The Quarterly Journal of Experimental Psychology

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: https://www.tandfonline.com/loi/pqje20

Where did that come from?—Identifying the source of a sample

Marcus Lindskog

To cite this article: Marcus Lindskog (2015) Where did that come from?—Identifying the source of a sample, The Quarterly Journal of Experimental Psychology, 68:3, 499-522, DOI: 10.1080/17470218.2014.959534

To link to this article: https://doi.org/10.1080/17470218.2014.959534

Accepted author version posted online: 28 Aug 2014.
Published online: 01 Oct 2014.

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
Vol. 68, No. 3, 499–522, http://dx.doi.org/10.1080/17470218.2014.959534

Routledge
Taylor & Francis Group

Where did that come from?—Identifying the source of a sample

Marcus Lindskog

Department of Psychology, Uppsala University, Uppsala, Sweden

People's ability to summarize their knowledge of an observed numerical variable has been extensively studied. However, many real-life situations require people to go beyond summary statistics and infer which process or distribution has generated a sample. The present study investigates the extent to which people can make such inferences when the experienced variable is continuous and when they have had previous experience with instances of the variable. It also tests specific predictions derived from three possible cognitive processes of how inferences about a generating distribution are made. The results indicate that participants are efficient and flexible intuitive statisticians, requiring only as little as four observations in a sample to successfully infer which distribution it came from. Further, the results indicate that the cognitive process supporting the inference uses statistical properties of both an experienced distribution and a presented test sample, as suggested by the Naïve Sampling Model (NSM).

Keywords: Intuitive statistics; Sample; Inference; Naïve intuitive statistician; Naïve sampling model

The information that people experience in everyday life is seldom complete, and often they have to settle for a small sample of data as a basis for judgements and decisions. The present study investigates the cognitive process underlying how people make inferences from sparse data within the framework of the naïve intuitive statistician (Fiedler, 2000; Fiedler & Juslin, 2006; Juslin, Winman, & Hansson, 2007; Lindskog, Winman, & Juslin, 2013a, 2013b).

Intuitive statistical judgements

Research investigating people's ability to estimate statistical properties of experienced data has mainly focused on the accuracy of estimates of descriptive properties, such as central tendency and variability. The general conclusion from research concerned with inductive inference in controlled laboratory settings has been that while estimates of some properties (e.g., central tendency) are accurate, estimates of others (e.g., variability) are consistently biased (Kareev, Arnon, & Horwitz-Zeliger, 2002; Lovie, 1978; Lovie & Lovie, 1976; Pollard, 1984). Further, several investigations concerned with general knowledge acquired outside of the laboratory have suggested that judgements of statistical properties are the result of fallible heuristics and prone to biases (Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974).

Correspondence should be addressed to Marcus Lindskog, Department of Psychology, Uppsala University, Uppsala, Sweden. E-mail: marcus.lindskog@psyk.uu.se

The author is indebted to Anders Winman, Peter Juslin, Håkan Nilsson, and Hedvig Söderlund for reading and commenting on earlier versions of the manuscript and to Johnny Halldin for help with the data collection of Experiments 1 and 3

The Swedish Research Council sponsored this research.

People's knowledge of higher order properties, like distribution shape, has been studied using variables supposedly encountered in everyday life (Fox & Thornton, 1993; Griffiths & Tenenbaum, 2006; Jako & Murphy, 1990; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978; Linville, Fischer, & Salovey, 1989; Nisbett, Krantz, Jepson, & Kunda, 1983; Nisbett & Kunda, 1985) and variables experienced on a trial-by-trial basis in controlled laboratory settings (Griffiths & Tenenbaum, 2011; Lindskog et al., 2013a, 2013b). The results are mixed. Some studies indicate remarkably accurate knowledge (e.g., Griffiths & Tenenbaum, 2006), whereas others suggest that people are biased by external information in the environment (e.g., Galesic, Olsson, & Rieskamp, 2012; Lichtenstein et al., 1978), by their own location in the distribution (Fiedler, 2000; Nisbett & Kunda, 1985), or by inherently biased estimators of the underlying distribution (Lindskog et al., 2013b). Further, even though people seem to have enough knowledge of distributions to make accurate predictions of future events (Griffiths & Tenenbaum, 2011), the accuracy of predictions are influenced by the shape of the underlying distribution (Lindskog et al., 2013b).

Judgements and decisions informed by samples

The research presented above indicates that people, at least under some conditions and for some properties, are equipped with an ability to summarize their knowledge of numerical values. However, many tasks require people to go beyond summary statistics and use the information in a sample to infer something about the underlying distribution. Indeed, several accounts of human cognition assume an internal sampling of information from memory prior to making a judgement or decision (e.g., Busemeyer, Myung, & McDaniel, 1993; Dougherty, Gettys, & Ogden, 1999; Dougherty & Hunter, 2003; Fiedler, 2000; Fiedler & Juslin, 2006; Gaissmaier, Schooler, & Rieskamp, 2006; Hansson, Rönnlund, Juslin, & Nilsson, 2008; Kahneman & Miller, 1986; Kareev et al., 2002; Lindskog et al., 2013b; Thomas, Dougherty, Sprenger, & Harbison, 2008). The naïve sampling model (NSM; Juslin et al., 2007; Lindskog et al., 2013b), for example, suggests that people use properties of a small internally generated sample as a proxy for the properties of the population from which the sample originates. More specifically, it has been shown that both judgements of confidence intervals (Juslin et al., 2007) and point estimates for unknown objects (Lindskog et al., 2013b) are informed by the statistical properties of a small internally generated sample. Similarly, in decision by sampling theory (Stewart, Chater, & Brown, 2006) the value of a target is determined by its relative rank in a small sample from memory. It is assumed that the sample from memory reflects both a distribution of values in a specific context and the underlying real-world distribution (Stewart et al., 2006). Thus, using the relative rank of a target within a series of internally generated samples will give information about where the target is positioned in the overall distribution and thereby its value. The internally generated samples thereby convey information about the statistical properties of the underlying distribution. Similarly, Dougherty and Hunter (2003) showed that when participants were to estimate the likelihood that a particular menu item would be ordered, their judgements were made relative to alternatives retrieved from long-term memory. Again, the alternatives retrieved from memory give information about the underlying distribution, which, in this case, can be used when judging likelihood. Finally, recent Bayesian accounts of human cognition have suggested that prior distributions are formed by drawing a limited number of samples from memory (Vul, Goodman, Griffiths, & Tenenbaum, 2009).

The cognitive processes that people use are sometimes considered to be adaptations to the environment in which they operate (e.g., Anderson, 1991; Brunswik, 1955). The observation that people in various tasks infer population properties from internally generated samples may therefore suggest that this process has evolved because it works well in several situations. This might, in turn, indicate that using the information in a sample to infer population properties is

something that people can do with reasonable accuracy. Studies investigating whether this is the case, however, report mixed results. Under some conditions, both infants (Gweon, Tenenbaum, & Schulz, 2010; Xu & Denison, 2009; Xu & Garcia, 2008) and adults (e.g., Evans & Pollard, 1985) show an ability to infer population properties from samples and even seem to take complex features of the sampling process into account (Gweon et al., 2010). However, a substantial body of research suggests that people are naïve with respect to several aspects of the processes that shape samples. For example, even though people under some conditions acknowledge that larger samples contain more information than smaller samples (Bar-Hillel, 1979; Chesney & Obrecht, 2012; Evans & Dusoir, 1977; Obrecht, Chapman, & Gelman, 2007; Obrecht & Chesney, 2013; Sedlmeier, 1998; Sedlmeier & Gigerenzer, 1997), they do not always integrate this information in their decisions (Evans & Pollard, 1985; Kahneman & Tversky, 1972; Obrecht et al., 2007). In addition, people seem to be naïve with respect to the conclusions that can be drawn from a sample (Fiedler, 2000; Fiedler & Juslin, 2006; Kareev et al., 2002; Lindskog et al., 2013b). Failing to appropriately evaluate the representativeness of a sample, for example, can lead to a number of apparent judgement biases (e.g., Fiedler, 2000; Fiedler & Juslin, 2006; Kareev et al., 2002; Lindskog et al., 2013b).

Features of inference tasks

Many of the studies concerned with people's ability to make inferences from small samples share three task features (see e.g., Beach, Wise, & Barclay, 1970; Griffin & Tversky, 1992; Phillips & Edwards, 1966). First, they predominantly use a binomial distribution. For example, a sample of red and white chips might be drawn from an urn with an unknown proportion of red and white chips, and participants are required to infer these unknown proportions. Even though the binomial case is interesting, most data sets that people experience in everyday life come from continuous distributions. Second, first-hand experience with the underlying distribution is generally withheld from participants. That is, prior to being presented with information about the sample, they have not been shown any instances from the distribution. While situations where little or nothing is known about the underlying distribution are not uncommon, people often have some prior knowledge or experience of the underlying distribution before the sample is presented. Finally, the sample is often presented to participants in a written summary description rather than by displaying all individual values of the sample. That is, participants might receive the information that a sample contains four red and three white chips in a written summary statement rather than observing each chip in the sample. Information in everyday life is, however, seldom experienced in descriptive summaries of data, and decisions based on description have been shown to deviate from equivalent decisions based on experience (e.g., Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Pleskac, 2010). In addition to these features, the task can be set up either as a yes/no recognition task where participants are asked to infer whether a sample has been drawn from a distribution or as a comparison task where the task is to infer which of two, or more, samples has been drawn from a distribution.

