Is the Intuitive Statistician Eager or Lazy?

Exploring the Cognitive Processes of Intuitive Statistical Judgments

MARCUS LINDSKOG
Numerical information is ubiquitous and people are continuously engaged in evaluating it by means of intuitive statistical judgments. Much research has evaluated if people’s judgments live up to the norms of statistical theory but directed far less attention to the cognitive processes that underlie the judgments.

The present thesis outlines, compares, and tests two cognitive models for intuitive statistical judgments, summarized in the metaphors of the lazy and eager intuitive statistician. In short, the lazy statistician postpones judgments to the time of a query when the properties of a small sample of values retrieved from memory serve as proxies for population properties. In contrast, the eager statistician abstracts summary representations of population properties online from incoming data.

Four empirical studies were conducted. Study I outlined the two models and investigated whether an eager or a lazy statistician best describes how people make intuitive statistical judgments. In general the results supported the notion that people spontaneously engage in a lazy process. Under certain specific conditions, however, participants were able to induce abstract representations of the experienced data. Study II and Study III extended the models to describe naive point estimates (Study II) and inference about a generating distribution (Study III). The results indicated that both the former and the latter type of judgment was better described by a lazy than an eager model. Finally, Study IV, building on the support in Studies I-III, investigated boundary conditions for a lazy model by exploring if statistical judgments are influenced by common memory effects (primacy and recency). The results indicated no such effects, suggesting that the sampling from long-term memory in a lazy process is not conditional on when the data is encountered.

The present thesis makes two major contributions. First, the lazy and eager models are first attempts at outlining a process model that could possibly be applied for a large variety of statistical judgments. Second, because a lazy process imposes boundary conditions on the accuracy of statistical judgments, the results suggest that the limitations of a lazy intuitive statistician would need to be taken into consideration in a variety of situations.

Keywords: Lazy intuitive statistician, Eager intuitive statistician, Intuitive statistics, Sampling model, Numerical cognition

© Marcus Lindskog 2013

ISSN 1652-9030
ISBN 978-91-554-8827-7
urn:nbn:se:uu:diva-211505 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-211505)
To Nils-Eric and Alfred
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.
Contents

Introduction ................................................................................................... 11

Background ................................................................................................... 15
  Intuitive Statistical Judgments .................................................................. 15
    Estimates of Descriptive Properties ..................................................... 15
    Intuitive Statistical Inference ............................................................... 20
    From an Intuitive to a Naïve Intuitive Statistician ............................... 23
    Bayesian Models of Intuitive Judgments ............................................. 26

Distinctions in Cognitive Science ............................................................ 28

An Eager vs. a Lazy Intuitive Statistician ................................................... 29
  The Eager Statistician .......................................................................... 30
  The Lazy Statistician ........................................................................... 31
  Eager or Lazy? ..................................................................................... 32
  Not Only Intuitive Statistics ................................................................ 33
  A Lazy Model for Naïve Point Estimation .......................................... 34

Empirical Work ............................................................................................. 38

Study I: Calculate or Wait: Is Man an Eager or a Lazy Intuitive Statistician? ......................................................................................... 38
  Predictions ........................................................................................... 39
  Method ................................................................................................. 40
  Results ................................................................................................. 41
  Discussion ............................................................................................ 44

Study II: Naïve Point Estimation .............................................................. 45
  Predictions ........................................................................................... 46
  Method ................................................................................................. 46
  Results ................................................................................................. 47
  Discussion ............................................................................................ 50

Study III: Where Did That Come From? – Identifying the Source of a Sample ................................................................. 51
  Predictions ........................................................................................... 53
  Method ................................................................................................. 55
  Results ................................................................................................. 56
  Discussion ............................................................................................ 58

Study IV: Are all Data Created Equal? - Exploring Boundary Conditions for a Lazy Intuitive Statistician ................................. 61
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>61</td>
</tr>
<tr>
<td>Results</td>
<td>63</td>
</tr>
<tr>
<td>Discussion</td>
<td>65</td>
</tr>
<tr>
<td>General Discussion</td>
<td>67</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>75</td>
</tr>
<tr>
<td>References</td>
<td>77</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>DR</td>
<td>Direct Retrieval</td>
</tr>
<tr>
<td>LTM</td>
<td>Long-term Memory</td>
</tr>
<tr>
<td>LTMC</td>
<td>Long-term Memory Constrained</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MR</td>
<td>Mean Absolute Ratio</td>
</tr>
<tr>
<td>NPE</td>
<td>Naïve Point Estimation</td>
</tr>
<tr>
<td>NSM</td>
<td>Naïve Sampling Model</td>
</tr>
<tr>
<td>OED</td>
<td>Objective Environmental Distribution</td>
</tr>
<tr>
<td>RFH</td>
<td>Recognition Failure Heuristic</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Squared Deviation</td>
</tr>
<tr>
<td>SED</td>
<td>Subjective Environmental Distribution</td>
</tr>
<tr>
<td>SI</td>
<td>Shape Index</td>
</tr>
<tr>
<td>SSD</td>
<td>Subjective Sample Distribution</td>
</tr>
<tr>
<td>STM</td>
<td>Short-term Memory</td>
</tr>
<tr>
<td>STMC</td>
<td>Short-term Memory Constrained</td>
</tr>
</tbody>
</table>
Introduction

In everyday life people often encounter numerical information. They learn about the prices of groceries at their local supermarket, compare salaries with their co-workers or experience the time it takes to reach the checkout counter. Often, these sources of information also provide the basis for judgments and decisions. You might, for example, estimate the average salary at your work place or how long it will take before you reach the end of a queue based on previous experiences. The formal mathematics needed to perform the explicit calculations for such judgments is often difficult, and sometimes even intractable, without computers and statistical training. People are, however, by and large accurate when making intuitive judgments, that is, when making judgments without explicit calculations, the aid of computational devices, or any statistical training.

The psychological study of intuitive statistical judgments investigates how people make these kinds of judgments and has focused on three main areas. First, it has approached the normative aspect of intuitive statistical judgments by investigating the extent to which people’s judgments coincide with what could be expected from statistical theory (e.g., Peterson & Beach, 1967; Pollard, 1984). For example, when people are asked to estimate the mean salary at their work place, how close are their estimates to the actual mean? Peterson and Beach (1967, p. 29) described this research strategy as examining the correspondence between observed intuitive judgments and judgments of a “statistical man”, a norm akin to the “ideal observer” in signal detection theory.

