different" tasks is usually operationalized as a display with a large percentage of identical items. Studies using this paradigm have shown that variability affects participants' accuracy (Young & Wasserman, 2001). We refer to this concept as *variety*.

Second, the word "range" was frequently mentioned in the association task and the related word "spread" in the definition task. In addition, many word pairs contained the related adjectives "highest," "lowest," and "max." We refer to this concept as *range*.

It is worth noting that these concepts, while conceptually distinct from statistical variance, will not necessarily be statistically independent from it. Being closest to the principle underlying statistical variance, deviation from the mean is likely to be often correlated with statistical variance. To some extent, that might also be the case for pairwise distance and range (but not for variety, because more unique elements in the sample do not necessarily add more variation to it).

2.2.3 Prevalence of the different concepts of variance. To gauge the prevalence of the different concepts of variance among the respondents, we asked two raters to classify the entries independently. In addition to the four concepts described above, the raters could use the following categories: "randomness" (a substantial proportion of respondents associated variance with some random process), "I don't know," and "other" (if the entry did not fit the categories above but referred to an eligible concept). Finally, if the entry did not contain an eligible concept, the raters labeled it "not usable." The interclass correlation between the raters' categorizations was 0.7 (confidence interval [CI]$= [0.615, 0.759]$). We included in our analyses the entries the raters agreed on (136 entries, 67%) as well as the entries that only one rater assigned to a category other than "other" or "not usable" (27 entries, 13%). The 38 (20%) entries where the raters disagreed on the substantive categories (i.e., those other than "other" and "not usable") were removed.[6]

[6] Because one fifth of the responses were thus excluded, we conducted robustness checks. Specifically, instead of using the restricted set of entries (i.e., only those that raters agreed on), we ran two sets of the same regressions (one

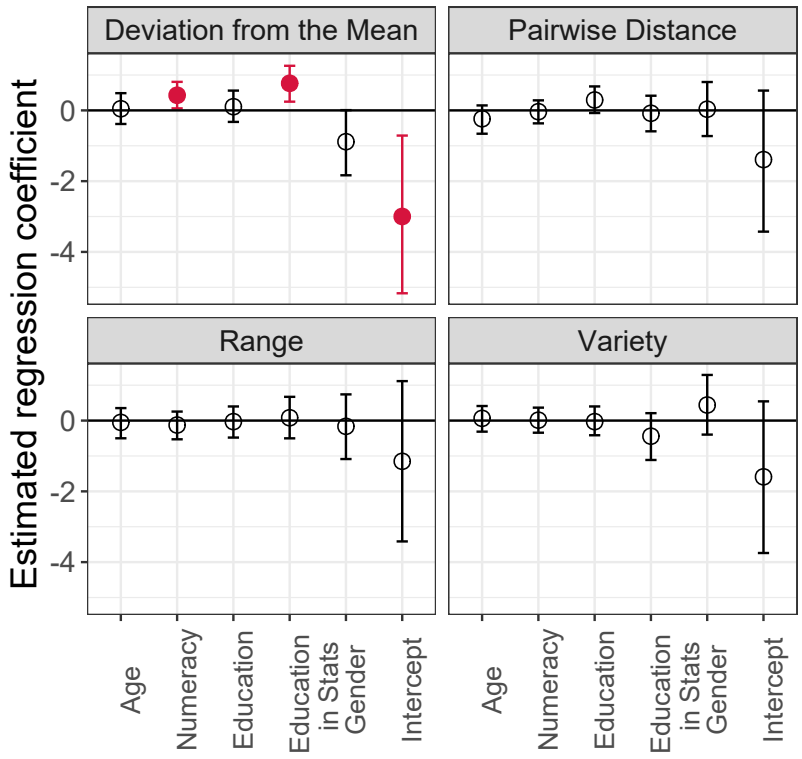

Figure 2. Results of the binomial logistic Bayesian regression analysis in Study 1. Each panel presents the results of a different model, using the respective concept of variance as the dependent variable; in each model, the value of the independent variable equaled 1 if the corresponding concept was mentioned in a given respondent's definition of variance and 0 otherwise. The error bars denote the 95% highest density interval ($HDI_{95\%}$) for each estimated regression coefficient. Estimates whose $HDI_{95\%}$ do not include 0 are printed in solid red.

Based on the raters' classifications, 89% of respondents provided some definition of variance, while 7% provided a response that was "not usable." (As we noted at the beginning of Section 2.2, the remaining 4% of respondents explicitly stated that they did not know or understand what variance is.) Of the 89%, 80% provided a definition of variance that is covered by one of the four concepts identified by the co-occurrence network analysis; 5% referred to randomness and 4% to some other concept. Overall, pairwise distance was the most prevalent concept (24.5%), followed by deviation from the mean (20.2%), variety (20.2%), and range

---

for each rater) that included all entries (for one rater, one observation had to be excluded because the rater made a mistake and did not fill the cell correctly). In other words, in one set of regressions, we included both the entries that raters agreed on and the entries that they disagreed on, where we used the label given by rater A. In the other set, we used the label given by rater B for the entries raters disagreed on. The main results remained the same.

(15.3%).

2.2.4 Were the different concepts of variance related to demographic variables?

Finally, we examined how the different concepts of variance were related to demographic variables. For each of the four most popular concepts (pairwise distance, deviation from the mean, variety, and range), we created a binary variable that equaled 1 if the concept was invoked by the respective respondent and 0 otherwise. This binary variable was used as a dependent variable in a Bayesian logistic regression, and the demographic variables were used as predictors.[7] For all estimations in this section and subsequent sections, we used the RStan (Stan Development Team, 2020) and brms (Bürkner, 2017) packages in R. For all parameters, normal distributions with a mean of 0 and a standard deviation of 8 were used as priors. As Figure 2 shows, a higher level of education in statistics and higher numeracy were positively associated with the probability of defining variance as deviation from the mean (the 95% highest density interval [$HDI_{95\%}$] for both parameters excludes 0). For all other concepts, the relationship was not credible.

3. Studies 2a and 2b: What Factors Influence People's Judgments of Variance?

Study 1 established that variance is a meaningful concept for most people and that it is, therefore, appropriate to use the term to elicit judgments of variability. Further, Study 1 provided insights into the content of people's concepts of variance, indicating that, along with a concept related to statistical variance—namely, deviation from the mean—pairwise distance (the difference between number pairs), variety (the number of unique elements in the sample), and range reflect some aspects of their perception of variance. In Studies 2a and 2b, we tested the extent to which these factors—along with mean and sample size, which (as summarized in Section 1) previous research has suggested impacts judgments of variance—predict judgments of variance when people are presented with sets of number sequences. As a benchmark, we also

[7] We did not include concepts mentioned by fewer than 10% of respondents because the small sample size for these concepts implies that estimating a logistic regression can be unreliable (for a discussion see King & Zeng, 2001). Furthermore, for this analysis, we removed the observations that were classified as "not usable"; these do not carry meaningful information and may therefore bias the estimations. Overall, the analysis included 146 observations. Apart from the binomial regressions, we also estimated multinomial regressions with each of the concepts as the reference level. The results of that analysis were very similar to those reported here.

include statistical variance as a predictor in the analysis.