To address the possible limitations of previous research, the present study uses task features that are more similar to how people are expected to experience data in real-world situations. This is done by using a continuous variable, by allowing participants to experience several values from the underlying distribution prior to making inferences, and by letting participants make inferences from samples with all observations in the sample present.

The naïve sampling model

Research evaluating people's ability to act as intuitive statisticians has primarily emphasized the degree to which judgements conform to the normative rules of statistics and probability theory, while largely disregarding the nature of the cognitive processes that lead up to a judgement. Recently, however, a series of related studies have outlined a framework for intuitive statistical judgements where people are considered naïve intuitive

statisticians (Fiedler, 2000; Fiedler & Juslin, 2006; Juslin et al., 2007; Lindskog et al., 2013a, 2013b). Within this framework, the NSM (Juslin et al., 2007) has been proposed as a process model for how some intuitive statistical judgements are formed. The NSM suggests that the generic process that people use to realize their knowledge of an experienced variable is post hoc sampling from long-term memory (LTM; Juslin et al., 2007). That is, intuitive statistical judgements are computed on a small sample of observations retrieved from memory at the time of judgement (Juslin et al., 2007; Lindskog et al., 2013a, 2013b). As a consequence, judgements will be influenced by constraints on the cognitive processes. For example, the samples of data retrieved from memory will have to be of a size that can be activated within working memory constraints. Further, there is extensive support for the notion that controlled judgements are sequential and additive (Anderson, 1991; Hogarth & Einhorn, 1992; Juslin, Karlsson, & Olsson, 2008; Juslin, Nilsson, Winman, & Lindskog, 2011; Nilsson, Winman, Juslin, & Hansson, 2009). This means that information in the environment is integrated by considering one piece of information at a time and that the predominant operation used to integrate previous knowledge with new data is additive in its nature. Thus, the information integration of a naïve intuitive statistician is likely to be constrained by the sequential real-time properties of a controlled judgement process (Juslin et al., 2007). Further, to create estimates of statistical properties, people will tend to use properties of the retrieved sample as a proxy for population properties. As is known from introductory statistics, some sample properties (e.g., mean and proportion) are unbiased estimators of population properties while others (e.g., variance and coverage) are not. Accordingly, which has been shown repeatedly in previous research (e.g., Pollard, 1984), people's estimates of the former will be more accurate than those of the latter.

The cognitive process of inference

Consider the following example to appreciate the type of inference task that the participants in the studies reported below are asked to do. Imagine working in a factory with two production lines (A and B), both producing the same type of chocolate bar differing only in size. Working at Line A you only experience bars from that line and have no experience of those from the other. While your line (A) is set up to satisfy the European market and produces bars the size of which follows a bimodal distribution with either large or small bars, the other (B) is set up to satisfy the American market and produces bars the size of which follows a unimodal distribution. The bars from each line are put in boxes of five bars, labelled with Europe or USA. Now, one day the labelling machine breaks down, and, equipped only with the sample of bars in each box and experience from your work at Line A, you need to decide whether a particular box should be shipped to Europe or the US. Thus, your task is to infer, equipped with the knowledge about the distribution of bars that Line A generates, which of the two processes (Line A or Line B) has generated the sample in a particular box.

People could respond to this inference task by means of one of at least three different, possible cognitive processes. The first process assumes that experienced values are stored in and retrieved from long-term memory and are compared to the concrete and specific values in a test sample. The other two processes are somewhat more sophisticated and assume that statistical properties from both the experienced distribution and the test sample are induced and used in the decision. The inference task, and the associated processes outlined below, all include three sets of values. The *objective distribution* (OD) describes the set of all values of a given variable. During learning, participants experience a subset of values, the *experienced distribution* (ED), from the OD. It is possible that the ED is either a representative or a biased subset of the OD. However, in the experiments reported below, the ED presented to participants is always representative of the OD. Finally, during test participants are presented with a *test sample* (TS) and are asked to infer whether the TS comes from the same OD as the ED.

Memory inference

A first possibility is that people make an inference by matching values in the test sample against values in the experienced distribution. A memory inference stores the values from the experienced distribution in long-term memory in the form of a *subjective distribution* (SUD). The SUD could possibly contain all the values of the ED but it is reasonable to assume that only a subset of values of the ED will actually be remembered. A later inference will be made by matching the values in the TS with the values in the SUD. The process of inferences would thus have participants decide whether a test sample has been drawn from the OD based on the proportion of matches between the values in the SUD and the values in the TS. For example, a worker at the chocolate factory presented with two boxes containing 10 chocolate bars each would count the proportion of bars in each box that are equal in size to the sizes stored in memory and conclude that the box with the highest proportion of matches is the one originating from the objective distribution. That is, a sufficient proportion of matches would indicate that the OD has generated the TS.

More formally, let S denote the set of values in the OD, and let s_s denote the subset of values from S that have been stored in LTM during exposure to S (i.e., the SUD). Further, let s_t be a set of values (TS) that potentially may have been drawn from the same distribution as S. To infer whether s_t has been drawn from the same distribution as S, the memory inference compares each value k of s_s to each value j of s_t and calculates the proportion of matches. Let $M(s_{s,k} : s_{t,j})$ denote a matching function, which takes value 1 if $s_{s,k} = s_{t,j}$ and 0 otherwise, and let n_{st} be the number of values in s_t. Then $PM = \sum_k \sum_j M(s_{s,k} : s_{t,j})/n_{st}$ is the proportion of matches. In a yes/no recognition task, the memory inference will have participants conclude that s_t has been drawn from the same distribution as S when $\mathrm{PM} > \theta$, where θ is some critical proportion of matches. In a comparison task with two samples, as in the experiments reported below, the memory inference will have participants conclude that sample s_1 rather than sample s_2 has been drawn from the same distribution as S if $\mathrm{PM}_1 > \mathrm{PM}_2$—that is, if the proportion of matches is larger for sample s_1 than for sample s_2.

The memory inference would be inefficient if the test sample includes values from the OD not previously experienced or if any noise is added to the values stored in memory during exposure. The two processes described next are more flexible with respect to these situations in that they both derive estimates of statistical properties of both the experienced distribution and the test sample rather than rely on the exact matching of values.

Inference from a large-sample representation

A second possibility is that an inference is made from a precomputed representation based on a large sample of the experienced variable. Precomputed large-sample representations could be generated from explicit attempts from a participant to abstract statistical properties during exposure (i.e., online) to the variable as when someone is keeping and updating a running mean as new observations are made. The representation could also, hypothetically, arise from corresponding preconscious and automatic computations (Zacks & Hasher, 2002). When presented with a test sample and asked to make an inference, people would have to base their judgement on the statistical properties of the large-sample representation because little or no information about the specific values is retained. Previous research indicates that central tendency and variability (e.g., Pollard, 1984) are properties that could potentially be included in the inference.

Formally, let S be the set of values of the OD, and let s_e be the values of the ED—that is, the subset of values from S that the participants experience. The values of s_e will serve as the basis for estimating the statistical properties of S. During exposure, a participant would engage in an online estimation of, for example, the mean (μ_S) and standard deviation (σ_S) of S by updating a running mean (μ_{se}) and standard deviation (σ_{se}) from the values of s_e as each new value is presented. The majority of the individual values of s_e will be disregarded while the statistical estimates will be stored in LTM. At the time of judgement the properties

of s_e will be compared to those of the test sample. In the case of a yes/no recognition task including one sample (s_1), it will be inferred that this sample comes from S if $SP(s_e, s_1) < \varepsilon$ where $SP(s_x, s_y)$ is a function that calculates the deviation between the statistical properties of two sets of values (s_x and s_y) and is given by

$$SP(s_x, s_y) = \theta \frac{\left|\mu_{sx} - \mu_{sy}\right|}{\mu_{sx}} + (1 - \theta)\frac{\left|\sigma_{sx} - \sigma_{sy}\right|}{\sigma_{sx}}, \quad (1)$$

with $0 \leq \theta \leq 1$. The division by μ_{sx} and σ_{sx} in the first and second terms, respectively, of Equation 1 is to ensure that both terms will be on equivalent similarity scales. In the case of a comparison task with two samples (s_1 and s_2), the inference that s_1 rather than s_2 has been drawn from S will be made if $SP(s_e, s_1) < SP(s_e, s_2)$.

According to the large-sample account, our chocolate factory worker would not store the size of each and every one of the bars passing by on Line A in memory. Rather, he or she would keep estimates of the mean size and the variability of the size in memory. As each new bar passes by, these estimates would be updated. To infer whether a box has been generated by Line A, our worker would compare the statistical properties stored in memory with those of the box. A small deviation for the compared properties would suggest that the box comes from Line A.

Inference from a small-sample representation

Finally, a third possibility follows from the NSM. The NSM suggests that intuitive statistical judgements are based on a small sample of values in a *memory sample* (MS), the properties of which are considered to be proxies for the properties of the population distribution. In principle, it is possible that all of the values of the SUD are retrieved at the time of judgement. However, we follow research suggesting that online judgements are constrained by short-term memory (e.g., Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Kareev et al., 2002; Stewart et al., 2006) and therefore that the MS will contain approximately 4 ± 2 observations (Cowan, 2000). Thus, in contrast to a large-sample representation, inferences from a small-sample representation use the properties of the MS rather than those of the ED when performing the inference. In all other aspects the inferences are equivalent.