The second approach has investigated judgments from a descriptive viewpoint by looking at the processes leading up to a statistical judgment (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999; Juslin, Winman, & Hansson, 2007; Kahneman & Tversky, 1982; Levin, 1974a, 1974b, 1975; Peterson & Miller, 1964; Tversky & Kahneman, 1974). That is, what strategies and algorithms do people use to reach their estimate of the mean salary? The third approach has investigated how general cognitive constraints interact with properties in the environment to form statistical judgments (e.g., Fiedler & Kareev, 2006; Fiedler, 2000; Juslin, Nilsson, Winman, & Lindskog, 2011; Juslin et al., 2007; Kareev, Lieberman, & Lev, 1997; Nilsson, Winman, Juslin, & Hansson, 2009). For example, if I can only remember the salary of four of my co-workers, how will that affect my estimate of the mean salary at my work place?
To what degree can people be considered intuitive statisticians? The general conclusions from previous research can be summarized as follows. While some statistical judgments are remarkably accurate (e.g., Griffiths & Tenenbaum, 2006; Peterson & Beach, 1967) others are consistently and systematically biased (e.g., Fiedler, 2000; Kareev, Arnon, & Horwitz-Zeligier, 2002; Pollard, 1984). Further, statistical judgments are constrained by cognitive limitations such as people having limited time, knowledge, and computational ability\(^1\) (i.e., are boundedly rational, Simon, 1990), working memory capacity (Dougherty & Hunter, 2003; Gaissmaier, Schooler, & Rieskamp, 2006; Hansson, Rönnlund, Juslin, & Nilsson, 2008; Kareev et al., 2002; Stewart, Chater, & Brown, 2006), and linear additive integration of information (Anderson, 1991; Hogarth & Einhorn, 1992; Juslin, Karlsson, & Olsson, 2008; Juslin et al., 2011; Nilsson et al., 2009). It is possible that people could adjust for the impact of cognitive limitations, at least to some extent, if they were aware of their existence. However, a third important finding in the literature on intuitive statistical judgments is that people suffer from **cognitive myopia** (e.g., Fiedler, 2000, 2008). That is, even though they may be accurate with respect to the given stimulus, they are naïve with respect to how the data have been created and thereby to any biases inherent in the data (Juslin et al., 2007; Kareev et al., 2002).

Although previous research has outlined descriptive process models for some specific judgments (e.g., Goldstein & Gigerenzer, 2002; Juslin et al., 2007; Tversky & Kahneman, 1974) there have been few attempts at outlining a more general process model for statistical judgments. Further, the models that have been suggested have mostly emphasized the processes operating on the data at the time of judgment (however, see Juslin et al., 2007), while much less attention has been given to \(i\) how experienced data are stored prior to a judgment and \(ii\) how data are retrieved from memory at the time of a judgment. The present thesis outlines and tests two models for statistical judgments that both address these issues. The first, summarized in the metaphor of the **Lazy Intuitive Statistician**, is a generalization and extension of the **Naïve Sampling Model** (NSM, Juslin et al., 2007) and proposes that data is stored as raw data points during exposure and that no operations are made before the time of a judgment. When a person is asked for an estimate of the experienced data the lazy model suggests that a small sample of stored data points are retrieved from memory. The statistical properties of this sample are then used as proxies for the population properties. The lazy intuitive statistician is contrasted against a model summarized in the metaphor of the **Eager Intuitive Statistician**. In short, the eager model proposes that people extract and store abstract summary representations from experienced data

---

\(^1\) Computational ability should here be conceived as the built in ability of the mind to perform intuitive calculations on experienced data rather than a mathematical ability acquired through formal schooling.
during exposure and that it is the summary representations that are retrieved from memory and used at the time of judgment.

The models draw inspiration from several classical distinctions in cognitive science and machine learning with respect to what algorithms operate on the data (lazy vs. eager; Aha, 1997; Juslin & Persson, 2002), when operations are made on the data (online vs. retrospective; Dougherty, Franco-Watkins, & Thomas, 2008; Hastie & Park, 1986), and how the data is stored in memory (episodic vs. semantic memory; Battaglia & Pennartz, 2011; Squire & Zola, 1996; Tulving, 2002). In addition the models are generalizations and extensions from models proposed within the framework of the naïve intuitive statistician (Fiedler & Juslin, 2006a; Fiedler, 2000; Juslin et al., 2007).

The interest in formulating the models described above is further highlighted by the interest in Bayesian accounts of human cognition (Oaksford & Chater, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). These accounts often assume that cognitive processes are adaptations to distributions in the environment and that people update their beliefs in the light of new data based on knowledge of prior distributions. Although it is required from a Bayesian agent to be able to both represent distributions of variables in the environment and to perform complex statistical computations, very little research has addressed how this might come about (but see, Shi, Griffiths, Feldman, & Sanborn, 2010; Vul, Goodman, Griffiths, & Tenenbaum, 2009). The distinction between the eager and lazy intuitive statistician may therefore serve as a first indication of how a Bayesian agent might be implemented in the mind.

The remainder of the thesis is organized as follows. The Background section first reviews the research on intuitive statistical judgments including the development of the naïve intuitive statistician and Bayesian models of cognition. The second part of the background discusses distinctions within cognitive science that are important for the two possible models of intuitive statistical judgments. The models are outlined in the third part of the background. Finally, the last part of the background illustrates how the models can be applied to a novel statistical judgment task when a model for naïve point estimation is derived.

In the Empirical Work section four studies will be presented that investigate the general question of whether intuitive statistical judgments are best described by the eager or lazy intuitive statistician. Study I investigates whether an eager or a lazy statistician best describe how people make intuitive statistical judgments. Study II and Study III test the two models for novel types of statistical judgment. Foreshadowing the results, the three first studies support the idea that people in general adopt a lazy cognitive algorithm for statistical judgments. Study IV, therefore, explores some of the assumptions and boundary conditions embedded in a lazy model. Finally, in
the General Discussion section the results of the four studies are discussed in
the light of previous research.
Background

Intuitive Statistical Judgments

Since the 1950s researchers in the field of cognitive psychology have investigated to what degree people live up to the norms of statistical theory (e.g., Beach & Swensson, 1966; Peterson & Beach, 1967) when dealing with artificially created or naturally occurring random variables. This research has addressed both the extent to which outcomes (i.e., estimates) and algorithms (i.e., processes) are normative.

Similar to the mathematical discipline concerned with the investigation of random variables, the psychological literature addressing people as intuitive statisticians has mainly focused on two sub-areas. The first, similar to the sub-discipline of probability theory, has investigated people’s notion of probabilities (Tversky & Kahneman, 1974) and to what extent normative rules such as Bayes’ theorem (Britney & Winkler, 1974; Edwards, Lindman, & Savage, 1963; Koehler, 1996; Tversky & Kahneman, 1974; Winkler, 1970) and the conjunction rule (Juslin, Nilsson, & Winman, 2009; Nilsson et al., 2009; Tversky & Kahneman, 1974) are adhered to. The second, similar to the sub-discipline of statistical theory, has emphasized the investigation of estimates of statistical properties of variables (e.g., Pollard, 1984), estimates of relational properties between variables (e.g., covariation; Allan & Tangen, 2005; Beach & Scopp, 1968; Cheng, 1997; Griffiths & Tenenbaum, 2005; Malmi, 1986), and the intuitive use of statistical inference (e.g., Slovic, Fischhoff, & Lichtenstein, 1977; Tversky & Kahneman, 1974). Even though all of these areas are interesting and relevant when evaluating the notion of people as intuitive statisticians it is a large undertaking to investigate them all in only a few studies. This thesis will therefore focus on and be limited to areas that are related to statistical theory. More specifically, it will focus on the sub-areas of estimates of statistical properties and intuitive statistical inference.

Estimates of Descriptive Properties

Consider having a list with the height, in meters, of all the children in a school class. In isolation these values tell you very little. The most basic operation of any statistician when encountering a set of values, like the height of children in a school class, is to calculate descriptive statistics.
There are several descriptive parameters that could be calculated on a set of values. Some, however, are more important than others in everyday life. For example, knowing the *mean*, or average, price of a cup of coffee makes it possible to evaluate whether a cup of coffee at your favorite café is expensive or not. Knowing the *kurtosis* of the price distribution might, however, be less informative. Research in the area of intuitive estimates of descriptive properties has mainly studied three basic and important properties; *central tendency*, *variability* and *higher order properties* (e.g., kurtosis, shape, and skew).