As we noted in Section 1, in many types of distributions the sample characteristics that we focus on here tend to be considerably intercorrelated (e.g., range and statistical variance). Therefore, in Studies 2a and 2b we designed specific distributions that enhanced the possible differences across candidate sample characteristics. We then employed Bayesian regression analysis to statistically control for the remaining interdependencies between the factors.

3.1 Materials and Methods

3.1.1 Participants. We recruited 151 (41% female) and 150 (55% female) participants on Prolific Academic for Study 2a and Study 2b, respectively. All participants received a flat fee of £2 as compensation. In both studies, the median age group was 25–34 years, the median level of education was an undergraduate degree, and the median level of statistical knowledge was the basics of statistics learned at school.

3.1.2 Procedure. Studies 2a and 2b had the same general structure. After providing consent, participants were informed that the study was concerned with intuitive perceptions of variance. The experiment consisted of a familiarization phase and a test part. In the familiarization part, participants were presented with the following text: "One way to think about variance is that it reflects the accuracy when guessing numbers that are randomly drawn from a set of numbers. Imagine that after seeing a set of numbers, five numbers are randomly drawn from this set one after another and you are asked to guess each number. The HIGHER the variance of a set is, the HIGHER will be the average error of your 5 guesses." A graphical illustration accompanied this text (see the Supplementary Material for details). Participants were then shown two sequences of numbers, one with high variance and one with low variance. The level of variance corresponded to the high/low variance levels that participants observed in the subsequent test phase (but with a different mean).

The test phase consisted of 12 trials. The sets of numbers presented in each trial were randomly drawn from a particular distribution; for some trials in Study 2a and all trials in Study

2b, the maximum and minimum number were added to that set of numbers (having the minimum and maximum in the sequence ensured that the range could be set to a particular value). The numbers of these quasi-random sets of values were presented sequentially to participants with a rate of 600ms per number (cf. Goldstein & Rothschild, 2014).

The distributions used in the studies were designed to vary in their mean, statistical variance, range, sample size, and variety. While independent variation in mean and sample size was easily achieved, this was more challenging to achieve for statistical variance, range, and variety. The design was guided by the creation of pairs of trials that contrasted different factors. There were one and three pairs of trials contrasting the mean in Study 2a and 2b, respectively. In both studies, one pair of trials contrasted the sample size; three and two different sample sizes were used in different trials in Study 2a and 2b, respectively. Because it was harder to isolate statistical variance, range, and variety, we next created distributions in which the predictions of the two factors in a pair of trials would specifically go in opposite directions or the same direction. For instance, in one pair of trials, the distribution of one trial had a higher variety than the distribution of the other trial but a lower variance (see Trial 3 and 4 in Study 2a and Trials 10 and 11 in Study 2b; Tables S1 and S2 in Supplemental Materials for details). An example of two factors making the same prediction is Trials 7 and 8 in Study 2a (see Supplemental Materials), in which variety and range are both higher in Trial 8 (see also Trials 4 and 9 in Study 2b). In Study 2b, we took a more systematic approach to the design of the distributions. Specifically, we manipulated the number of distinct outcomes (2, 3, or multiple) and the shape of the distribution (uniform, unimodal, bimodal). Manipulating the shape of the distribution creates changes in one factor while keeping others on a similar level. For instance, a unimodal distribution has lower variance than a uniform distribution but their range and variety are similar (e.g., Trials 2 and 11 in Study 2b). Detailed information about the distributions used in both experiments can be found in the Supplementary Material (Tables S1 and S2). Note that although the design aims at disentangling the different factors, the factors are not perfectly orthogonal to each other across the different trials—that is, the design is not factorial.

Directly after all numbers in a sequence had been presented, participants were asked to judge the variance of the sequence on a scale from 0 ("very low variance") to 100 ("very high variance").[8] Additionally, participants were asked to provide the same demographic information as in Study 1 (i.e., age group, gender, level of education, and level of statistical knowledge).

3.2 Results

Because the factors were not perfectly orthogonal across trials, participants' variance judgments were analyzed with a set of Bayesian mixed-effects linear regressions, using the sample characteristics—the mean, variance, deviation from the mean, range, sample size, variety, and pairwise distance, as derived from the samples that each participant actually observed—as fixed effects. All models included random intercepts for each participant. This approach allows us to take into account the remaining correlation between the factors as well as take advantage of the variability of factors across trials.

To assess the impact of each individual factor and determine the predictive power of each, we first estimated a set of simple regression models where each of the factors was the only predictor. All factors except the mean were positively associated with judged variance (see the left panels on Figures 3 and 4). In Study 2a, the mean was unrelated to judged variance ($b = 0.03$, $HDI_{95\%}$ = [-0.05, 0.11]) but was negatively related in Study 2b ($b = -0.11$, $HDI_{95\%}$ = [-0.16, -0.06]). In order to evaluate the individual factors, we determined the predictive accuracy of the simple regression models using the leave-one-out technique, quantifying predictive accuracy as the expected log predictive density (ELPD; Vehtari, Gelman, & Gabry, 2017).[9] Moreover, we estimated the Bayesian equivalent of R^2 to access the performance of the models on the absolute scale.[10] From the individual models, those with variety, range, and pairwise distance were better

[8] The exact wording of the task was: "You just observed a set of numbers. Please indicate on the scale how much variance, in your view, this set has. To indicate your judgment, please click on the respective location on the scale and adjust the bar to match your assessment."

[9] Leave-one-out cross-validation is a technique in which one value of the dependent variable is removed in the estimation and then predicted by the estimated model. This procedure is usually iterated across all values of the dependent variable and and overall predictive accuracy is calculated.