The process could formally be described as follows. Let S and s_s denote the set of values in the OD and SUD, respectively. Further, let s_m be a sample of approximately four values drawn randomly from the s_s. In the case of yes/no recognition task with one sample (s_1), it will be inferred that s_1 has been drawn from S if $SP(s_m, s_1) < \varepsilon$ (see Equation 1). Further, in the two-sample (s_1 and s_2) comparison task case it will be inferred that s_1 rather than s_2 has been drawn from S if $SP(s_m, s_1) < SP(s_m, s_2)$.

Predictions

The following section derives specific predictions if one or the other of the above suggested processes are used to solve the inference task. While the two latter accounts suggest that people use statistical properties of the test sample to infer which distribution it has been drawn from, the memory inference does not. As such, the memory inference will be sensitive to whether or not the values in the test samples have been previously experienced or not, and there should be a distinct difference in performance depending on whether the data points are new or old, with better performance for old than for new data points. In contrast, if judgements are based on statistical properties of the test samples, all other things equal, performance is predicted to be equally good when the test samples contain new as when they contain old data points from the OD. Another prediction from the memory inference model is that performance will be independent of the shape of the distribution, or of the statistical properties of the distribution, from which the values are drawn.

The statistical properties of the test sample will reflect the "true" distribution of values with more or less precision. This precision is denoted by the

statistical term of standard error. Standard error is determined by the variance of the distribution and the size of the sample. Samples drawn from populations with low variance (e.g., unimodal rather than bimodal) and consisting of a large number of observations will have smaller standard errors than others. In addition to the standard error of the test sample, participants that rely on a small-sample representation will suffer from a second source of error; the sample drawn from long-term memory will reflect the total pool of values stored in memory with a certain precision, also determined by a standard error.

In the small-sample account, the properties (mean and standard deviation, see Equation 1) of the MS and the TS are compared. From this comparison follows the general prediction that performance will increase as the standard errors of the MS and TS decrease. This occurs because when standard errors are small the difference between the sample means and sample standard deviation in the MS and TS will be small, if they are drawn from the same distribution.

Previous research has indicated that statistical judgements based on a small-sample representation are more accurate for variables with a unimodal than with a bimodal distribution (Lindskog et al., 2013a, 2013b). However, this has not been tested for the task of inferring the generating distribution of a sample. Because, other things equal, samples from a unimodal distribution have smaller standard errors than those from a bimodal distribution, a small-sample account predicts a higher proportion of correct responses when the underlying distribution is unimodal than when it is bimodal. Notably, this should occur even when the two distributions are equally well learnt. Because a large-sample representation is similar to a small-sample representation with a large MS, the same unimodal–bimodal difference is predicted in the case of inferences being based on large-sample representations. However, the effect of distribution should be smaller than in the large-sample case because it is only the standard errors of the TS that are affected by the underlying distribution. In contrast, there should be no such differences between the two distributions for a memory inference.

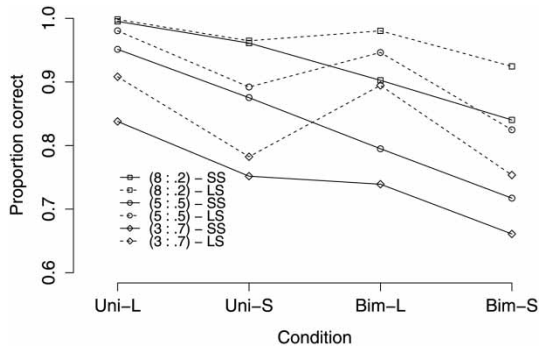

Figure 1. *Predicted proportion correct in the four experimental conditions [unimodal (uni) or bimodal (bim) distribution with large (L) or small (S) sample sizes] under three different sets of parameter values for the underlying distributions [beta(8, 8) vs. beta(.2, .2); beta(5, 5) vs. beta(.5, .5); beta(3, 3) vs. beta(.7, .7)]. Dashed lines depict predicted performance for the large-sample (LS) account while solid lines depict predicted performance for the small-sample (SS) account.*

In general, larger samples include more information than smaller samples. It is therefore reasonable to expect that people should find it easier to solve the inference task if the test samples are large rather than small (i.e., contain many as opposed to few data points). In the large-sample and small-sample case, this occurs because the standard errors of the TS become smaller as sample size increases. Notice that because sample proportion is an unbiased estimator of population proportion there should be no effect of sample size if participants use a memory inference.

The three previous predictions are able to differentiate between the memory inference and the two accounts based on sample representation. They do not, however, differentiate the large-sample and small-sample account. Under certain conditions, the two-sample accounts do, however, predict qualitatively different patterns in performance. Figure 1 illustrates the predicted performance of the two accounts, derived from computer simulations, under a set of conditions including those used in the experiments reported below [distributions: unimodal: beta(5, 5) and bimodal: beta (.5, .5); sample sizes: small: 5 or 4 and large: 10 or 8]. Details of the simulations are presented in the Appendix. The figure illustrates both the

effects of distribution and sample size predicted for the large-sample and small-sample accounts. More importantly, the figure illustrates a qualitative difference in the predicted pattern of performance in the unimodal–small and bimodal–large conditions. While the small sample account predicts better performance in the unimodal–small than the bimodal–large condition, the large-sample account makes the opposite prediction. The different patterns occur because a change in the variability of the underlying distribution has less impact on a large-sample than a small-sample representation and can therefore be compensated by an increase in sample size.

THE PRESENT STUDY

The aim of the present study was to address three main questions. First, to what extent are people able to solve an inference task that uses a continuous variable, allows them to experience values from the objective distribution, and allows them to experience all values in the sample? Second, do people use the statistical properties of the objective distribution and the test samples to solve the inference task or are their judgements based on a memory inference? Finally, does the possible use of statistical properties involve a large sample or a small sample (similar to the NSM) representation?

In Experiments 1 and 2, participants learned the distribution of a single variable from trial-by-trial experience. They were later asked to identify which of two test samples had been drawn from the experienced distribution. Experiment 1 was designed to investigate the influence of sample size and the shape of the underlying distribution on the accuracy of inference. By allowing both old and new values in the test samples used in Experiment 2, this experiment investigated the role of memory processes in solving the inference task.

In Experiment 3, participants learned the distributions of two variables simultaneously. Later, they were asked to identify which of two test samples came from either of the two variables. Thus, Experiment 3 investigated how the ability to draw inferences from samples is influenced by the presence of multiple distributions.

EXPERIMENT 1: SAMPLE SIZE AND DISTRIBUTION

During learning, participants observed 100 consumer ratings for a fictitious product with either a bimodal or a unimodal distribution of the ratings. They were later presented with two test samples of consumer ratings, both containing either 5 or 10 values, and were asked to decide which of the two had been generated by the same distribution seen during learning. Distribution (unimodal/bimodal) was manipulated between subjects and sample size (5/10) within subjects. The experiment was thus designed to investigate the influence of sample size and the shape of the objective distribution on performance in the inference task. To estimate participants' knowledge of the objective distribution, they also reproduced the distribution by means of frequency estimates and gave estimates of central tendency and variability. Finally, to control for proficiency to handle numbers and an ability to remember numbers, measures of numeracy and long-term memory for numbers were collected.

Method

Participants

Participants were 36 undergraduate students (15 male and 21 female) from Uppsala University ($M = 22.8$ years, $SD = 3.7$). They received a movie voucher or course credits as compensation for participating in the study.

Materials and procedure

The computerized task consisted of a *learning phase* and a *test phase*. On each trial of the learning phase, participants observed a numerical value between 1 and 1000 described as a consumer rating of a fictitious product. The cover story informed participants that the ratings came from a market survey using a representative sample, and participants were instructed to observe the ratings carefully in

Table 1. *Characteristics in terms of central tendency, range, and variability of the sets of values used in the three experiments*

			Central tendency		*Range*		*Variability*	
Experiment	*Distribution*	*Sample type*	*Mean*	*Median*	*Min*	*Max*	SD	*MAD*
Exp. 1 and 3	Bimodal		498	500	3	997	345	311
	Unimodal		500	502	122	883	160	130
Exp. 2	Bimodal	Old	502	506	7	991	351	325
	Bimodal	New	500	468	14	996	352	325
	Unimodal	Old	499	499	35	1000	210	173
	Unimodal	New	499	500	13	24	213	175

Note: MAD = mean absolute deviation.

order to be able to give a first advisory opinion on the outcome of the market survey. Two sets of 100 unique values, drawn from a symmetric unimodal distribution [beta distribution with $\alpha = \beta = 5$, beta(5, 5)] and a symmetric bimodal distribution [beta(.5, .5)] and transformed to the range [1, 1000], defined the unimodal and bimodal conditions, respectively. The properties of the two sets are shown in Table 1. Each of the 100 values was observed once in an individually randomized order during learning and was presented together with a fictitious consumer identification tag. The presentation was self-paced, but each value remained on the screen for a minimum of 2 s before participants could proceed to the next trial.

In the test phase, participants completed a sample identification task and a production task, the order of which was counterbalanced, and gave estimates of descriptive statistics (central tendency and variability). Finally, after the main experiment, participants performed a long-term memory test for numbers and completed a questionnaire measuring numeracy.