**Central tendency**

The central tendency of a variable is a basic property that has received a lot of attention in the cognitive literature (e.g., Levin, 1974a, 1974b; Peterson & Beach, 1967; Peterson & Miller, 1964; Spencer, 1961, 1963; Wolfe, 1975). The research has focused on four main questions. First, there are several normative measures that can be used to describe the central tendency of a variable (e.g., mean, median, mid-range etc.). One line of research has investigated which of these that people’s estimates coincide with. When estimating data from only one variable, people’s estimates seem to coincide with the mean (Malmi & Samson, 1983). When estimating a population mean based on two samples the evidence suggest that people use a weighted average (Levin, 1974a), but with the weights not always sufficiently adjusted for sample size (Pollard, 1984). Further, researchers have investigated the extent to which people’s estimates for a given type of measure coincide with the normative central tendency. Pollard (1984) reviewed almost twenty years of research on this question and concluded that when people are exposed to numeric variables on a trial-by-trial basis and are asked to estimate central tendency, the estimates generally coincide decently with what could be expected from statistical theory. A third line of research has investigated the extent to which estimates of central tendency are sensitive to higher order moments. That is, whether properties such as variance, skewness, kurtosis, etc. influence estimates. The main findings have been that the accuracy of estimates of central tendency decreases as variability increases (Beach & Swensson, 1966; Wolfe, 1975) and that even though the number of presented values and the shape of the distribution both influence accuracy (Malmi & Samson, 1983; Smith & Price, 2010; Wolfe, 1975) they do so to a small extent. Smith and Price (2010), for example, showed that estimates of central tendency (mean) increased as a function of the number of values that were to be included in the estimate. There is also evidence suggesting that people can monitor factors (e.g., variability) that influence the difficulty of estimation (Laestadius, 1970). For example, Laestadius (1970) had participants specify an interval of acceptable error for intuitive estimates of the mean for a set of values with either high or low variance, a factor that has been shown (Beach & Swensson, 1966; Spencer, 1961, 1963; Wolfe, 1975) to influence
the accuracy of estimates. The participants reported wider intervals of acceptable error for the high variance set than for the low variance set, indicating that they were responsive to task features that influence the accuracy of estimates. Finally, researchers have investigated whether people can, post hoc, estimate the central tendency of a subset of the experienced variable. Malmi and Samson (1983), for example, had participants observe 48 values from two numerical variables. After exposure participants were asked to either report the central tendency of all presented values or to disregard a subset of the values and to report the central tendency of the remaining subset. They found that participants’ estimates were accurate both when reporting the central tendency for all values and when reporting it for the subset. These results suggest that people not only monitor the central tendency as they are presented with values, they also have access to the values at the time of judgment.

**Variability**

Another important statistical property is variability. Variability is a measure of the extent to which each value in a set deviates from the central tendency of that set. As with central tendency, there are several different ways in which variability can be expressed (e.g., variance, standard deviation, mean absolute deviation, coefficient of variation etc.). Research within this area has addressed whether or not people respond to variability (i.e., whether judgments and decisions are influenced by the level of variability) and to what degree estimates of variability are correct. Whether people are sensitive to and respond to variability, often expressed as risk, is, in fact, at the heart of a large literature concerned with people’s risk attitudes and risk preferences (see e.g., Fox & Tannenbaum, 2011; Kahneman, Slovic, & Tversky, 1982; March, 1996). The general conclusion here has been that people are indeed sensitive to and respond to variability (Kareev et al., 2002; Weber, Shafir, & Blais, 2004; Weber, 2010). With respect to the accuracy of estimates, previous research has indicated that estimates of variability are often inaccurate (Kareev et al., 2002; Lovie & Lovie, 1976; Lovie, 1978; Pollard, 1984). More specifically, this body of research suggests that people often underestimate variance. Kareev et al. (2002), for example, had people observe paper cylinders colored up to different heights. Participants underestimated the variance of the heights both when comparing the variance of an observed population to two comparison populations and when choosing a population for a lottery. Thus, even though people seem to be sensitive to variability, they underestimate it. With only a few exceptions (Lovie & Lovie, 1976; Lovie, 1978) participants in the studies reported above never give explicit numerical estimates of variance but rather estimate it implicitly. For example, in several tasks the participants in the Kareev et al. (2002) study made comparative judgments between sets based on variance. In one task participants were first presented with an original set of paper cylinders col-
ored up to different heights. They were then presented with two comparison sets and were to decide which of the two sets that best resembled the original set. While one comparison set was identical, item for item, to the original set the other comparison set had the same mean but unequal variance. Another of the tasks in Kareev et al. (2002) had participants chose one of two sets for a subsequent draw in a lottery, including a reward, of two items. The reward of the draw was dependent on the similarity of the two drawn items. Thus, in both of these tasks participants are expected to base their judgment on the variability of the experienced data, but in neither are any explicit estimates of variability elicited. Further, the stimuli used are almost exclusively perceptual. The main reason for both the use of perceptual stimuli and implicit estimates of variance is that variance is not scale invariant. That is, it is not straightforward to compare estimates of variance for height with estimates of variance of weight. In fact, even just changing the unit of interest from, for example, meters to centimeters scales the variance in a non-intuitive way. Weber and colleagues (2004; 2010) have argued that this scale invariance suggests that people are likely to use a measure of variability that is more akin to the coefficient of variation (the variance scaled against the mean) than to actual variance. It should be noted that within the study of estimates of variability the influence of higher order properties have received little attention. That is, few studies have investigated the extent to which properties such as the shape of the distribution influence estimates of variability.

Higher order properties

Central tendency and variability are sometimes referred to as the first and second moment of a random variable, respectively. However, even though central tendency and variability are important when describing a variable they are in no way exhaustive. Often, it is also possible to calculate higher order moments (e.g., skewness, kurtosis, shape), or properties, of a variable that also add information. The two variables in Figure 1, illustrated by their probability density functions, for example, share the same mean and median but obviously have very different properties. The most salient difference is the shape of the distributions (although the variance also differs highly), a higher order property that is not sufficiently described by descriptive statistics such as central tendency and variance. As mentioned previously, there are several higher order properties that may be used to describe a variable (e.g., skewness, kurtosis etc.). This thesis will focus on one of these properties, the shape of the distribution.

If it is possible to determine which class of distribution (e.g., normal, log-normal, exponential etc.) that best describes a variable it will also be possible to parameterize it. That is, it will be possible to fully describe the variable with only a few pieces of information called parameters. For example, under the assumption that a variable is normally distributed it would be sufficient to know the mean and variance in order to fully determine its proper-
ties. However, making assumptions about what distribution that has generated a variable is not always straightforward from an intuitive statistical viewpoint. Because the variable illustrated in Figure 1A shares some features with the normal distribution, often covered in introductory statistics courses, you might assume, for example, that it is normally distributed. However, there are several other possible distributions, which people in general might not have heard of, that have similar features. In fact, the variables depicted in Figure 1 are both from a beta-distribution, where the psychologically not so accessible $\alpha$- and $\beta$-parameters determine the distribution, illustrating that making an assumption of what distribution that generates a variable is not always easy.