[10] Specifically, we use a statistic developed by Gelman, Goodrich, Gabry, and Vehtari (2019) that divides the variance

predictors of judged variance than those with statistical variance in both studies (see the right panels in Figures 3 and 4). The model with variety performed best in both studies (with coefficients of $b = 2.73$, $HDI_{95\%}$ = [2.56, 2.91], ELPD= -7758.2, $R^2 = 0.473$ and $b = 3.13$, $HDI_{95\%}$ = [2.94, 3.30], ELPD= -7945.9, $R^2 = 0.492$ in Studies 2a and 2b, respectively), followed by the model with range (with coefficients of $b = 1.92$, $HDI_{95\%}$ = [1.76, 2.07], ELPD= -7866.4, $R^2 = 0.406$ and $b = 3$, $HDI_{95\%}$ = [2.79, 3.22], ELPD= -8075.3, $R^2 = 0.423$ in Studies 2a and 2b, respectively); in Study 2b the model with range had similar predictive accuracy to the model with sample size. The model with pairwise distance was the third-best performing model in Study 2a and the fourth-best in Study 2b (with coefficients of $b = 5.58$, $HDI_{95\%}$ = [5.07, 6.10], ELPD= -7927.8, $R^2 = 0.363$ and $b = 6.15$, $HDI_{95\%}$ = [5.52, 6.80], ELPD= -8235.4, $R^2 = 0.295$ in Studies 2a and 2b, respectively). The model with deviation from the mean was the second worst in terms of predictive accuracy in both studies (with coefficients of $b = 4.42$, $HDI_{95\%}$ = [3.75, 5.11], ELPD= -8063.6, $R^2 = 0.258$ and $b = 4.12$, $HDI_{95\%}$ = [3.38, 4.88], ELPD= -8350.3, $R^2 = 0.197$ in Studies 2a and 2b, respectively). The models had similar levels of R^2 across studies, ranging from $0.47 - 0.49$ for the best-performing models to $0.2 - 0.25$ for the worst-performing models.

In a second step and in order to mutually control for the factors' contributions in the prediction of judged variance, we estimated a set of more complex models that contained multiple factors as predictors. With this approach it is important to be sensitive to the potential issue of multicollinearity (i.e., high correlation between different predictors), which might lead to a trade-off between different coefficients in the parameter estimation and can increase the uncertainty in the estimate (reflected in a very wide posterior distribution of the regression coefficients). As it turned out (see Figures S9 and S10 in the Supplementary Material for correlograms), range was highly correlated with statistical variance ($r = 0.77$), pairwise distance

---

of the predicted values by the variance of the predicted values plus the expected variance of errors.

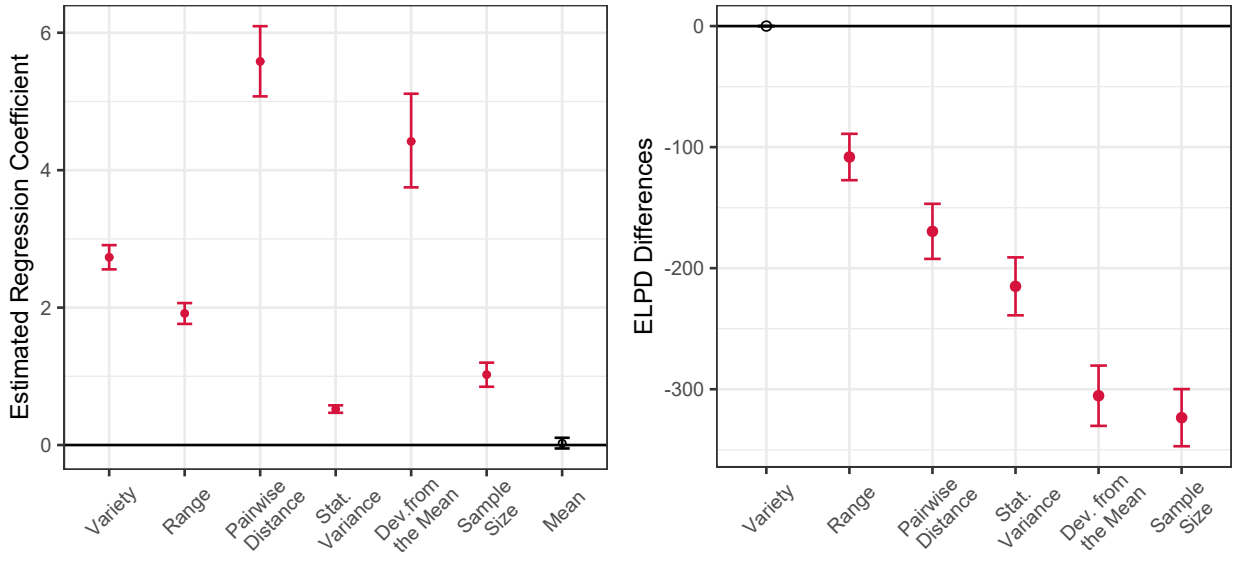

Figure 3. Results of the linear Bayesian regression analysis predicting people's variance judgments based on single predictors in Study 2a. Points on the left panel are the estimated regression coefficients of the different predictors (shown on the *x* axis). Error bars denote the 95% highest density interval ($HDI_{95\%}$) for the estimated coefficient. Estimates whose $HDI_{95\%}$ did not include 0 are printed in solid red. The right panel shows the predictive accuracy of every model relative to the best model at 0. The measure of predictive accuracy is expected log predicted density (ELPD).

($r = 0.78$), and variety ($r = 0.79$) in Study 2a, and variety was highly correlated with range ($r = 0.86$) and sample size ($r = 0.83$) in Study 2b. Furthermore, statistical variance was highly correlated with both deviation from the mean ($r = 0.93$ and 0.95 in Studies 2a and 2b, respectively) and pairwise distance ($r = 0.98$ and 0.95 in Studies 2a and 2b, respectively). Based on computer simulations, McElreath (2018) concluded that multicollinearity in estimating regression coefficients is a serious problem with intercorrelations $r > .7$. In light of this result, we ran separate models for predictors with intercorrelations $r > .7$. In total, we estimated four models for Study 2a and six models for Study 2b.[11] Tables 3 and 4 show the results. In both

[11] The full model would have included all seven predictors. However, some predictors could not be included in the same model due to high intercorrelations. In both studies, we did not run models in which statistical variance, pairwise distance, and deviation from the mean were simultaneously included. Furthermore, in Study 2a, we did not run any models that simultaneously contained range and variety, range and statistical variance, or range and pairwise distance. This resulted in only one model that contained range (Model 4). Models 1–3 contained variety and either statistical variance, pairwise distance, or deviation from the mean. In Study 2b, we did not run models that contained variety simultaneously with range or sample size. This resulted in models that included either variety (Models 1–3)