Sample identification. On each of the 40 trials of the sample identification task, participants were presented with two test samples (Sample 1 and Sample 2) of consumer ratings and were to decide which of the two came from the distribution of values experienced during learning. The two test samples on each trial had an equal number of values; half of the trials had 10 values in each sample, and half had 5. One of the samples (target sample) consisted of values randomly drawn from the set of values seen during learning, while the other (distractor sample) included values drawn from the set of values not seen during learning. That is, a participant in the unimodal condition would have target samples drawn from the unimodal set and distractor samples drawn from the bimodal set, while a participant in the bimodal condition would have the opposite. The presentation of the test samples is illustrated in Figure 2. Which of the two test samples (Sample 1 or 2) that was the target and distractor, respectively, was randomized for each trial. Participants made their decision by pressing the button located directly below the chosen sample.

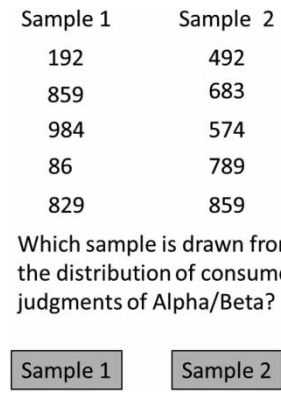

Figure 2. *Illustration of the presentation of the two test samples in the sample identification task.*

Prior to the task, we took care to explain how the instructions should be interpreted. If a test sample came from the experienced distribution, participants were told that it should have the same

properties as the experienced distribution but that this might occur without any one single value being recognized. Conversely, participants were instructed that recognizing one or more values would not necessarily be sufficient for concluding that the sample came from the same distribution.

Production. In the production task, participants were given 10 equally wide intervals on the range [1, 1000] (1–100, 101–200, . . . 901–1000) and were to state how many of the 100 consumer ratings from the learning phase fell into each interval. Frequencies were required to sum to 100.

Descriptive statistics. Participants gave estimates of central tendency (mean and median) and variability (mean absolute deviation) for the consumer ratings. The estimates were preceded by a brief definition of the measures in terms of an explicit exemplification of its calculation (e.g., "The mean for a set of numbers is the sum of the numbers divided by the count of the numbers. For example, the mean of 4, 8, 12 is 8 because $(4 + 8 + 12)/3 = 8$."). Mean absolute deviation was explained as "The mean of the distances of each value from their total mean").

Proficiency to handle numbers. After the main experiment, participants carried out a test designed to measure long-term memory capacity for numbers. In addition they completed a questionnaire measuring numeracy.

Long-term memory for numbers. In the test for long-term memory for numbers, participants saw 30 numbers presented individually for five seconds each during an exposure phase. Fifteen of the numbers were two-digit numbers (e.g., 45), and 15 were three-digit numbers (e.g., 543). They were told that their memory for these numbers would be tested later. In a recognition test, participants were shown 60 numbers, half of which they had seen during exposure, and were to decide if the number had been shown during exposure or not. The retention time was approximately 25–30 min. During the retention time, participants filled out a numeracy questionnaire and completed a set of unrelated tasks not reported here.

Table 2. *Performance in terms of mean proportion correct in the four sample size by distribution conditions in Experiments 1 and 2 and pooled over the two experiments*

	Unimodal		*Bimodal*	
Experiment	*Large*	*Small*	*Large*	*Small*
Exp. 1	.99 (.02)	.98 (.05)	.87 (.15)	.76 (.19)
Exp. 2	.87 (.17)	.85 (.13)	.76 (.27)	.68 (.26)
Exp. 1 and Exp. 2	.94 (.13)	.92 (.11)	.82 (.22)	.73 (.22)

Note: Standard deviations in parentheses.

Numeracy questionnaire. Participants completed a questionnaire consisting of 11 items measuring numeracy. The questionnaire was a Swedish translation of the questionnaire used by Lipkus, Samsa, and Rimer (2001; see also, Lipkus & Peters, 2009; Peters, Slovic, Västfjäll, & Mertz, 2008; Peters et al., 2006).

Design

The experiment used a mixed design with distribution (unimodal/bimodal) as independent between-subjects variable and sample size [small (5)/large (10)] as independent within-subjects variable. Participants were randomly assigned to the experimental conditions. The approximate length of the experiment was 120 min.

Results

The proportion of correct answers was used as a measure of performance in the sample identification task. Performance in the four conditions is summarized in Table 2. The three accounts of inference primarily make predictions with respect to the effects of *distribution* and *sample size* and with respect to the difference between the unimodal–small and bimodal–large conditions. Therefore, the following analyses focus on these three aspects of the data. As is evident from Table 2, the variance in the four conditions is not homogeneous. This fact suggests that standard parametric statistical tests should be avoided. In the following, nonparametric approaches are therefore used when analysing the data from the sample task.

Effect of distribution

Both the small-sample and the large-sample accounts predicted better performance when the underlying distribution is unimodal than when it is bimodal. In line with this prediction, participants performed significantly better (Wilcoxon: $W = 289.5$, $p < .001$) in the unimodal ($Mdn = 1$) than in the bimodal condition ($Mdn = .85$).

Effect of sample size

If participants made inferences from a small-sample representation or a large-sample representation it was predicted that they would perform better with large than with small samples. No such difference was predicted if participants use a memory inference. In accordance with the prediction from the two former accounts, performance was significantly better (Mann–Whitney: $V = 153$, $p < .001$) with large ($Mdn = 1$) than with small samples ($Mdn = .91$).

Unimodal–small versus bimodal–large

The large-sample and small-sample accounts predicted a critical difference between the unimodal–small and bimodal–large conditions. While the small-sample account predicted better performance in the unimodal–small than in the bimodal–large condition, the large-sample account predicted the opposite. Proportion correct in each of the four conditions is summarized in Table 2. As can be seen in the table, the critical difference is consistent with the prediction of a small-sample account, and comparing performance in the unimodal–small and bimodal–large conditions revealed a significant difference (Wilcoxon: $W = 234.5$, $p = .009$) in the direction predicted by the small-sample account.

Influence of knowledge of distribution properties

It is possible that the effect of distribution is due to participants in the bimodal condition having less accurate knowledge of the objective distribution than participants in the unimodal condition. To investigate this possibility, four separate measures of participants' knowledge of the objective distribution were calculated. First, we calculated the mean absolute error (MAE) between the rated and actual frequency in the production task. MAE is given by

$$\mathrm{MAE} = \frac{\sum_{n=1}^{10} |r_n - a_n|}{10}, \qquad (2)$$

where r_n and a_n are the rated and actual frequencies of interval n, respectively. Second, three measures of knowledge of the distribution parameters were calculated as the signed deviation between actual and estimated parameter values of mean, median, and mean absolute deviation (MAD). Each measure was scanned for outlier responses ($|z| > 2.5$) but no data points had to be excluded from the analyses. Figure 3 presents the average assessed proportion consumer ratings (grey bars) in each interval of the production task together with the proportions from the objective distribution (black bars) for the bimodal (Figure 3a) and unimodal (Figure 3b) conditions, respectively. The figure illustrates that participants in both conditions reproduced the overall shape of the distribution quite well. With respect to MAE, however, the difference between the two conditions approached significance [$t(34) = 2.00$, $p = .054$; bimodal: $M = 3.89$, $SD = 1.75$ vs. unimodal: $M = 2.96$, $SD = 0.94$]. The difference between the two conditions for estimates of MAD was significant [$t(34) = 3.08$, $p = .004$; bimodal: $M = -123.8$, $SD = 138.8$ vs. unimodal: $M = 7.6$, $SD = 115.3$]. There was, however, no difference between the two distribution conditions with respect to estimates of central tendency (mean and median; both ps $> .62$). Accordingly, MAE and the signed deviation for MAD were entered as covariates into a permutation test for one-way analysis of covariance (ANCOVA) with distribution (unimodal/bimodal) as between-subjects independent variable and proportion correct as dependent variable.[1] The analysis showed that effect of distribution ($p < .005$) remained after controlling for the covariates.

Influence of individual differences. In both the LTM task and the numeracy test we calculated the

[1]Permutation tests are a class of nonparametric test that rely on permutation and resampling techniques and make minimal assumptions about the data.