Figure 1. Illustration of a Unimodal (Panel A) and Bimodal (Panel B) distribution (created with beta distributions).

The research addressing if people have knowledge of statistical properties beyond descriptive estimator statistics has mainly addressed if people can estimate the distribution shape of variables they have encountered in their everyday lives (Fox & Thornton, 1993; Galesic, Olsson, & Rieskamp, 2012; Griffiths & Tenenbaum, 2006; Jako & Murphy, 1990; Linville, Fischer, & Salovey, 1989; Nisbett, Krantz, Jepson, & Kunda, 1983; Nisbett & Kunda, 1985). In addition, some studies have used laboratory settings with variables being experienced on a trial-by-trial basis (Griffiths & Tenenbaum, 2011). Nisbett et al. (1983), for example, asked students to estimate the distribution of grade point averages among their peers and Griffiths and Tenenbaum (2006) were concerned partly with distributions of baking times of pastries and movie runtimes. The results are, at best, mixed. In some cases, people’s knowledge of the distributions appear remarkably accurate (Griffiths & Tenenbaum, 2006, 2011; see also, Mozer, Pashler, & Homaei, 2008) in others, judgments seem to be biased by the external information in the environ-
ment (Galesic et al., 2012; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978) and/or where people find themselves in the distribution (Fiedler, 2000; Nisbett & Kunda, 1985).

Whether people have accurate access to higher order distribution properties is of interest because in several research areas it is assumed that people are influenced by knowledge of distributional shape. For example, in economic theory, an expert forecaster is assumed to use the central tendency of his or her subjective probability distribution as point prediction for a variable (e.g., Engelberg, Manski, & Williams, 2009), under the assumption that the distribution is normal. Some research also suggests that people enter category-learning tasks with expectations of normal distributions of exemplars (Flannagan, Fried, & Holyoak, 1986) and models such as the category density model (Fried & Holyoak, 1984) assume that people expect unimodal distributions. Further, Bayesian models of human cognition often assume that people update their beliefs in the light of new data based on knowledge of prior distributions (Chater, Tenenbaum, & Yuille, 2006; Tenenbaum et al., 2011). It is reasonable to assume that the properties of these distributions need to somehow be specified. Some studies, however, suggest that people store the information in a “raw” form, more like non-parametric frequency counts in “mental histograms” (Malmi & Samson, 1983), rather than as distributions.

The research reviewed above has been enlightening but leaves a number of questions unanswered. It does not address the question of how the properties of a numerical variable are represented. That is, how the numerical information is stored in memory. Further, the research does not investigate any boundary conditions that might influence such a representation. For example, if some class of distributions is easier to learn than others. Finally, few of the available studies have been conducted as controlled experiments, making it difficult to determine how veridical the knowledge of the participant is with respect to the information that he or she has actually experienced. For example, in the studies by Nisbett et al. (1983) and Griffiths and Tenenbaum (2006) there is no control over what instances of the variables participants actually have encountered. Thus, it is difficult to know whether people extract a notion of the distributional shape by experience of instances of the variable alone, or whether their idea of the distributional shape is influenced by other top-down sources of knowledge. It might be, for example, that students having taken an introductory course in statistics have learned that grade point averages are normally distributed around some mean value.

Intuitive Statistical Inference

Consider being a school nurse who has reviewed the height of 200 children at a primary school. Now you encounter a new child, the height of which you do not know, from the same school. What would be your best guess of
the height of the child? Predicting the height of the child is an example of statistical inference. Based on previous knowledge you infer the unknown state of a variable. In statistics, as well as in everyday life, there are several different situations where people are required to make statistical inferences. For example, estimating the remaining waiting time in the supermarket cash queue, given that you’ve waited five minutes, or estimating the height of the next child that steps into your office are both examples of statistical inference. There exist an extensive body of research investigating several different types of intuitive statistical inferences (Evans & Pollard, 1985; Griffiths & Tenenbaum, 2006; Kahneman & Tversky, 1982; Nisbett et al., 1983). This thesis will focus on two types of statistical inference in particular, namely point estimates and inference of a generating distribution.

Intuitive point estimation
Point estimation involves using data from a sample to calculate a best guess of an unknown parameter or an unknown observation of a variable. People regularly perform such operations on everyday data. For example the school nurse presented above would have to make a point estimate of the most probable height of the next child based on an observed sample of children. Even though there appears to be no single normative principle that can guide point estimates (Lehmann & Casella, 1998), researchers in psychology have investigated how people make point estimates of unknown quantities. In their metrics and mapping framework, Brown and Siegler (1993) emphasized the importance of knowledge both of the metric properties of the quantity (e.g., its mean, variance, and distribution) and the mapping properties of the quantity, the latter involving knowledge about the ordinal relations within the domain, based on domain-specific knowledge and heuristics. While mapping knowledge has been emphasized in research on heuristics (Gilovich, Griffin, & Kahnean, 2002), multiple-cue judgment (Juslin et al., 2008; von Helversen & Rieskamp, 2008), and function learning (DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004), in models of how people use cues to predict a criterion, the effects of metric knowledge on point estimates have largely been ignored (but see, Pitz, Leung, Hamilos, & Terpening, 1976). In the literature on forecasting (e.g., Goodwin, 1996; Lawrence, Goodwin, O’Connor, & Önkal, 2006) and in economic theory (e.g., Engelberg et al., 2009) people are often assumed to use the central tendency of a subjective probability distribution as their point estimate, where this distribution manifests both their metric and their mapping knowledge. However, it is still unclear how metric knowledge, that is statistical properties, influences point estimates and as of yet there is no formal model of how metric knowledge is used to form point estimates. One goal of this thesis was to outline such a model. This is done in Study II.
Intuitive inference of a generating distribution

Another take on inference from samples is found in tasks that require people to infer the distribution or process, which has generated the sample. The generating distribution is thus the underlying distribution from which a sequence of random numbers has been drawn (cf. Brown & Steyvers, 2009). This situation would be similar to hypothesis testing found in classical statistics or the estimation of a likelihood as found in Bayesian statistics.

There are several accounts of human cognition that assume an internal sampling of information from memory prior to making a judgment or decision (Busemeyer & Townsend, 1993; Dougherty & Hunter, 2003; Fiedler & Justlin, 2006a; Fiedler, 2000; Gaissmaier et al., 2006; Hansson et al., 2008; Justlin et al., 2007; Kahneman & Miller, 1986; Kareev et al., 2002; Vul et al., 2009). That people are prone to use internally generated samples as a basis for judgments and decisions suggests an ability to use the information contained in samples appropriately when making inferences of the generating distribution. 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, even taking 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, 2011; Evans & Dusoir, 1977; Obrecht, Chapman, & Gelman, 2007; Obrecht & Chesney, 2013; Sedlmeier & Gigerenzer, 1997; Sedlmeier, 1998) 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 & Justlin, 2006a; Fiedler, 2000; Kareev et al., 2002). Failing to appropriately evaluate the representativeness of a sample, for example, can lead to a number of apparent judgment biases (Fiedler & Justlin, 2006a; Fiedler, 2000; Kareev et al., 2002). While previous research has been informative with respect to how people utilize sample information it has to a large extent ignored which processes that govern how the scarce information in a small sample is used to make an inference about a generating distribution. This question is addressed in Study III of the present thesis.