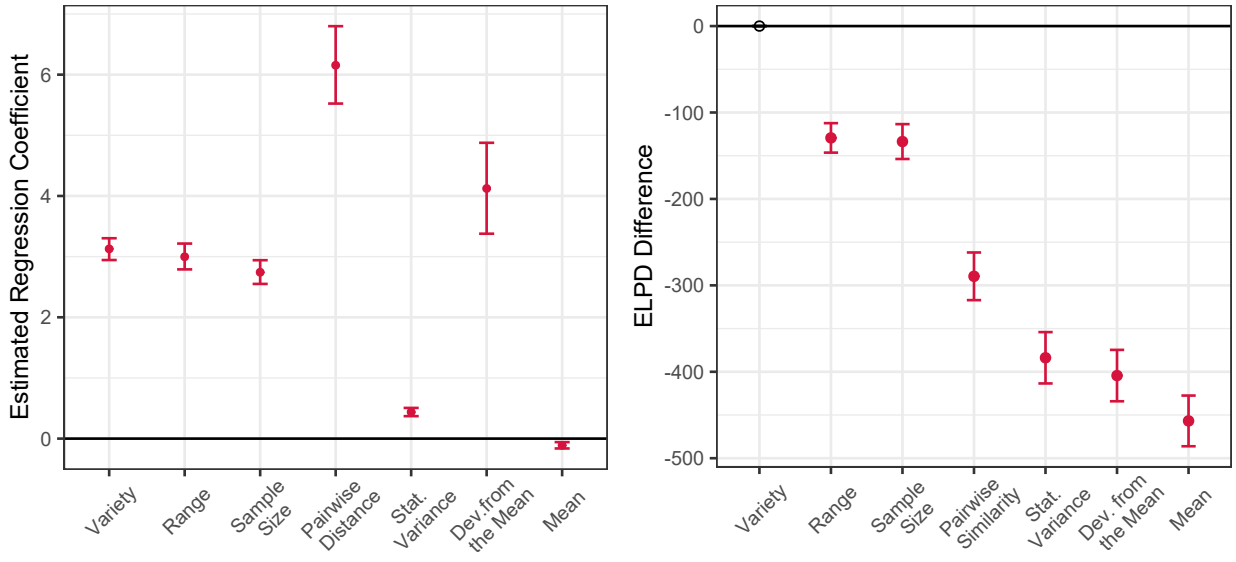

Figure 4. Results of the linear Bayesian regression analysis predicting people's variance judgments based on single predictors in Study 2b. Points on the left panel are the estimated regression coefficients of the different predictors (shown on the *x* axis). Error bars denote the 95% highest density interval ($HDI_{95\%}$) for the estimated coefficient. Estimates whose $HDI_{95\%}$ did not include 0 are printed in solid red. The right panel shows the predictive accuracy of every model relative to the best model at 0. The measure of predictive accuracy is expected log predicted density (ELPD).

studies, the best-performing model was the model containing variety and pairwise distance (Model 1 in both studies); in Study 2b, that model had similar predictive accuracy to the one containing variety and statistical variance (Model 2). Models containing range and sample size (Model 4 in Study 2a and Models 4–6 in Study 2b) performed worse that models containing variety in both studies. For all models, the R^2 did not fall below 0.43 and, for the best-performing models, R^2 was 0.491 and 0.498 for Studies 2a and 2b, respectively.

A comparison with the results of the simple models reported above also points to an interaction between mean and other factors: Although the mean was unrelated (Study 2a) or negatively related (Study 2b) to judged variance in the simple models, the relationship was positive when other factors were controlled for (in the models with multiple predictors). For sample size, there seemed to be an interaction with other factors: Although its positive

---

or range and sample size (Models 3–6), and either statistical variance, pairwise distance or deviation from the mean.

Table 3

Results of the linear Bayesian regression analysis predicting people's variance judgments in Study 2a based on multiple predictors. The interval in the parenthesis denotes the 95% highest density interval ($HDI_{95\%}$) for the estimated coefficient. Coefficients whose $HDI_{95\%}$ do not contain 0 are in boldface. Bayesian R^2, the predictive accuracy of each model (in terms of expected log predicted density; ELPD) as well as the difference in ELPD to Model 1 are reported in the last three rows.

	Model 1	Model 2	Model 3	Model 4
Variety	**2.21** [1.99,2.44]	**2.34** [2.12,2.54]	**2.57** [2.36,2.75]	
Range				**2.17** [1.96,2.37]
Pairwise distance	**2.54** [1.85,3.24]			
Statistical variance		**0.24** [0.17,0.31]		
Deviation from the mean			**2.19** [1.46,2.87]	**1.49** [0.71,2.25]
Sample size	-0.16 [-0.36,0.03]	-0.18 [-0.38,0.02]	-0.12 [-0.31,0.08]	**-0.53** [-0.77, -0.31]
Mean	**0.09** [0.03,0.16]	**0.09** [0.03,0.16]	**0.08** [0.13,0.27]	**0.21** [0.14,0.28]
Bayesian R^2	0.491	0.49	0.486	0.431
ELPD	-7731.2	-7733.7	-7740	-7830.7
ELPD difference	–	-2.60 [-4.6,-0.6]	-8.8 [-11.8,-6]	-99.5 [-115.2,-83.8]

relationship with judged variance in the simple model also held when controlling for the other factors in Study 2b, in Study 2a the relationship disappeared or was negative when other factors were controlled for (see Model 4 in Table 3).

Taken together, the results of Studies 2a and 2b show that several of the concepts articulated by respondents in Study 1 as aspects of variance may serve as cues when people actually judge variance. Variety, pairwise similarity, and range were found, along with statistical variance, to have strong positive associations with judgments of variance. In both studies, simple models containing variety or range as single predictors performed better than the model containing statistical variance, which was also outperformed by the model containing pairwise distance in Study 2a. In Study 2b, the models had similar predictive accuracy. The analysis of model performance indicates that variety and pairwise distance were better predictors than the

Table 4
Results of the linear Bayesian regression analysis predicting people's variance judgments in Study 2b based on multiple predictors. The interval in the parenthesis denotes the 95% highest density interval ($HDI_{95\%}$) for the estimated coefficients. Coefficients whose $HDI_{95\%}$ do not contain 0 are in boldface. Bayesian R^2, the predictive accuracy of each model (in terms of expected log predicted density; ELPD) as well as the difference in ELPD to Model 1 are reported in the last three rows.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Variety	**2.94** [2.71,3.16]	**3.04** [2.85,3.24]	**3.08** [2.88,3.27]			
Range				**1.92** [1.52,2.31]	**2.08** [1.72,2.44]	**2.20** [1.8,2.59]
Pairwise distance	**1.22** [0.54,1.88]			**0.95** [0.19,1.69]		
Statistical variance		**0.11** [0.05,0.17]				-0.01 [-0.09,0.06]
Deviation from the mean			**0.95** [0.32,1.58]		0.44 [-0.23,1.12]	
Sample size				**1.40** [1.13,1.69]	**1.41** [1.14,1.7]	**1.39** [1.1,1.67]
Mean	**0.06** [0.01,0.1]	**0.06** [0.02,0.11]	**0.05** [0.01,0.097]	**0.09** [0.04,0.14]	**0.09** [0.04,0.14]	**0.09** [0.04,0.14]
Bayesian R^2	0.498	0.497	0.496	0.469	0.468	0.467
ELPD	-7938.9	-7939.1	-7941.4	-7989.5	-7991.7	-7992.8
ELPD difference	–	-0.20 [-1,0.6]	-2.5 [-3.4,-1.6]	-50.50 [-62.5,38.5]	-52.80 [-64.9,-40.7]	-53.80 [-65.9,-41.7]

other factors. The model containing these two factors was the best-performing one in Study 2a. In Study 2b, it had a similar predictive accuracy to the model containing variety and statistical variance. The associations among mean and sample size with people's judgments were more complex. The bivariate relationships are consistent with previous research: Mean and sample size are negatively and positively associated with judged variance, respectively. Our analysis, however, indicates that these factors interact with other factors.