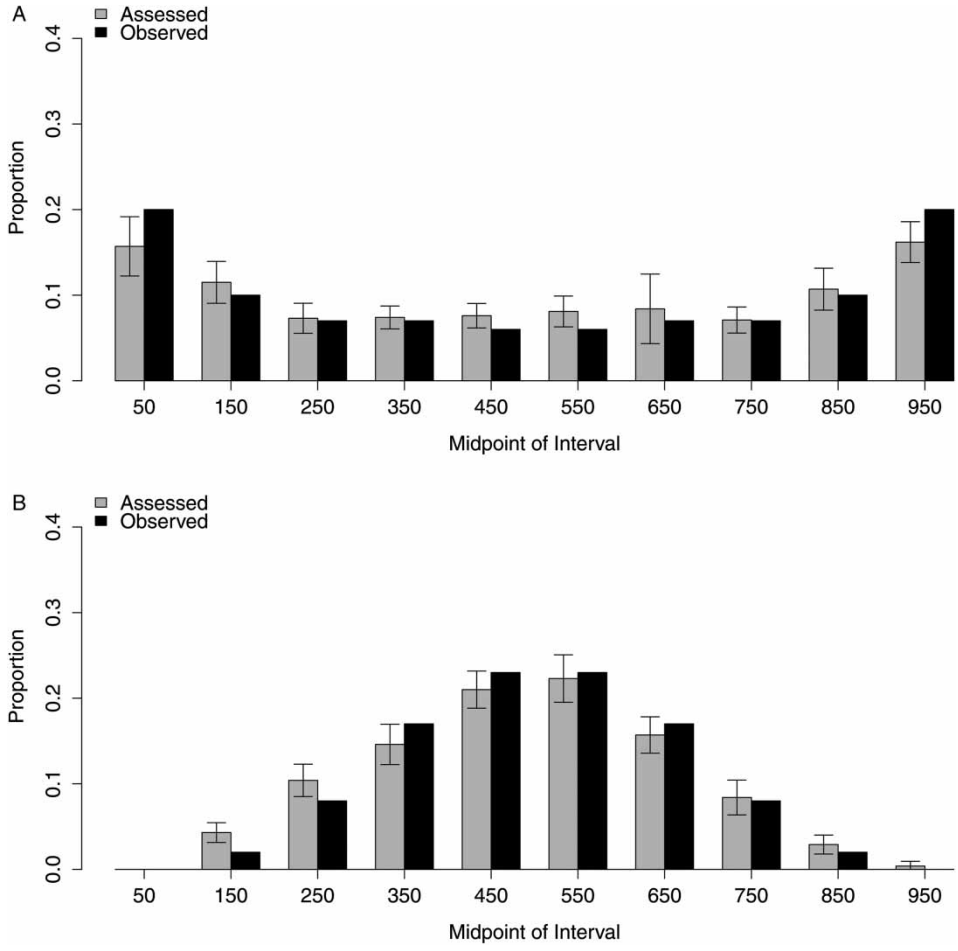

Figure 3. *Assessed proportion of consumer ratings (grey bars) and the underlying distribution (black bars) for the bimodal (Panel A) and the unimodal (Panel B) conditions separately in Experiment 1. Whiskers denote 95% confidence intervals.*

number of correct responses as a measure of performance. The relationship between individual measures of long-term memory for numbers and numeracy and performance in the sample identification task was investigated by two separate Spearman correlations. Neither the LTM–sample relationship, $r_s(34) = .21$, $p = .22$, nor the numeracy–sample relationship, $r_s(33) = -.1$, $p = .62$, reached significance.

Discussion

Experiment 1 was designed to investigate the influence of sample size and the shape of the objective distribution on participants' ability to infer which of two test samples originated from an experienced distribution. Both the large-sample and small-sample accounts predicted a difference in performance depending on the shape of the objective distribution, while the memory inference did not. In line with this prediction, there was an effect of distribution in the sample task with better performance in the unimodal than in the bimodal condition. It is possible that the effect was due to the unimodal distribution being more easily learned than the bimodal distribution. However, the effect remained when controlling for two measures of distribution knowledge that differentiated between

the two conditions. Thus, the advantage for the unimodal distribution was not merely an effect of better knowledge. Further, neither the LTM–sample nor the numeracy–sample correlation reached significance, indicating that the difference was not related to an individual proficiency with numbers.

Both the large-sample and the small-sample account predicted an effect of sample size. Consistent with this prediction, participants performed significantly better with a large than with a small sample size. The two-sample accounts predicted different orders of the unimodal–small and bimodal–large conditions with respect to proportion correct. The observed pattern of results, as indicated by the significant difference between the unimodal–small and bimodal–large conditions, was consistent with a prediction from a small-sample but not a large-sample account.

In all four conditions participants performed very well. It is possible that such good performance, coming close to ceiling in some of the conditions, might influence the conclusions that can be drawn from Experiment 1. Even though the observed levels of performance are similar to those found in preliminary simulations of the models, and even though the ordering of the four conditions is consistent with the prediction from a small-sample account, it would be valuable to show that the effects are similar even under different conditions with levels of performance that are further away from the ceiling. In Experiment 2, reducing the sample size to four and eight values created such a situation.

Taken together, the results suggest that people use statistical properties of the test samples to make an inference about which of the two was drawn from the objective distribution. Further, the results support the notion that the inference is based on a small-sample, rather than a large-sample, representation. However, because the values in the test sample had all been observed during learning, it is still possible that the task could potentially be solved by a memory inference. Therefore, Experiment 2 was designed to investigate the possible use of a memory inference further by including samples with values not seen during exposure.

EXPERIMENT 2: OLD VERSUS NEW SAMPLES

In Experiment 2, participants learned the distributions as in Experiment 1. The test phase, however, included samples with both values seen during exposure (old) and values from the same distribution but not previously experienced (new). If people rely on a memory inference to solve the inference task a difference in performance between new and old samples was expected regardless of the objective distribution. Experiment 1 indicated good performance in the inference task already at a sample size of five values. To further investigate the boundary conditions of participants' ability to solve the task and to investigate a possible ceiling effect, sample sizes in Experiment 2 were reduced to four and eight, respectively. In Experiment 2, distribution (unimodal/bimodal) was manipulated between subjects while sample size (4/8) and sample type (new/old) were manipulated within subjects.

Method

Participants

Participants were 31 undergraduate students (10 male and 21 female) from Uppsala University ($M = 25.1$ years, $SD = 4$). They received a movie voucher or course credits as compensation for participating in the study.

Materials and procedure

The learning phase of Experiment 2 was the same as that of Experiment 1. The test phase used the same tasks (sample identification, production, and descriptive statistics) as those in Experiment 1, with a slightly altered sample identification task (described below). In addition to the main experiment, the long-term memory for numbers test was included as an additional behavioural measure while the numeracy measure was excluded. The properties of the stimulus material used in Experiment 2 are illustrated in Table 1.

Sample identification. Half of the 40 trials had samples with eight values, and half had samples

with four values. On half of the trials the values for the target samples were drawn from the 100 values seen during learning (old sample) while the other half of the trials had values drawn from the same distribution but values not seen during learning (new sample). As in Experiment 1, the distractor sample was always randomly drawn from the distribution not seen during learning.

Design

Experiment 2 used a mixed 2 × 2 × 2 design with distribution (unimodal/bimodal) as an independent between-subjects variable and sample size [small (4)/large (8)] and sample type (new/old) as independent within-subjects variables. Participants were randomly assigned to the experimental conditions, and the approximate length of the experiment was 120 minutes.

Results

As in Experiment 1, the measure of performance in the sample identification task was proportion correct. Even though, as can be seen in Table 2, the variance in the four conditions is closer to being homogeneous than they were in Experiment 1, the differences still motivate the use of nonparametric tests. Following the predictions of the three accounts, the analyses below focus on the effects of distribution, sample size, old versus new samples, and the ordering of the four conditions.

Effect of distribution

The small-sample and large-sample accounts both predicted better performance by participants in the unimodal condition than by those in the bimodal condition. The memory account predicted no such difference. In line with this prediction, performance was better in the unimodal ($Mdn = .93$) than in the bimodal ($Mdn = .86$) condition. However, the difference was only marginally significant (Wilcoxon: $W = 167.5$, $p = .06$).

Effect of sample size

In Experiment 2 the sample sizes were reduced to four and eight values in the test samples. Nevertheless, the small-sample and large-sample accounts predicted better performance with larger than with smaller samples. Comparing performance with the two sample sizes revealed significantly better (Mann–Whitney: $V = 71$, $p = .01$) performance with large ($Mdn = .95$) than with small samples ($Mdn = .85$).

Old versus new samples

The values in the test samples in the sample task of Experiment 2 were either new or old. If participants use a memory account to make inferences, there should be a difference in performance between the two sample types. More specifically, the use of a memory heuristic would predict better performance if samples are old than if they are new. In contrast to this prediction, there was no effect of the old/new manipulation (Mann–Whitney: $V = 140$, $p = .8$; old, $Mdn = .9$ vs. new, $Mdn = .9$)

Unimodal–small versus bimodal–large

The order with respect to proportion correct over the four sample size by distribution conditions is summarized in Table 2. From the table it is obvious that the order in both experiments follows what could be expected from a small-sample account. To achieve better statistical power, the data were collapsed over the two experiments. Figure 4 depicts the predictions from the small-sample and large-sample accounts [for underlying distributions of beta(5, 5) and beta(.5, .5)] together with the pooled data from Experiments 1 and 2. From the figure it is obvious that the data follow a pattern that could be predicted from a small-sample but not a large-sample account. Further, the critical difference between the unimodal–small and the bimodal–large conditions was significant, $t(65) = 2.25$, $p = .03$, in the direction predicted by the small-sample account.