Common features of inference tasks

Studies investigating intuitive inductive inference often share three task features that may influence what conclusions that can be drawn (see e.g., Beach, Wise, & Barclay, 1970; Griffin & Tversky, 1992; Phillips & Ed-
wards, 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, several data sets that people experience in their everyday life come from continuous distributions. Second, participants are generally withheld firsthand experience with the underlying distribution. 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 expectation of the underlying distribution before the sample is presented. Finally, the sample is often described to participants rather than experienced firsthand. That is, participants might receive the information that a sample contains four red and three white chips in a written summary statement rather than experiencing each chip in the sample. Information in everyday life is, however, seldom experienced in descriptive summaries of data. Further, 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 Study III the features of the task are chosen to address the possible limitations of previous research.

From an Intuitive to a Naïve Intuitive Statistician

The research on statistical judgments has been strongly influenced by three metaphors of the human mind. At least since the work of Brunswik (1955) people have repeatedly been likened to intuitive statisticians (Gigerenzer & Murray, 1987; Peterson & Beach, 1967). The core claim in this metaphor is that people are able to accurately estimate statistical properties in their environment and draw statistical inferences without the aid of calculators, statistical software or formal training in statistics. The early work showed promising results indicating that people have a surprising ability to correctly estimate statistical properties of the presented information (e.g., Peterson & Beach, 1967; Spencer, 1961, 1963).

Even though the metaphor of humans as intuitive statisticians is appealing, a long line of research has demonstrated that people have difficulties with a large number of judgment tasks involving statistical judgments. Such limitations have been explored within the heuristics and biases approach to statistical judgments (Gilovich et al., 2002; Kahneman et al., 1982). This approach questions the ability of people as intuitive statisticians by showing a broad range of instances where judgments and inferences far from coincide with what could be expected from statistical theory. In these instances, statistical judgments are the result of fallible heuristics and prone to biases (however see, Cosmides & Tooby, 1996; Gigerenzer & Murray, 1987). The ap-
proach emphasizes people’s reliance on heuristics or “rules of thumb” due to limitations in time, knowledge, and computational ability. The heuristics that people use often aid in decision situations but also produce characteristic biases (Gilovich et al., 2002; Kahneman et al., 1982). The apparent difficulty for people to act as intuitive statisticians, accompanied by a tendency to give biased judgments in a broad range of tasks even led some researchers to suggest that rather than being efficient intuitive statisticians people were more of cognitive cripples (Slovic, 1972). A different take on heuristics, and the biases they give rise to, has been offered by research concerned with fast and frugal heuristics (e.g., Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1999). According to this approach people have developed a set of simple and robust algorithms, available in an “adaptive toolbox”, that are ecologically rational. That is, heuristics that are adaptations to and exploit the structure of the environment in which people operate. For example, the recognition heuristic (Goldstein & Gigerenzer, 2002) is a decision rule that exploits the relationship between recognition memory and an object’s criterion value where recognized objects are more likely to have high criterion value. In contrast to the heuristics and bias approach the fast and frugal framework considers biases to be artifacts of otherwise adaptive and well-functioning decision rules.

Both the intuitive statistician and the two heuristics approaches focus on performance, but with respect to different focal areas. While research within the intuitive statistician approach has focused on the agreement between normative models and judgments acquired from inductive experience in controlled laboratory environments, research on heuristics and biases has highlighted biases attributed to the use of heuristics on knowledge acquired outside of the laboratory.

To resolve the apparent conflict between early research highlighting people’s ability to act as intuitive statisticians and the subsequent heuristics and biases approach a new framework has recently been suggested (Fiedler & Juslin, 2006b; Fiedler, 2000; Juslin et al., 2007). Within this framework people are considered naïve intuitive statisticians (Fiedler & Juslin, 2006b; Fiedler, 2000). The framework incorporates findings from the two previous approaches by suggesting that people have an ability to process information correctly, accompanied by an inability to evaluate the representativeness of that information (Juslin et al., 2007). The naïve intuitive statistician is defined by three assumptions (Fiedler & Juslin, 2006b; Juslin et al., 2007). First, people can store frequencies as both natural and relative frequencies and their judgments are accurate expressions of these frequencies. Second, people are naïve with respect to the effects of the processes that generate the samples they experience and generally assume that samples are more representative than they are. Finally, people are naïve with respect to properties of statistical estimators and do not correct for the fact that some estimators are biased. Several recent studies (Fiedler, 2000; Juslin et al., 2007; Kareev et
al., 2002) have demonstrated that the naïvety described above will give rise to apparent biases even though the operations performed on the data are normative. Kareev et al. (2002), for example, demonstrated that the tendency for people to underestimate population variance could be accredited to their failure to correct estimates of sample variance by $n/(n-1)$.

**Sampling**

An important process within the naïve intuitive statistician framework is sampling. That is, the acquisition of a subset of values from a variable in the environment. In general, people seem to be prone to base judgments on small samples from memory (Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Kareev et al., 2002) and there are several accounts of human judgment and decision making that assume sampling of information from memory prior to making a judgment or decision (e.g., Busemeyer & Townsend, 1993; Fiedler, 2000; Kahneman & Miller, 1986; Stewart et al., 2006; Tversky & Koehler, 1994).

In these accounts, samples are retrieved from memory. It is, however, possible to consider variations of the sampling process. Fiedler and Juslin (2006a), for example, suggested a three-dimensional taxonomy for the sampling process where origin, degree of conditionality and unit size constitute the three dimensions. The first dimension, origin, captures whether information comes from internal or external sources. That is, whether information is sampled from memory or the environment. Degree of conditionality, on the other hand, is thought to capture the extent to which the accessible information depends on factors such as a position in time and space or if a sender has made a selection of information before transmitting it. Finally, the dimension of unit size is used to capture the complexity of the information. That is, whether it is made up of elementary pieces of information or comes from a more complex knowledge structure (Fiedler & Juslin, 2006a). Further, Denrell (2005, 2007) illustrated how initial experiences can lead to selective sampling of environmental data. There are thus many factors that might influence the degree to which sample properties coincide with population properties. Regardless of how samples are generated they are, however, most often constrained by limitations in cognitive ability.

**Cognitive constraints**

The concept of bounded rationality (Simon, 1956, 1990), that is the insight that people’s limited time, knowledge, and computational ability will need to be considered when evaluating the rationality of a judgment, has often motivated research on judgment and decision making (see e.g., Gigerenzer & Goldstein, 1996). However, the implications of these constraints, such that people often have to rely on small samples (Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Juslin et al., 2007; Kareev et
al., 2002; Stewart et al., 2006) or linear additive integration of information (Juslin et al., 2009, 2011; Nilsson et al., 2009), are more rarely investigated.

There are, at least, two specific cognitive constraints that have the potential to influence intuitive statistical judgments. When statistical properties are calculated on values of a numerical variable, this operation is carried out within short-term memory (STM), a memory module with limited capacity. The number of observations that can be activated in STM is typically estimated to approximate 4±2 observations (Cowan, 2000). Thus, an intuitive statistician would have access to approximately four values simultaneously. Consider for example the task of estimating the average salary at your workplace. When performing such a calculation you would be able to activate no more than approximately four of your fellow co-workers’ salaries in STM at a time. Obviously, such a limitation would render your estimate of the average salary less accurate than if you had the possibility to perform your calculations on a larger sample.