4. Study 3: Variance Judgments with Naturally Varying Stimuli

The goal of the targeted design of the stimuli used in Studies 2a and 2b was to disentangle possible determinants of judged variance. For instance, we used distributions that had a similar range but differed in variance. This approach may have resulted in somewhat unrealistic stimuli: In samples drawn from some common distributions, such as the normal distribution, range and

variance are usually correlated. In Study 3 we examined whether our conclusions from Studies 2a and 2b generalize to less construed value distributions. To that end, all samples were randomly drawn from the same normal distribution, and we took advantage of the natural variation due to sampling error in mean, range, variance, and variety of the individual samples to examine each characteristic's influence on people's variance judgments.

4.1 Materials and Methods

4.1.1 Participants. We recruited 150 (39% female) participants via Prolific Academic; they received a flat fee of £2 as compensation. The median age group was 25–34 years, the median level of education was an undergraduate degree, and the median level of statistical knowledge was the basics of statistics learned at school.

4.1.2 Procedure. Study 3 had the same overall structure as Studies 2a and 2b but the test phase consisted of four rather than 12 trials. In all trials, the numbers in the sample were drawn from a discretized normal distribution with a mean of 50 and a standard deviation of 10. Sample size was manipulated across four levels, such that 5, 10, 15, or 20 numbers were drawn from the distribution on a given trial. The characteristics of the generating distribution were chosen such that the samples were comparable to those used in Studies 2a and 2b (e.g., all numbers were between 20 and 80). Participants were asked to provide the same demographic information as in the previous studies.

4.2 Results

We used the same linear Bayesian regression analysis as in Studies 2a and 2b to analyze participants' variance judgments. First, we examined bivariate relationships between the factors and judged variance (see Figure 5) using simple models with a single predictor. As in the previous studies, these simple models showed that all factors except for the mean were positively associated with judged variance; the mean did not have an effect on the judgments ($b = 0.01$,

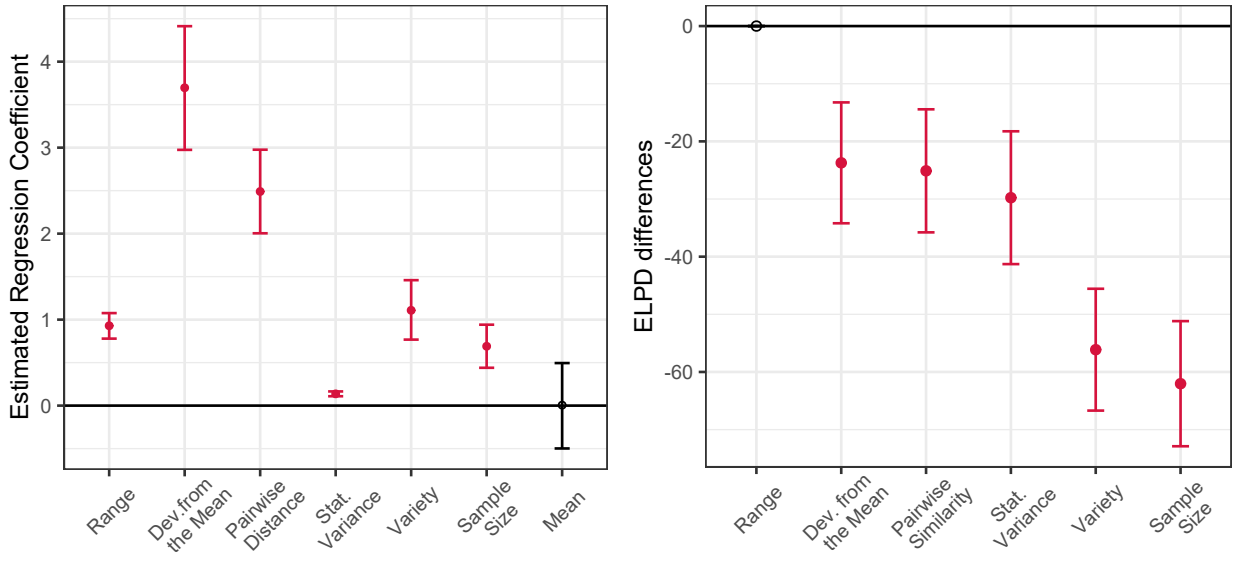

Figure 5. Results of the linear Bayesian regression analysis predicting people's variance judgments in Study 3 based on the single predictors. Points on the left panel represent the estimated regression coefficients of the different predictors (shown on the x axis). The error bars denote the 95% highest density interval ($HDI_{95\%}$) for the respective estimated coefficient. Estimates whose $HDI_{95\%}$ did not include 0 are printed in solid red. The right panel shows the predictive accuracy of every model relative to the best model at 0. The measure of predictive accuracy is expected log predicted density (ELPD).

$\text{HDI}_{95\%} = [-0.50, 0.49]$). Based on the models' predictive accuracy, range was the best single predictor ($b = 0.93$, $\text{HDI}_{95\%} = [0.78, 1.08]$, ELPD = -2566.9, $R^2 = 0.492$), with deviation from the mean ($b = 3.7$, $\text{HDI}_{95\%} = [2.97, 4.41]$, ELPD = -2590.7, $R^2 = 0.446$), pairwise distance ($b = 2.49$, $\text{HDI}_{95\%} = [2.00, 2.98]$, ELPD=$-2592$, $R^2 = 0.444$), and statistical variance ($b = 0.14$, $\text{HDI}_{95\%} = [0.11, 0.17]$, ELPD=$-2596.7$, $R^2 = 0.436$) in second place. The models with variety ($b = 1.11$, $\text{HDI}_{95\%} = [0.77, 1.46]$, ELPD=$-2623.1$, $R^2 = 0.378$) and sample size ($b = 0.69$, $\text{HDI}_{95\%} = [0.44, 0.94]$, ELPD=$-2629.0$, $R^2 = 0.364$) performed on the similar level, but the models with these factors were clearly outperformed by the other models. The levels of R^2 ranged from 0.49 for the best-performing model to 0.364 for the worst-performing one.