Influence of knowledge of distribution properties

The same four measures of knowledge of distribution properties as those in Experiment 1 were calculated, and knowledge of the two distribution conditions was compared using four separate *t*-tests. Each measure was scanned for outlier

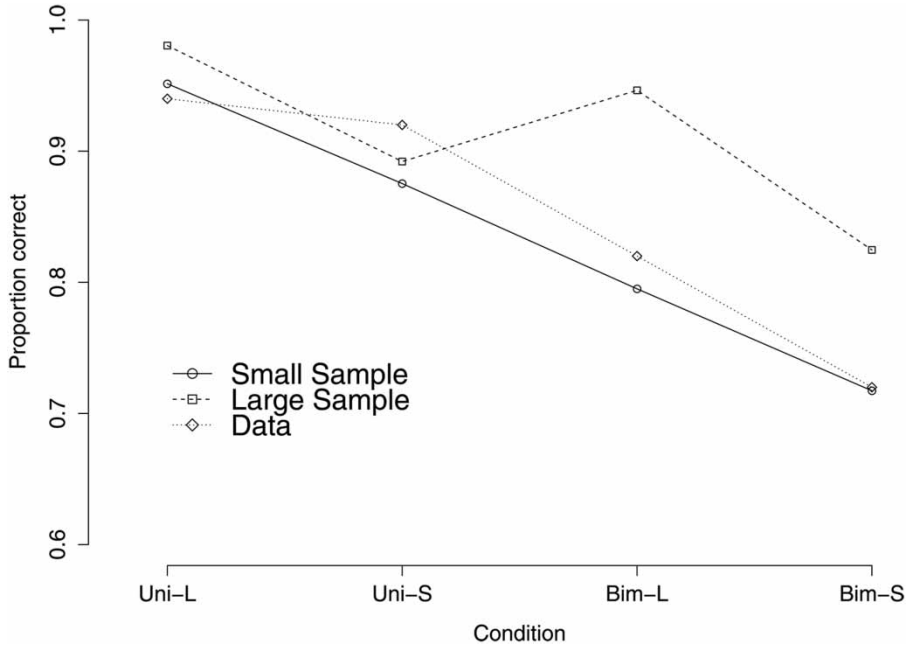

Figure 4. *Predicted proportion correct in the four experimental conditions [unimodal (uni) vs. bimodal (bim) with large (L) or small (S) sample sizes] under the parameters for the underlying distributions that was used in the current experiments [beta(5, 5) vs. beta(.5, .5)] for the small-sample and large-sample accounts together with the pooled data from Experiments 1 and 2.*

responses ($|z| > 2.5$), and as a result 2.5% of data points were excluded from the analyses. As illustrated in Figure 5, and similar to Experiment 1, participants reproduced the general shape of the underlying distribution in both conditions.

With respect to MAE, there was no difference between the two conditions [$t(29) = 1.05$, $p = .30$; bimodal: $M = 3.99$, $SD = 2.24$ vs. unimodal: $M = 3.24$, $SD = 1.63$]. The significant difference for estimates of MAD seen in Experiment 1 was replicated [$t(29) = 2.66$, $p = .012$; bimodal: $M = -174.1$, $SD = 127.7$ vs. unimodal: $M = -63.3$, $SD = 102.0$]. Both the difference for estimates of median, $t(27) = 1.92$, $p = .065$, and the difference for estimates of mean, $t(27) = 1.86$, $p = .074$, approached significance.

The possible influence of knowledge of distribution properties on performance in the sample identification task was accordingly examined by entering estimates of MAD, mean, and median as covariates into a permutation test for one-way

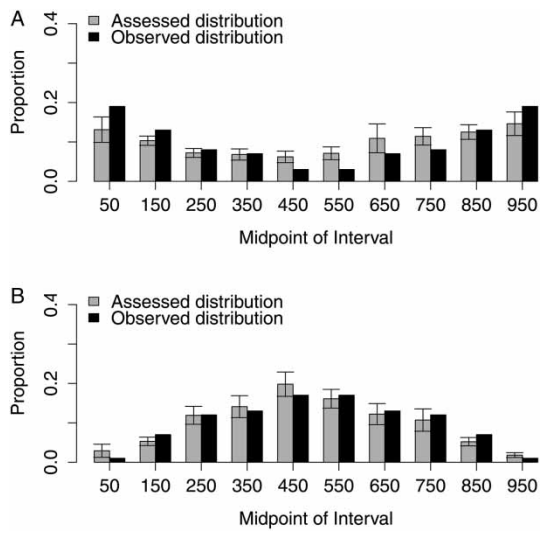

Figure 5. *Assessed proportion of consumer ratings (grey bars) and the underlying distribution (black bars) for the bimodal (Panel A) and the unimodal (Panel B) conditions separately in Experiment 2. Whiskers denote 95% confidence intervals.*

ANCOVA with distribution (unimodal/bimodal) as between-subjects independent variable and proportion correct as dependent variable. The results indicated that the marginally significant effect of distribution was reduced ($p = .15$) when controlling for the covariates.

Influence of individual differences

A Spearman correlation showed that there was no relationship between long-term memory for numbers and performance in the sample identification task, $r_s(28) = .16$, $p = .41$.

Discussion

The pattern of results seen in Experiment 1 indicated that people solve the sample identification task by a process that uses the statistical properties of both the test sample and the objective distribution rather than by a memory inference. Even though the memory inference cannot explain the observed effect of distribution seen in Experiment 1, the test samples used in the first experiment did not allow for ruling out the possibility of a memory inference. Therefore, to fully rule out the possible use of a memory inference, Experiment 2 included test samples with old values seen during training and new values not previously encountered. If participants use a memory inference, this should manifest itself as an effect of the old/new manipulation. There was no effect of old versus new values, indicating that participants do indeed use statistical properties rather than the individual values when choosing between test samples.

Experiment 2 replicated the effect of sample size, from Experiment 1, with better performance when samples contained eight values than when they contained four. This effect was predicted by the both the large-sample and small-sample account. Thus, and perhaps not that surprising, the more information contained in a sample the easier it is to identify its source. The overall performance in Experiment 2 was poorer than that in Experiment 1. This was expected because as the test samples become smaller it will be more difficult to infer which distribution they are drawn from due to the increase in the standard errors of the TS.

The two statistic-based accounts predicted different orders with respect to proportion correct over the unimodal–small and bimodal–large conditions. Pooling the data from both experiments gave support for the ordering suggested by a small-sample account as opposed to a large-sample account. Further, the critical difference between the unimodal–small and bimodal–large conditions was significant in the direction predicted by a small-sample account.

If inferences are based on statistical properties, an effect of distribution was expected. The effect of distribution was only marginally significant in the predicted direction. It is possible that this is due to the generally lower level of performance in Experiment 2. Participants did, however, perform above chance in all conditions, signalling an impressive ability to capitalize on small samples.

Concerning knowledge about the properties of the experienced values, there was no effect of distribution on estimates of central tendency. However, and replicating the results of Experiment 1, participants in the bimodal condition gave worse estimates of variability than did participants in the unimodal condition. Further, the estimates of variability underestimated the true variability to a large extent. These findings replicate previous research (Lindskog et al., 2013a) showing that when the stimulus is a continuous numerical variable the accuracy of variability estimates is dependent on the shape of the underlying distribution. The results are also consistent with the notion that intuitive statistical judgements are based on small samples.

The results from Experiments 1 and 2 suggest three major conclusions. First, people are well equipped to solve the inference task and need only a small amount of information in the samples to perform well above chance. Second, the ability to solve the task, independent of which strategy participants use, does not seem to be related to a more general proficiency to use or remember numbers. Finally, people seem to use the statistical properties of the objective distribution and the test samples to solve the inference

task rather than base their judgements on a memory inference. Further, the pattern of data supports the idea that a process similar to the one suggested by the NSM forms the statistical properties.

The good performance of participants in Experiments 1 and 2, even with very small samples, raises the question of whether people could solve the inference task in more complex situations. In many real-life situations, people are not constrained to experiencing a single numerical variable. Rather, they experience and learn several variables in parallel. For example, it might be the case that the worker in the chocolate factory alternates between the two production lines, thereby gaining knowledge of both types of chocolate bars. Experiment 3 was designed to investigate the extent to which learning an additional variable would influence performance in the inference task.

EXPERIMENT 3: MULTIPLE DISTRIBUTIONS

In the first two experiments, participants learned only one distribution, and the test samples would come either from that distribution or not. Experiment 3 explores a situation where participants experience and learn values from a unimodal and bimodal distribution simultaneously and are later asked which of two test samples comes from one or the other of the two distributions. The experiment thereby investigates boundary conditions for people's ability to use the information in small samples. It is possible that presenting two variables at the same time will lead to interference where participants can no longer separate the values from the two distributions. If this is the case, performance should be worse than in Experiment 1. On the other hand, if participants are able to keep the two distributions separate, they might, for example, benefit in the sample task from being able to compare both test samples to both distributions. Experiment 3 manipulated focal distribution (unimodal/bimodal) and sample size (5/10) within subjects.

Method

Participants

Participants were 18 undergraduate students (7 male and 11 female) from Uppsala University ($M = 22.1$ years, $SD = 2.2$). They received a movie voucher or course credits as compensation for participating in the study.

Materials and procedure

Experiment 3 used the same procedure and stimulus materials as those in Experiment 1 (see Table 1). However, on each trial during the learning phase participants were presented with two consumer ratings, rather than one, labelled with separate product names (alpha/beta). The consumer ratings for one of the products followed a unimodal distribution while the ratings for the other followed a bimodal distribution. The label associated with each distribution was counterbalanced over participants. With some minor alteration, described below, participants carried out the same tasks as those in Experiment 1 during the test phase.