Second, there is extensive support for the notion that controlled judgments are sequential and additive (Anderson, 1991; Hogarth & Einhorn, 1992; Juslin et al., 2008, 2011; Nilsson et al., 2009). The sequential nature of controlled thought suggests that information in the environment is integrated by considering one piece of information at a time. Further, the predominant operation that integrates previous knowledge with new data is additive in its nature. Because of the evidence suggesting that judgments in several areas are sequential and additive it is reasonable to assume that statistical judgments would also need to adhere to such a constraint. Thus, when creating your estimate of the average salary at your workplace you would need to consider one piece of information activated in STM at a time and your estimate is likely a weighted average of all of the activated values.

Bayesian Models of Intuitive Judgments

Although some of the early research on intuitive statistical judgments (e.g., Peterson & Beach, 1967) painted a fairly positive picture of people as intuitive statisticians the large majority of more recent research has been negative, to say the least. In contrast to this rather gloomy picture of people’s prospects as intuitive statisticians, recent rational accounts of human cognition (Oaksford & Chater, 2009; Tenenbaum et al., 2011) have indicated that people are rational Bayesian agents with a remarkable ability to integrate information in accordance with the laws of statistical theory. For example, Bayesian models have had success in predicting people’s behavior in such diverse areas as language acquisition (Kalish, Griffiths, & Lewandowsky, 2007), visual perception (Yuille & Kersten, 2006), sensorimotor learning (Körding & Wolpert, 2004), memory (Anderson & Milson, 1989), intuitive physics (Teglas et al., 2011), and intuitive statistics (Griffiths & Tenenbaum, 2006, 2011; Gweon et al., 2010). In a very general sense being a Bayesian
agent encompasses updating beliefs about a latent variable given some new data (Tenenbaum et al., 2011). More specifically, Bayesian statistics describe how an individual’s belief in a hypothesis \( h \) \( (P(h)) \) is updated in the light of new data \( d \). This is done according to Bayes’ rule, Eq. 1, which states that the posterior belief \( (P(h|d)) \) in hypothesis \( h \) given the data \( d \) is given by the likelihood \( (P(d|h)) \) of the data \( d \), given the hypothesis \( h \) and the prior belief in \( h \).

\[
P(h \mid d) = \frac{P(d \mid h)P(h)}{\int P(d \mid h)P(h)dh} \propto P(d \mid h)P(h)
\]

Thus, according to Eq. 1 the posterior belief in a hypothesis given observed data is proportional to the prior belief in the hypothesis times the likelihood of the observed data given that the hypothesis is true. In the current context it is important to notice that \( P(h|d), P(d|h) \) and \( P(h) \) are all probability distributions.

The large majority of Bayesian models investigate and specify the computational problems that people face and make no or few claims about the cognitive algorithms that people might implement (Griffiths, Chater, Norris, & Pouget, 2012). They merely suggest that people have some implementation that approximates Bayesian inference (e.g., Griffiths & Tenenbaum, 2011; Tenenbaum et al., 2011). However, it is reasonable to expect that any cognitive algorithm that aims to implement a Bayesian inference somehow needs to specify how the prior, likelihood, and posterior distributions are represented. Recent research has shown that a possible candidate process could be a sampling mechanism based on Monte Carlo (MC)-sampling (Shi et al., 2010; Vul et al., 2009). Shi and colleagues (2010) for example showed how exemplar models together with a type of MC algorithm called importance sampling could be a possible basis for Bayesian inference. However, although some attempts have been made at describing possible algorithms by combining documented cognitive processes (e.g., exemplar memory) with algorithms from Bayesian statistics (e.g., importance sampling) little research has focused on exploring what cognitive algorithms that people actually use. The present thesis contributes to the research investigating the processes underlyin Bayesian inference by suggesting which cognitive processes are plausible and which boundary conditions that should be taken into consideration. Whether people can actually implement and perform Bayesian inference is still a question under debate (see e.g., Bowers & Davis, 2012; Griffiths et al., 2012; Marcus & Davis, 2013). However, to the extent that people can at least approximate Bayesian inferences, the results of the present thesis could be a first indication of how people might realize their knowledge of the distributions \( P(d|h), P(h|d) \), and \( P(h) \) in such a process.
Distinctions in Cognitive Science

Dichotomies are ubiquitous in cognitive science and researchers have used them to describe processes in reasoning, judgment and social cognition (Evans, 2008) as well as in, for example, categorization (Ashby & Maddox, 2005). In the following, distinctions from cognitive science are reviewed with respect to different stages of an intuitive statistical judgment.

Any model that aims to describe intuitive statistical judgments need to specify which processes that operate in at least three different stages. First, the model needs to specify which, if any, operations that operate on the incoming data as it is experienced. Second, it needs to specify how the data is stored in memory while awaiting a judgment. Finally, the model is required to describe which processes that operate on the data at the time of judgment with respect to i) how data is retrieved from storage and ii) what, if any, operations that are performed on the retrieved data.

Both research within artificial intelligence/machine learning (Aha, 1997) and cognitive science (Juslin & Persson, 2002) have made the distinction between eager and lazy learning algorithms. According to Aha (1997) there are three distinguishing characteristics of lazy algorithms. First, they postpone any computations on the data until the time of a query and the data are stored as exemplars. Second, the stored data is combined to give a reply to a query. Finally, the reply, and any intermediate results, is discarded after answering the query. In contrast, an eager learning algorithm will make computations directly on the data as it is presented and store only the results of such computations while any input is discarded.

Similarly, research on frequency judgments (Dougherty et al., 2008; Hasher & Zacks, 1984; Zacks, Hasher, & Sanft, 1982) has suggested a distinction between online and retrospective models of frequency estimation (Dougherty et al., 2008). Here, online models (e.g., Hasher & Zacks, 1979) assume that frequency estimates are updated dynamically after each new data point is observed and that a judgment therefore merely requires a readout of pre-stored frequencies. Retrospective models (e.g., Brown, 1995; Hintzman, 1988) on the other hand do not record frequencies but rather infer them from individual events stored in memory (Dougherty et al., 2008).

A third similar distinction is found in research on categorization learning (e.g., Ashby & Maddox, 2005; Nosofsky & Johansen, 2000). Some models of categorization learning assume that during training people actively induce abstract summary representations of the categories, like prototypes or classification rules. Other models assume that people store representations of the concrete category exemplars in memory, which are retrieved at the time of classification, and the similarity to these exemplars are used for the classification.

Finally, research on human memory has suggested a taxonomy for different types of memory (e.g., Squire & Zola, 1996). More specifically, re-
searchers have suggested a distinction between semantic and episodic memories for declarative (explicit) memories (e.g., Battaglia & Pennartz, 2011; Squire & Zola, 1996). Semantic memory is a highly structured set of facts about the world that have been extracted inductively from statistical regularities in sparse and noisy data (Battaglia & Pennartz, 2011; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Episodic memory, on the other hand, is an autobiographical stream rich in contextual information (Tulving, 2002). The distinction between semantic and episodic memory is thus similar to that between exemplars and prototypes in categorization learning. Further, and related to how data for statistical judgments are stored, it has been suggested that semantic memories are formed by abstracting similarities from episodic memories (e.g., Battaglia & Pennartz, 2011; Wagner, Gais, Haider, Verleger, & Born, 2004).