In a second step, we ran more complex regression models that contained multiple predictors. As shown in Figure S11 in the Supplementary Material, in Study 3 variety and sample size were highly intercorrelated ($r = 0.96$), and range was highly correlated with statistical

Table 5

Results of the linear Bayesian regression analysis predicting people's variance judgments in Study 3 based on multiple predictors. The interval in the parenthesis denotes the 95% highest density interval ($HDI_{95\%}$) for the estimated coefficients. Bayesian R^2, the predictive accuracy of each model (in terms of expected log predicted density; ELPD) and the difference in ELPD to Model 1 are reported in the last three rows.

	Model 1	Model 2	Model 3	Model 4
Variety	**1.09**	-0.06	**1.11**	**0.89**
	[0.78,1.41]	[-0.45,0.33]	[0.8,1.42]	[0.57,1.21]
Range		**0.95**		
		[0.76,1.13]		
Pairwise distance	**2.48**			
	[2.02,2.95]			
Statistical variance			**0.14**	
			[0.11,0.16]	
Deviation from the mean				**3.41**
				[2.69,4.12]
Mean	-0.10	-0.10	-0.14	-0.10
	[-0.55,0.33]	[-0.54,0.34]	[-0.58,0.29]	[-0.54,0.34]
Bayesian R^2	0.497	0.493	0.491	0.48
ELPD	-2566.6	-2569.3	-2570.8	-2576.1
ELPD difference	–	-2.60	-4.10	-9.4
		[-9.5,4.3]	[-7.7,-0.5]	[-12.4,-6.4]

variance, pairwise distance, and deviation from the mean ($r = 0.72$, 0.71, and 0.71, respectively). Moreover, as in the previous studies, statistical variance was highly correlated with both pairwise distance (0.98) and deviation from the mean (0.94). Overall, we estimated four models (see Table 5).[12]

As in Studies 2a and 2b, the model with variety and pairwise distance as predictors (Model 1) performed best. However, the simple model with only range (as well as Model 2, which included both range and variety) showed similar predictive accuracy (its ELPD was $0.3[-6.5, 7.1]$ higher than that of the model with variety and pairwise distance). Although statistical variance and pairwise distance predicted people's variance judgments well as single

[12] Because of these high intercorrelations, we did not run models that included range simultaneously with statistical variance, pairwise distance, or deviation from the mean. Models 1, 3, and 4 contained variety and either statistical variance, pairwise distance, or deviation from the mean. Model 2 contained range and variety. As all models contained variety, none contained sample size (we examine the role of sample size in a separate analysis further in this section).

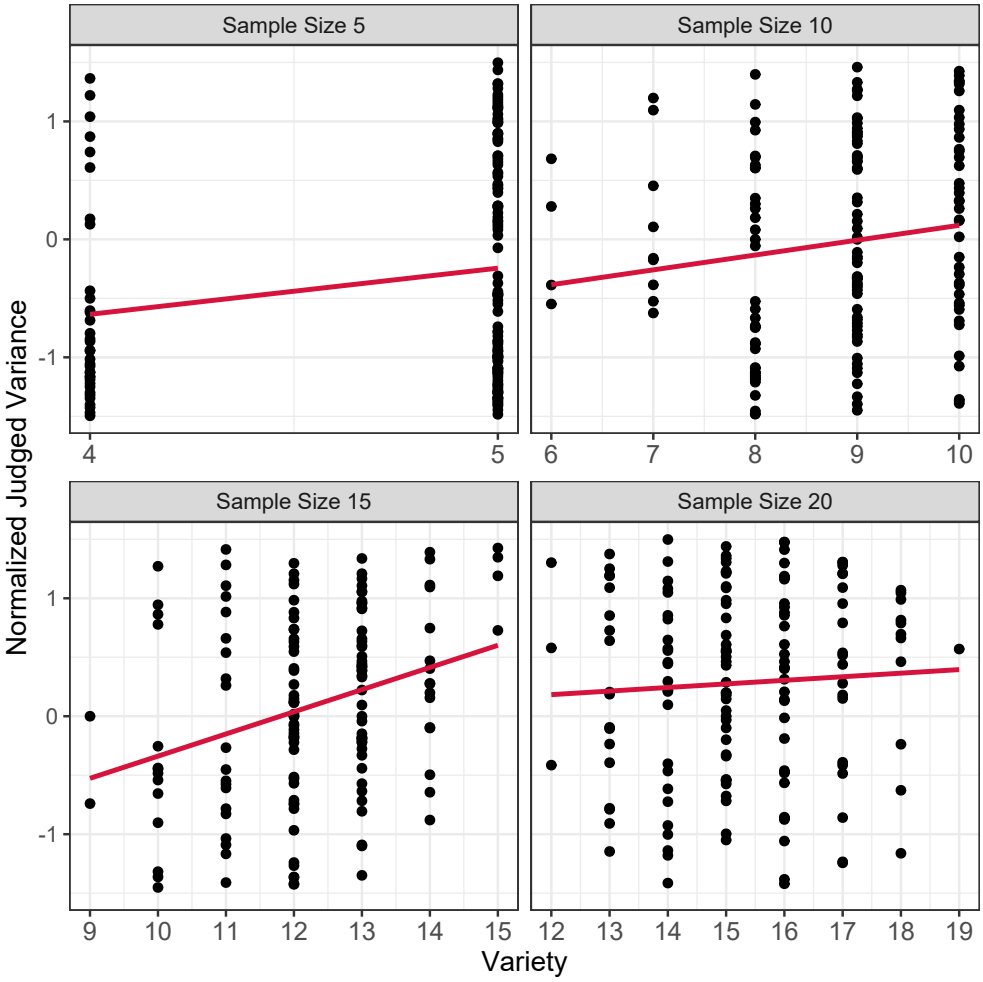

Figure 6. Relationship between variety and judged variance (normalized for each participant) in Study 3. Each panel shows one of four trials where sample size was held constant. The red line shows the simple regression line between the two variables.

predictors in the simple regression models, in the multiple regression analysis the models that contained pairwise distance performed better than those that contained statistical variance. For all models, the R^2 did not fall below 0.48 and, for the best-performing model, R^2 was 0.497. Note that because of the high intercorrelation between variety and sample size, the latter was not included into any of the models we ran. Unlike in the previous studies, the effect of variety and sample size could not be disentangled in the regression analysis: the very high correlation meant that the effect of variety on judged variance is confounded with the effect of sample size in these regression models. We turn to this issue shortly.