Sample identification. On each of the 40 trials in the sample identification task participants were presented with two test samples. Values for one of the test samples were drawn from the unimodal distribution, and values for the other test sample were drawn from the bimodal distribution. On half of the 40 trials both samples contained five values, and on the other half both contained 10 values. Further, on half of the trials the participants' task was to identify which of the two test samples was drawn from the unimodal distribution and on the other half which of the two samples was drawn from the bimodal distribution. Thus, half of the trials had the unimodal distribution as the focal distribution while the other half had the bimodal distribution as focal distribution. The two distributions seen during learning were labelled with the fictitious product names alpha and beta. Accordingly, focal distribution was manipulated by on half of the trials asking participants to indicate which of the two test samples was representative of the alpha product and on the other half

asking which was representative of the beta product.

Production task. In the production task, participants produced frequency distributions for the unimodal and bimodal distribution separately.

Results

The effects of sample size and focal distribution on sample identification performance were evaluated by entering proportion correct as the dependent measure into a dependent 2 × 2 analysis of variance (ANOVA) with sample size [small(5)/large(10)] and focal distribution (unimodal/bimodal) as within-subjects independent variables. Neither the main effect of sample size [$F(1, 17) = 0.37$, $MSE = .03$, $p = .55$; small: $M = .83$, $SD = .26$; large: $M = .85$, $SD = .22$] nor the main effect of focal distribution reached significance [$F(1, 17) = 0.50$, $MSE = .04$, $p = .49$; unimodal: $M = .82$, $SD = .27$; bimodal: $M = .86$, $SD = .21$]. Further, the sample size by distribution interaction was not significant ($F < 1$). Thus, when learning two distributions there were no effects of sample size or of focal distribution.

Knowledge of distribution properties

The two previous experiments revealed effects of distribution on some of the measures of knowledge of distribution properties. In Experiment 3, participants made separate estimates of the four properties for both of the experienced distributions. Estimates of the four properties—mean, median, MAD, and MAE—were compared between the two experienced distributions. None of the four comparisons revealed a significant difference (all ps $> .27$). Thus, when learning two distributions simultaneously, participants are able to give equally accurate estimates for properties of both variables.

Influence of individual differences

The influence of individual differences on performance in the sample identification task was examined by two separate Pearson correlations. Neither the numeracy–sample nor the LTM–sample correlation reached significance (both ps $> .18$).

One versus two variables—Samples

Because Experiment 3 used the same stimulus material and procedure as those in Experiment 1, it is possible to compare performance in the two experiments. The effect of condition (unimodal/bimodal/mixed) was investigated by means of a Kruskal–Wallis nonparametric ANOVA. The analysis revealed a significant effect of condition, $H(2) = 21.39$, $p < .001$, and post hoc analysis showed that performance in the unimodal condition ($Mdn = 1.0$) was better than that in both of the other two (bimodal: $Mdn = .85$; mixed: $Mdn = .91$) conditions, which did not differ significantly from each other.

One versus two variables—Distribution properties

Because there was no difference in accuracy of estimates of distribution properties between the bimodal and unimodal experienced distribution in Experiment 3, the data were collapsed into one accuracy measure for each property. This was done by taking the mean of the accuracy of the two estimates for each property. The accuracy of estimates of distribution properties in Experiments 1 and 3 were compared by means of four separate one-way ANOVAs, one for each of the four measures of distribution knowledge, with condition (unimodal/bimodal/mixed) as independent between-subjects variable. The analyses showed that there were no significant differences between the conditions with respect to estimates of central tendency (both ps $> .28$). The effect of condition on estimates of MAD was significant, $F(2, 51) = 5.4$, $MSE = 14{,}820.2$, $p = .007$. Post hoc tests showed that the difference between the unimodal and bimodal conditions was significant, as documented in Experiment 1, but that no other pairwise comparisons reached significance. Finally, the ANOVA with MAE as the dependent measure, and follow-up post hoc tests, showed that the performance in Experiment 3 on the production task was significantly worse, $F(2, 51) = 8.6$, $MSE = 3.17$, $p < .001$, than that in both conditions of Experiment 1.

Discussion

Experiment 3 was designed to investigate whether people are able to learn two numerical variables simultaneously and whether they can accurately separate and maintain knowledge of both variables when evaluating from which of the two presented test samples come. The accuracy of explicit estimates of variability and central tendency indicated that participants learned the target variables as accurately as in Experiment 1, where only one target variable was used. Similarly, comparing performance in the sample identification task with Experiment 1 showed that participants in Experiment 3 were as accurate as those in the bimodal condition of Experiment 1. Adding an extra variable only marginally affected the ability to accurately identify the source of a sample. The results are thus a first indication that people can accurately separate and maintain knowledge of the properties of two simultaneously experienced variables and use this knowledge to infer the origin of the two presented samples.

In contrast to the two previous experiments, there was no effect of sample size. There are, at least, two possible explanations for this difference. First, if people use either a large-sample or a small-sample representation to solve the task it is possible that the somewhat lower performance in Experiment 3 may have pushed down performance to a region where there is no longer a difference between the two sample sizes. However, the sample size effect was evident in Experiment 2 where the level of performance was similar to that in Experiment 3. Another possibility is that people are able to evaluate the two test samples against both experienced distributions. Thus, deciding which of the two samples was drawn from the bimodal distribution could be done by recognizing which sample was drawn from the unimodal distribution and choosing the other one. It is possible that this type of strategy would influence the effect of sample size on performance. Such a strategy would require people to have a considerable ability to engage in metacognitive monitoring of their own decisions.

Experiment 3 further lacked an effect of distribution. However, focal distribution rather than experienced distribution was manipulated in Experiment 3 in contrast to Experiments 1 and 2. Because the formulation of the small- and large-sample accounts does not yet include possible processes for the two-distribution case, it is difficult to make strong predictions about whether an effect of focal distribution should be expected. The design of the present study did not allow for any conclusions as to which strategy participants used to solve the two-distribution task. It will, however, be an interesting question for future research to investigate whether people are capable of the metacognitive monitoring required for the task to be solved by means of the elimination strategy suggested above and the extent to which the use of such a process would influence the effects of focal distribution and sample size.

GENERAL DISCUSSION

A long line of research has explored people's ability to be intuitive statisticians. In general, this research has investigated the extent to which people are able to accurately summarize their experience of a variable using some statistic. However, in several situations we are required to go beyond summary statistics and infer which process or distribution has generated a set of data. While previous research has investigated how internally generated samples from memory are utilized in decisions (e.g., Busemeyer & Townsend, 1993; Stewart et al., 2006) and which information in samples people use (e.g., Bar-Hillel, 1979; Chesney & Obrecht, 2012; Kareev et al., 2002; Obrecht & Chesney, 2013), little attention has been given to the cognitive processes that govern inference from samples. Further, previous research has rarely addressed inference in situations with a continuous variable that people have some experience with. The present study extended previous research by addressing three main questions: First, can people solve an inference task that uses a continuous variable, allows them to experience values from the objective distribution, and allows them to experience all values in the test sample? Second, are the statistical properties of the underlying distribution and test

sample used to solve the task or are judgements based on a memory inference? Finally, does the possible use of statistical properties involve a large- or a small-sample representation? In addition, the present study extended previous research by outlining and testing predictions from three possible processes that people might engage in to solve the inference task.

In all three experiments, participants performed well in the sample identification task, indicating that they were able to correctly identify, well above chance levels, which of two test samples had been drawn from the same distribution as previously experienced values. In fact, even in a situation where the test samples contained only four values, none of which had been shown previously, as in Experiment 2, participants still performed well above chance. Thus, and with respect to the first research question, this indicates that people are well equipped to make inductive inferences from continuous variables that they have experienced. While it is inherently difficult to make inferences with a high level of certainty about distribution properties from as little as four data points (e.g., Kacelnik & Bateson, 1996) previous research has indicated that people often stop sampling long before they have any real possibility of knowing distribution properties (e.g., Hertwig & Pleskac, 2010). The results of the present study suggest that people might be equipped with strategies that allow them to infer distribution properties, at a satisfactory level of confidence, already at very small sample sizes. It is an interesting venue for future research to explore the extent to which people's accuracy and confidence vary with sample size.

The suggested processes relying on statistical properties of the underlying distribution (small-sample and large-sample account) made different predictions with respect to the effects of distribution, sample size, and new versus old values than did the process relying on a memory heuristic. Both Experiments 1 and 2 revealed strong sample size effects with a direction predicted by the small- and large-sample accounts. Further, in both experiments the effect of distribution was in the direction suggested by the two accounts relying on statistical properties. The effect was significant in Experiment 1 but only marginally significant in Experiment 2. Controlling for measures of knowledge of distribution parameters left the effect of distribution more or less unaltered, suggesting that it is not due to values in one of the distribution being learnt more accurately. Finally, in contrast to what could be expected from a memory heuristic, there was no old–new effect in Experiment 2. Taken together, these results suggests that people use the statistical properties of the underlying distribution and the presented test sample to solve the inference task.