To summarize: a model of statistical judgments needs to specify which processes that operate on at least three different stages and the distinctions highlighted above suggest which processes that are to be expected at each stage. First, previous research suggests that we could expect either that no operations are made on the data or that some summaries are extracted from the data as each new datum is presented during exposure. Second, several of the distinctions suggest that data could be stored either as raw data points similar to exemplars (episodic memory) or as summary representations similar to prototypes (semantic memory). However, it should not be ruled out that a combination of the two could occur. Finally, data could either be retrieved in the form of exemplars or as summary representations. Further, it is possible that no additional operations are made on the data at the time of judgment or that statistical properties are extracted from the retrieved exemplars. In addition, the suggested relationship between episodic and semantic memory indicates that it is possible to create more abstract summary representations (semantic memories) by combining the information in related episodic memory (see Kahneman & Miller, 1986, for a similar argument related to norms). It is reasonable to assume that the distinctions outlined above are not independent from one another and the constructs included in each of them are most likely overlapping. However, while the relationships between the constructs and distinctions are interesting it is beyond the scope of the present thesis to analyze the specific nature of them. In the following, I will concentrate on the initial distinction between lazy and eager algorithms.

An Eager vs. a Lazy Intuitive Statistician

With the important distinctions described above as a starting point the following two sections outline two possible models for intuitive statistical judgments. The first, summarized in the metaphor of the Eager intuitive
**statistician**, involves large-sample representations that capitalize on extensive experience while the second, the *Lazy intuitive statistician*, involves small-sample representations constrained by Short-term memory capacity. The models share two assumptions. First, variables that exist in the environment can be described by Objective Environmental Distributions (OED). Second, when a person observes the values from the OED a subset of the observations becomes encoded in long-term memory in the form of a Subjective Environmental Distribution (SED; Juslin et al., 2007). Three things regarding these assumptions are important to notice. First, at this stage there are no restrictions as to the nature of the SED. As will become evident later the SED can be represented in several different ways. Second, the two assumptions are silent on the question of whether operations are made on the data during exposure and merely states that the values, in some form, are encoded into long-term memory (LTM). Finally, the values that make up the SED could be either a representative or biased subset of the values in the OED.

**The Eager Statistician**

The first model, an eager cognitive algorithm, suggests that people have the cognitive capacity to spontaneously induce abstract representations of the properties of the OED from experience with the values of the OED. The process includes four steps from exposure to judgment. During exposure to the variable the eager model computes summary estimates of the statistical properties of the OED online, much like a spontaneous calculation of the intuitive equivalents of running estimates of the mean and the variance. This is done as each additional observation is presented. In the second step of the model (storage) these summary estimates are stored in LTM. While it is possible, and perhaps likely, that at least some of the individual values seen during exposure also become stored in LTM, the model suggests that these individual values are not included in the representation of the OED. Because extracted summary parameters cannot, as previously discussed, by themself support knowledge of the entire distribution it is likely that the formation of the representation also entails a priori “assumptions” about the distribution shape that become stored in LTM. To the extent that beliefs about distribution shape are explicit it is not expected that they are precise and quantitative, but have a rough, qualitative character capturing prototypical distribution shapes, like uniform distribution (e.g., “all values are equally likely”), unimodal distribution (e.g., “most values are in the middle”), and bimodal distribution (e.g., “most values are at the extremes”). Such a qualitative assumption together with summary statistics roughly specifies the distribution. There is evidence that any assumption about the distribution shape is likely to involve a normal (or unimodal) distribution (Flannagan et al., 1986; Fried & Holyoak, 1984). This thesis does not further address the specific nature of
such putative abstract representations of distribution shapes, or the viability of different processes whereby people acquire such representations. The representation of the OED in the eager model thus consists of two components, the extracted summary information and assumptions of distribution shape. At the judgment stage of the process, when a person is asked for an estimate of a property of the OED, the summary statistics along with assumptions of distribution shape are retrieved and used as a basis for a judgment.

The Lazy Statistician

The second model, a lazy cognitive algorithm, is a generalization and extension of the NSM as outlined in Juslin et al. (2007). The lazy model is a generalization of the NSM because while the NSM was developed for intuitive confidence intervals the present model aims to be model that can describe a range of intuitive statistical judgments. The lazy model suggests that, similarly to other lazy cognitive algorithms (e.g., Dougherty, Gettys, & Ogden, 1999; Hintzman, 1988; Juslin & Persson, 2002), no operations are made on the data as each new value is being experienced. During the second step in the model (storage) each data point is stored in memory as an individual memory trace. While there is no summary information extracted during encoding it is likely that the values are accompanied by some label which indicates that they belong to the same set or share a set of features. In contrast to the eager model, the SED of the lazy model contains only individual values possibly accompanied by a set of features. At the time of judgment, a random sample of observations (Subjective Sample Distribution; SSD) is retrieved from the SED, temporarily becoming active in short-term memory, and a property of this sample is used as a direct proxy for the population property (Juslin et al., 2007). For example, when asked to estimate the mean of the OED a SSD would be sampled from the SED and the mean of the SSD would be used as a proxy for the mean of the OED. The number of observations that could become included in the SSD would typically be limited by constraints on short-term memory (Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Kareev et al., 2002; Stewart et al., 2006). The sample of active observations in short-term memory is typically estimated to approximate 4±2 observations (Cowan, 2000). To illustrate, a person may not know the exact average salary of people working in his or her work place, but may retrieve a number of known salaries and, on this basis, estimate the average salary.

The qualification “naïve” in the NSM refers to the presumption that sample properties can be taken directly to describe population properties. While some sample properties, like mean and proportion, are unbiased, other sample properties, like variance and coverage, are biased. That is, while the expected value of the former coincides with the corresponding population
property under repeated random sampling, the expected value of the latter systematically distorts population properties (which is why sample variance needs to be corrected by $n/(n-1)$ to be an unbiased estimate of the population variance). The naïve presumption that a sample property can be used as a proxy for the population property thus affords accurate judgments with some sample properties (mean, proportion), but poor judgments with other sample properties (variance, coverage). The implication is that the judgment is constrained, and sometimes biased, by being naïvely projected in this way from a small sample.

Eager or Lazy?

Even though the distinction between an eager and a lazy intuitive statistician draws heavily on both previous theoretical and empirical work within cognitive science it has, as of yet, received little attention within the research of intuitive statistical judgments. Because previous research is scarce it is not evident a priori which of the two to expect. There are, in fact, theoretical and empirical arguments in favor of both models.

First, on a theoretical note, lazy algorithms, rather than pre-computing statistical summaries for every conceivable future demand, postpone computations until the specific need of computations are specified by a query. In contrast, eager algorithms pre-compute summary statistics before any demand is known. The pre-computed summary statistics of eager algorithms may succeed in closed and well defined environments, such as those given in laboratory tasks. It has, however, been suggested (Dougherty et al., 2008; Juslin & Persson, 2002) that lazy algorithms, such as exemplar models (Nosofsky & Johansen, 2000), the NSM (Juslin et al., 2007), the Minerva 2 model (Hintzman, 1988), and the enumeration model (Brown, 1995), afford greater efficiency and flexibility when the environment becomes more complex. It is possible to perform the computations needed with an eager algorithm efficiently when there are few variables. However, as the environment becomes complex the number of possibilities that need to be pre-computed for grows rapidly. It has therefore been suggested that models relying on online computations of statistical properties (e.g., cue validities) quickly become computationally intractable as the complexity of the environment increases (Dougherty et al., 2008).