The mean was unrelated to judged variance when other factors were controlled for. However, because the mean was not experimentally manipulated but rather varied across trials due to sampling error, the differences in mean across trials were smaller than in Studies 2a and 2b: In Study 3, 89% of the sample means were between 45 and 55, or close to 50 (the mean of the normal distribution used to generate the samples). Consequently, the differences between the majority of sample means in Study 3 were no more than 10. In Studies 2a and 2b, by contrast, this difference was at least 38.1 and 36, respectively.

To address the confound between variety and sample size, we conducted additional analyses in which we plotted the association between within-participant normalized judgments of variance and variety separately for each level of sample size (thus holding sample size constant).[13] As can be seen in Figure 6, in all four trials the association between the two factors was positive. This observation was supported by a mixed-effects regression model in which we predicted judged variance based on variety and included, for each level of sample size, a separate random intercept and random slope. In this model—consistent with the pattern in Figure 6—variety was positively associated with judged variance: $b = 1.71$, $\text{HDI}_{95\%} = [0.69, 3.14]$.

In sum, like Studies 2a and 2b, the results of Study 3 show that range, pairwise distance, and variety had a strong positive association with judged variance. Unlike in Studies 2a and 2b, however, the strongest predictor was range (rather than variety). This result indicates that when judging variance people may pay attention to different aspects depending on context. Moreover, statistical variance was outperformed by range as a single predictor and the best-performing model contained pairwise distance instead of statistical variance. Overall, the results of Studies 2a, 2b, and 3 demonstrate that pairwise distance, variety, and range—identified in Study 1 as central to people's conceptualization of variance—also predict people's actual judgments of variance.

---

[13] We normalized the judgments within the participant in order to control for individual differences in the overall level of judgments, in parallel to the random intercepts for participants in the mixed-effects regression models.

5. Discussion

Statistical variance is a frequently invoked concept in models of cognition to denote people's possible sensitivity to variability of stimuli in the environment (e.g., Stewart & Chater, 2002; Weber et al., 2004; Markowitz, 1952). To what extent does people's concept of variance overlap with statistical variance? If the statistical concept of variance only approximates people's perception of variance, which features of the distribution of stimuli feed into their intuitive concept? And, importantly, are people's actual judgments of variance sensitive to these features?

We first elicited people's conceptualizations of variance and conducted a co-occurrence analysis of the resulting descriptions and associations. This analysis showed that although most people were able to spontaneously articulate a concept of variance, that concept deviated from the statistical definition. Instead, in addition to deviation from the mean (the concept most similar to the statistical definition), three other concepts emerged as important aspects of people's conceptualization of variance: pairwise distance, range, and variety.

The results of three experimental studies then showed that these factors also predicted people's judgments of variance. In all experiments, pairwise distance, range, and variety were strong predictors of judged variance. In Studies 2a and 2b, models containing these factors as single predictors had higher predictive accuracy than those containing statistical variance as a single predictor, and the more complex model containing both variety and pairwise distance performed better than or on similarly to the model containing statistical variance. In Study 3, the model with variety and pairwise distance and the model containing range performed best, and range outperformed statistical variance as a single predictor of people's judgments of variance.

Although the general results were similar between studies, the relative importance of the factors in terms of predictive accuracy in regression models differed between Studies 2a, 2b, and 3. A possible reason for this was the difference in the stimuli used. In Studies 2a and 2b, the numbers were generated from distributions that differed in mean, variance, range, and variety. In Study 3, the numbers for all trials were drawn from the same normal distribution. Due to this difference in design, variation across trials in range was much higher in Study 3 (84.02) than in

Studies 2a (29.95) and 2b (20.57), whereas variation in variety was smaller (16.57 in Study 3 relative to 19.31 and 24.5 in Studies 2a and 2b, respectively). Furthermore, range and variety were less strongly correlated in Study 3 ($r = 0.55$) than in Studies 2a and 2b ($r = 0.79$ and 0.86, respectively). As people have been shown to pay more attention to attributes with higher variability than to attributes with lower variation (Einhorn, Kleinmuntz, & Kleinmuntz, 1979), it is possible that the higher variation of range in Study 3 increased its salience when people constructed their variance judgments, and that the lower variation of variety decreased its salience.

Our conclusions regarding the effect of the mean on variance judgments expand on those drawn in previous studies, which had observed that variance is judged to be lower for stimuli of larger magnitude (Lathrop, 1967; Weber et al., 2004). Consistent with these findings, our analysis of the bivariate relationship showed a negative association between the mean and people's judgments in Study 2b; in Study 2a and Study 3, however, the mean was unrelated to judged variance. In addition, further analyses suggested that the mean interacted with the other factors: Its association with judged variance was positive when other factors were taken into account. There was no evidence for such an interaction in Study 3, but note that here the variation of the means across the trials was smaller than in Studies 2a and 2b. One potential explanation for our results only partly replicating previous results on the impact of mean on judged variance relates to methodological differences. In the studies by Weber et al. (2004), participants chose between a two-outcome gamble and a sure value after experiencing each gamble's payoff distribution through sequential sampling. The authors used gambles with highly skewed payoff distributions (e.g., 10 with probability 0.1, 0 otherwise), meaning that the means of samples from the gambles were extremely low relative to the mean magnitude of the outcomes people encountered (the expected value of 10 with probability 0.1, 0 otherwise, is 1). The distributions we used in our experiments, by contrast, were symmetric, meaning that the mean value was a good representative of the outcomes people encountered. This difference in skewness might explain the distinct effects of the mean in our experiments and in Weber et al. (2004).

Finally, we found that, when a bivariate relationship is concerned, a larger sample size was

associated with high perceived variance in all three studies. At the same time, we observed an interaction effect in Study 2a, where the association between sample size and judged variance disappeared or even became negative when other factors were accounted for. In Studies 2b and 3, the relationship remained positive. Our results are thus, to some extent, consistent with previous studies that manipulated sample size and found that it affected judgments of variability (Konovalova & Le Mens, 2020). At the same time, they indicate that the effect of sample size might be mediated by other factors, such as variety or range.

Our studies established that not only statistical variance but other sample characteristics such as variety, range, pairwise distance, and sample size are important predictors of judged variance. It should be noted, however, that even here the impacts of the different factors on the perception of variance could not be completely disentangled from each other. Furthermore, the importance of each factor seemed to vary somewhat across the studies, indicating that variance perception might depend to some extent on the composition of the stimuli and therefore on context. Future work should investigate this potential context dependency in variance judgments more systematically and comprehensively.