While the two statistics-based accounts predicted similar effects of distribution, sample size, and an old–new difference, they made different predictions with respect to the ordering of the conditions. More specifically, they predicted a critical difference between the unimodal–small and bimodal–large conditions. This critical difference was significant in the direction predicted by the small-sample account. Further, because of the formulation of the SP-function (Equation 1), it is expected that the relative size of an effect of distribution and test sample size will be similar if participants use a large-sample representation. In contrast, if people are using a small-sample representation we should expect a larger effect of distribution than of test sample size. The reason for this is that with a small-sample representation the distribution will influence all four components (both MS and TS) in the SP-function while only two of the components (the TS part) will be influenced by sample size. Over the first two experiments, the relative reduction in proportion correct caused by the sample size manipulation was 6%. In contrast, the corresponding reduction caused by the distribution manipulation was 16%. These two pieces of evidence together suggest that the statistics-based process that participants use involves a small-sample representation.

Experiment 3 extended the findings from the two previous experiments by investigating whether participants would be able to solve the inference task after experiencing both the unimodal and the bimodal distributions simultaneously during learning. Comparing performance for

explicit estimates of variability, central tendency, and the production task showed that participants learned the two distributions as well as did participants in Experiment 1, who only experienced a single variable. Further, the results from the sample identification task revealed a similar level of performance to that of the bimodal condition of Experiment 1. Thus participants were able to keep the values from the two distributions separated throughout learning and access properties specific to each distribution in the test phase. Experiment 3 found no effect of sample size. This could be due to the somewhat lower overall performance in Experiment 3 pushing down performance to a region where there is no longer a difference between the two sample sizes. The similar performance of participants in Experiment 3, the bimodal condition in Experiment 1, and Experiment 2, where the sample size effect was evident, however, makes this less likely. A second possibility is that participants were able to utilize their knowledge of both underlying distributions when making inferences. For example, when deciding which of the two samples were drawn from the bimodal distribution, participants could have recognized which of the two were drawn from the unimodal distribution and choose the other one (see, Dougherty et al., 1999; Thomas et al., 2008, for a similar suggestion of a conditional memory search process). They could thus have benefited from being able to make the, according to Experiment 1, somewhat easier inference first. What process supports inference in the two-distribution case is unclear and should be an interesting question for future research to explore.

All three experiments are limited to the use of two extreme distributions. In real-life situations, people may, however, experience distributions that are quite different from those used here. If the type of distributions that they experience in everyday life influences the cognitive processes that people have developed to make inferences, the distributions used here might not fully capture the inference process. The unimodal and bimodal distributions and the respective sample sizes were, however, chosen because they predict different qualitative patterns of results with respect to the above suggested cognitive processes. Future research should investigate whether the findings here extend to situations where the presented data follow distributions similar to what could be found in real-life situations. Further, in the inference task participants were always presented with two samples. This situation allows participants to use the information both in the target sample and in the distractor sample to infer which of the two was drawn from the experienced distribution. Future research should explore this possibility by contrasting inference in the presence of a second sample (i.e., a comparison task) with inference from only one sample (i.e., a yes/no recognition task).

The present study makes two contributions to previous research. First, it shows that when people have first-hand experience with a continuous variable they apparently do not need a lot of information to infer whether a sample is drawn from an experienced distribution. Also, the results suggest that the cognitive process supporting such inferences is one where computations are made post hoc on small samples from memory retrieved at the time of judgement, a process that would be especially efficient and flexible when no prior information is given about how experienced data will be used. Thus, to summarize, when facing an inference from a small sample people seem to be efficient and flexible intuitive statisticians.

Original manuscript received 30 June 2013
Accepted revision received 23 June 2014
First published online 1 October 2014

REFERENCES

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*, 245–257.

Beach, L. R., Wise, J. A., & Barclay, S. (1970). Sample proportions and subjective probability revisions. *Organizational Behavior and Human Performance*, *5*, 183–190.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science*, *4*, 190–195.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision field-theory: A dynamic cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Chesney, D. L., & Obrecht, N. A. (2012). Statistical judgments are influenced by the implied likelihood that samples represent the same population. *Memory & Cognition*, *40*, 420–433.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*, 263–282.

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

Evans, J. S. B. T., & Pollard, P. (1985). Intuitive statistical inferences about normally distributed data. *Acta Psychologica*, *60*, 57–71.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*, 659–676.

Fiedler, K. & Juslin, P. (Eds.). (2006). *Information sampling and adaptive cognition*. Cambridge, UK: Cambridge University Press.

Fox, S., & Thornton, G. C. (1993). Implicit distribution theory: The influence of cognitive representation of differentiation on actual ratings. *Perceptual and Motor Skills*, *76*, 259–276.

Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 966–982.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science*, *23*, 1515–1523.

Gilovich, T., Griffin, D. W., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Krieger Publishing Company.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*, 725–743.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*, 9066–9071.

Hansson, P., Rönnlund, M., Juslin, P., & Nilsson, L.-G. (2008). Adult age differences in the realism of confidence judgments: Overconfidence, format dependence, and cognitive predictors. *Psychology and Aging*, *23*, 531–544.

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

Hertwig, R. & Pleskac, T. J. (2010). Decisions from experience: Why small samples?. *Cognition*, *115*, 225–237.

Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Jako, R. A. & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology*, *75*, 500–505.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 259–298.

Juslin, P., Nilsson, H., Winman, A., & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition*, *120*, 248–267.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*, 678–703.

Kacelnik, A., & Bateson, M. (1996). Risky theories—the effects of variance on foraging decisions. *American Zoologist*, *36*, 402–434.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136–153.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

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

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*, 287–297.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 551–578.

Lindskog, M., Winman, A., & Juslin, P. (2013a). Calculate or wait: Is man an eager or a lazy intuitive statistician?. *Journal of Cognitive Psychology*, *25*, 994–1014.

Lindskog, M., Winman, A., & Juslin, P. (2013b). Naïve point estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 782–800.

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*, 165–188.

Lipkus, I. M. & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. *Health Education & Behavior*, *36*, 1065–1081.

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

Lovie, P. (1978). Teaching intuitive statistics II. aiding the estimation of standard deviations. *International Journal of Mathematical Education in Science and Technology*, *9*, 213–219.

Lovie, P. & Lovie, A. D. (1976). Teaching intuitive statistics i: Estimating means and variances. *International Journal of Mathematical Education in Science and Technology*, *7*, 29–39.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*, 517–534.

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.

Nisbett, R. E., & Kunda, Z. (1985). Perception of social distributions. *Journal of Personality and Social Psychology*, *48*, 297–311.

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. & Chesney, D. L. (2013). Sample representativeness affects whether judgments are influenced by base rate or sample size. *Acta Psychologica*, *142*, 370–382.

Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making*, *3*, 619–635.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.

Phillips, L. D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychology*, *3*, 5–18.

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

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

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1–26.

Thomas, R. P., Dougherty, M. R. P., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Vul, W., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 66–72). Austin, TX: Cognitive Science Society.

Xu, F. & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*, 97–104.

Xu, F. & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*, 5012–5015.

Zacks, R. T. & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *ETC. frequency processing and cognition* (pp. 21–36). New York, NY, US: Oxford University Press.

APPENDIX

Predicted performance by the large-sample and small-sample accounts

This appendix outlines the computer simulations used to derive the predicted performance of the large-sample and small-sample accounts. The outcomes of three separate simulations are illustrated in Figure 1.

Similar to the experiments, the simulation creates two separate sets of values that are considered experienced distributions (EDs) by random sampling 100 values from two separate beta distributions, beta(α, β), with parameters α and β. The parameters α and β are chosen so that one distribution is bimodal, and the other is unimodal. Figure 1 shows the results of three simulations where α and β are chosen to be closer to [beta(3, 3) and beta(.7, .7)], further from [beta(8, 8) and beta (.2, .2)], or equally far from [beta(5, 5) and beta(.5, .5)] the uniform distribution as the distributions used in the experiments.

The simulation then simulates 2000 agents that first acquire knowledge about one of the EDs and then conducts the sample task described in the experiments. Learning is simulated by allowing each agent to retain a random subset of the 100 values in the respective ED (i.e., a subjective distribution; SUD). Half of the simulated agents receive experience with the unimodal and half with the bimodal ED. The size of the SUD is governed by a memory parameter (γ). In the current simulations $\gamma = .6$, which is motivated by the performance seen by participants in similar experiments (see e.g., Lindskog et al., 2013a, 2013b).

In the simulated sample task each agent makes 100 choices between two test samples (TS_1 and TS_2). For each choice, TS_1 and TS_2 are drawn randomly from the two EDs. That is, both test samples are old in the sense that they contain values that could possibly be in the SUD. The choice between TS_1 and TS_2 is made by means of the *SP*-function (Equation 1). Accordingly, the agent concludes that TS_1 rather than TS_2 is drawn from the OD if $\mathrm{SP}(C, T_1) - \mathrm{SP}(C, T_2) > 0$ and the opposite if $\mathrm{SP}(C, T_1) - \mathrm{SP}(C, T_2) < 0$. The difference between the large-sample account and the small-sample account is found in how C is defined. For the large-sample account, C is defined as the entire SUD. In the small-sample case, C is defined by a random sample of six observations drawn from the SUD for each choice.

The *SP*-function includes a parameter (θ) that determines the relative weight between the difference of means and the differences of the standard deviations. Because there is no a priori reason to expect that people would prefer one to the other, the simulations illustrated in Figure 1 used $\theta = .5$. It should be noted, however, that for the distributions used here the same qualitative pattern of predicted performance is found for $\theta \leq .5$. When θ becomes larger than .5, performance rapidly begins to decrease with the current distributions because both ODs have the same mean.