Our results suggest that pairwise similarity, range, and variety also play an important role in people's intuitive conceptual representation of variance. Why might these characteristics be attractive cues for people judging variance in the environment? One possible answer is that they are easy to compute. Whereas complex calculations are needed to compute statistical variance, computing range and variety is much more straightforward. To assess range, it is merely necessary to identify and keep track of the highest and lowest numbers in a set of numbers. Similarly, to assess variety, one only needs to keep track of the frequency of distinct elements in the sample, a task that seems to pose little cognitive cost to the mind (Alba, Chromiak, Hasher, & Attig, 1980; Hasher & Zacks, 1984; Manis, Shedler, Jonides, & Nelson, 1993). What about pairwise distance? Although it is more complex to compute than variety or range, it may still be easier than statistical variance. One simple strategy for gauging pairwise distance that is applicable in the context of sequential sampling is to track by how much two adjacently sampled

numbers differ from each other. For example, if the number sequence is 50, 48, 52, 50, then the differences would be -2, 4, and -2, which would suggest that numbers are close to each other and that the set is therefore not very variable (several respondents in Study 1 mentioned such a strategy).

Another possible factor contributing to the psychological attractiveness of range, variety, and pairwise distance is that they may be useful cues when making predictions in natural environments. Consider a foraging animal. Range, variety, and pairwise distance carry information that is important for its decision which patch of land to visit. Range indicates the lower and upper limits of the amount of food that a patch can produce and thus implies that values outside of the range are very unlikely. Variety refers to the number of different amounts of food that a patch can produce within those limits. This implies that new, previously unseen amounts are unlikely. Pairwise distance refers to how closely the food amounts are to each other, which indicates which of the existing food amounts within the limits of the production of a particular patch are more likely. An fascinating issue for future research is to test the roles of range, variety, and pairwise distance in real-world prediction tasks.

Several prominent models of decision making, ranging from risky choice (e.g., risk–return model; Markowitz, 1952; Weber et al., 2004) to categorization (Tenenbaum & Griffiths, 2001), rely on statistical variance to describe human behavior. To the extent that people's judgments of variance deviate in systematic ways from statistical variance—as we have shown in this article—and are sensitive to cognitively more achievable variables such as range and variety, our findings suggest ways of developing psychologically more realistic models of decision making. Range and variety could be easily integrated into these models. For pairwise distance, by contrast, some modifications may be necessary. Specifically, findings by Beach and Scopp (1968) indicate that people weight smaller deviations more heavily than larger ones.

Finally, to the extent that decision making that involves assessments of stimulus variability is sensitive to range, variety, and pairwise distance—characteristics that, though related, are imperfect cues to statistical variance—this also implies a novel perspective on the rationality of

decision making. Statistical variance is a prominent element in statistics. Therefore, it may seem a natural starting point to use this index of environmental variability in models of cognition and choice (cf. Gigerenzer, 1991). As our studies and others show, however, statistical variance only partly captures how people actually respond to variability in the environment. People appear to pay attention to other sample characteristics, and might even rely on different characteristics in different contexts. This invites a further question: How should people estimate variability in the environment? Rather than relying on one specific notion of a "rational" response to variability—which implies a specific set of goals (or loss function)—future work might focus on how people construe and understand a given decision context and adjust their reliance on ecological cues for variance judgment accordingly (Szollosi & Newell, 2020).

Conclusion

According to the influential notion of the mind as an intuitive statistician, people's responses to variability of stimuli in the environment can be well described by the notion of statistical variance. Although it has long been known that statistical variance only approximates people's judgments of variance, there have been few attempts to understand which indices of variability underlie people's judgments of variance. By unpacking subjective characterizations and assessments of variability, our findings lay the groundwork for a more cognitively rooted approach to variance perception.

Acknowledgements

We are grateful for discussions with and comments by the members of the Center for Adaptive Rationality. We thank Susannah Goss and Deborah Ain for editing the manuscript.

References

Alba, J. W., Chromiak, W., Hasher, L., & Attig, M. S. (1980). Automatic encoding of category size information. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(4), 370–378. doi: 10.1037/0278-7393.6.4.370

Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360. doi: 10.1016/j.tics.2013.04.010

Beach, L. R., & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance*, *3*(2), 109–123. doi: 10.1016/0030-5073(68)90001-9

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526. doi: 10.3758/BF03193020

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*(1), 25–47.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465. doi: 10.1037/0033-295X.86.5.465

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for bayesian regression models. *The American Statistician*.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1–14.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case

of frequency of occurrence. *American Psychologist*, *39*(12), 1372–1388. doi: 10.1037/0003-066X.39.12.1372

Hills, T. T., & Pachur, T. (2012). Dynamic search and working memory in social recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 218.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, *7*(3), 513–534. doi: 10.1111/tops.12151

Jia, C., Carson, M. B., Wang, X., & Yu, J. (2018). Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition*, *76*, 691–703. doi: 10.1016/j.patcog.2017.09.045

Kareev, Y. (2004). On the perception of consistency. *Psychology of Learning and Motivation*, *44*, 261–286. doi: 10.1016/S0079-7421(03)44008-5

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. doi: 10.1093/oxfordjournals.pan.a004868

Konovalova, E., & Le Mens, G. (2018). Learning variability from experience. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1942–1947). Austin, TX: Cognitive Science Society.

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, *127*(1), 47–73. doi: 10.1037/rev0000160

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology*, *73*(4), 498–502. doi: 10.1037/h0024344

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, *57*(2), 165–188. doi: 10.1037/0022-3514.57.2.165

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*(2), 203–208. doi: 10.3758/BF03204766

Mane, K. K., & Börner, K. (2004). Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences*, *101*(SUPPL. 1), 5287–5290. doi: 10.1073/pnas.0307626100

Manis, M., Shedler, J., Jonides, J., & Nelson, T. E. (1993). Availability heuristic in judgments of set size and frequency of occurrence. *Journal of Personality and Social Psychology*, *65*(3), 448–457. doi: 10.1037/0022-3514.65.3.448

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, *7*(1), 77–91. doi: 10.1111/j.1540-6261.1952.tb01525.x

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press, Taylor & Francis Group.

Peterson, C. R., & Beach, L. R. (1967). Man as intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46. doi: 10.1037/h0024722

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychological Research and Reviews*, *3*(1), 5–18. doi: 10.1007/BF02686528

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Science Software*, *1*(3), Article 37. doi: 10.21105/joss.00037

Stan Development Team. (2020). *Rstan: The R interface to Stan.* Retrieved from `http://mc-stan.org`

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 893–907. doi: 10.1037//0278-7393.28.5.893

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008–1018. doi: doi.org/10.1016/j.tics.2020.09.005

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi: 10.1017/S0140525X01000061

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using

leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, *111*(2), 430–445. doi: 10.1037/0033-295X.111.2.430

Young, M. E., & Wasserman, E. A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(1), 278–293. doi: 10.1037/0278-7393.27.1.278