An important feature of cognitive models is that they need to be able to scale to environments outside of the lab. That is, they need to be applicable not only in the stripped down tasks used in the laboratory but also in real world environments. As indicated above, while a lazy algorithm has the potential to scale to more complex environments than those investigated here, an eager algorithm is inherently more vulnerable to the combinatorial explosion in a complex environment.
While a lazy algorithm affords computational flexibility it does, however, place large demands on storage. Most lazy algorithms would require all, are almost all, encountered data points to be stored and available in memory at the time of a query. In contrast, eager models would store considerably less information because individual data points are discarded when summary information is extracted. Thus, in terms of storage, an eager statistician would afford much better cognitive economy than a lazy statistician.

On a more empirical note, previous research has shown that people are good at storing frequencies (Estes, 1976; Gigerenzer & Murray, 1987; Peterson & Beach, 1967; Zacks & Hasher, 2002) and have access to data after encoding (e.g., Malmi & Samson, 1983). Further, estimates of descriptive parameters seem to be constrained by STM capacity (Hendrick & Costantini, 1970). Also, statistical judgments such as confidence intervals (Juslin et al., 2007) seem to be generated by a lazy cognitive algorithm. However, previous research has also indicated that people seem to have expectations about the structure in presented data. For example, Brehmer (1974; see also, Kalish et al., 2007) showed that people expect positive linear relationships in function learning tasks and other studies indicated that people expect normal (or at least unimodal) distributions in a variety of tasks (Flannagan et al., 1986; Fried & Holyoak, 1984). Such expectations of structure would not be necessary for a lazy model but required for an eager model. In addition, several studies have indicated that pre-computed statistics are used by people to inform decisions. For example, Gigerenzer and colleagues (e.g., Gigerenzer & Goldstein, 1996) have argued that pre-computed cue validities are used to make decisions in a variety of tasks. There is thus some support for both models. The studies in the present thesis aim to investigate the viability of the two possible models of intuitive statistical judgments explicitly, thereby extending previous research.

Not Only Intuitive Statistics

This thesis is concerned mainly with statistical judgments. However, the question of whether these specific judgments are generated from a lazy or an eager process is not exclusively of interest to the area of statistical judgments. Several related areas assume that people have access to or can generate estimates of statistical properties, but do not explicitly specify the process by which this is done. For example, the Bayesian accounts of human cognition outlined above (Oaksford & Chater, 2009; Tenenbaum et al., 2011) often assume that cognitive processes are adaptations to distributions in the environment, the properties of which somehow need to be represented in memory. Further, research concerned with binary gambles has seen a recent interest in tasks where probabilities and outcomes are learned experientially (e.g., Hertwig et al., 2004), rather than explicitly stated (e.g., Kahneman & Tversky, 1979), requiring participants to generate statistical judg-
ments from memory at the time of a query. Finally, recent work on social judgments has suggested that people infer how other people are doing (e.g., the distribution of income) by sampling data available in their immediate social environment (Galesic et al., 2012). Whether people use a lazy or an eager cognitive algorithm might enforce boundary conditions on judgments made in all of these areas, because they all include the production of a statistical judgment.

A Lazy Model for Naïve Point Estimation

While the sections above outlined two possible processes for how intuitive statistical judgments could be formed, this section outlines a specific model derived from the lazy model. The model, introduced and tested in Study II, describes how people produce point estimates for novel objects. The presentation follows the development of the model, as described in Study II, by first introducing an initial model based on the NSM (Juslin et al., 2007) followed by a revised model that includes a response type discovered in Experiment 1 of Study II. Remember that point estimation involves using data from a sample to generate a best guess of an unknown observation. The model is used to predict the distributions of point estimates for “old” and “new” exemplars, that is exemplars previously experienced and exemplars seen for the first time at the moment of judgment.

The initial model

The initial model assumes two possible processes when an individual is presented with a probe $X$ and asked to give a best estimate of the value of $X$. First, if it is possible to retrieve the value of $X$ from LTM, this value is given as the point estimate by Direct Retrieval (DR). Second, if the value cannot be retrieved from LTM, the response is derived by Naïve Point Estimation (NPE). With NPE, a small sample (SSD) of $n$ observations is retrieved from the SED at the time of judgment, temporarily becoming active in short-term memory. The mean of the SSD is then reported as the point estimate (e.g., “the value of $X$ is probably close to the sample mean, like most of the other values in the sample distribution”). As discussed previously the size $n$ of the sample is constrained by short-term memory capacity and typically estimated to approximate 4±2 observations (Cowan, 2000). A qualification to this is, however, that if people store abstractions rather than raw data points they could possibly benefit from estimates that are based on larger samples of data.
Figure 2. Schematic outline of the naïve point estimation model including three types of responses; Direct retrieval (DR), Naïve Point Estimation (NPE) and Recognition failure heuristic (RFH). The bottom two rows of probability density functions illustrate the predicted densities of the three respective processes when the underlying distribution is unimodal (top row) and bimodal (bottom row). The density function of the entire process will be a weighted average of these three densities. The weight each response type receives is determined by $N_{SED}$ and $s$.

The revised model
The revised model, illustrated in Figure 2 and formalized below, includes both point estimates by DR (left branch) and naïve point estimates by NPE
The results of Experiment 1 of Study II, however, disclosed a third process similar to the recognition heuristic (Goldstein & Gigerenzer, 2002), for point estimates of “new” exemplars. The recognition heuristic could be formulated as follows: “If there are N alternatives, then rank all n recognized alternatives higher on the criterion than the N-n unrecognized ones.” (Marewski, Pohl, & Vitouch, 2010, p. 207, italics in original). It is thus a decision rule that could help people make decisions when recognition memory is correlated with an object’s criterion value (Marewski et al., 2010). Given the assumption that in most real-world distributions (not least, distributions of revenues, as used in the present experiments) large values are more salient and memorable, it would be useful for participants to conclude that if an object is not remembered it is likely to have a low value. Such a process was incorporated in the model in terms of a Recognition Failure Heuristic (RFH, right branch) where the lowest value in the SSD is reported (e.g., “Small (X) if not recognized(X)”) as the point estimate. This process was incorporated in the revised model to fully account for the data.

**The revised model formalized**

This section formalizes the revised model as described above and illustrated in Figure 2. The model is used to predict the proportion of point estimates ($p_i$) in a subinterval $i$ of length $l$, with starting point $\gamma_i$, end point $\gamma_i + l$, of the range of the underlying distribution. In Study II the underlying distribution is always a beta distribution with parameters $\alpha$ and $\beta$ (Beta($\alpha, \beta$)) defined on [0, 1]. The model has three parameters; $N_{SED}$ is the number of exemplars stored in LTM. With $N_{Tot}$ being the total number of items in the training distribution $N_{SED}/N_{Tot}$ is thus, given random sampling, the probability that a given probe can be retrieved from memory. The probability that participants will use the choice strategy NPE over the strategy RFH when they cannot retrieve the value of a probe from memory is given by $s$. Finally, $n$ is the number of items contained in the SSD when sampling from memory and is thus equivalent to short-term memory capacity.

According to the model the proportion of responses in an interval $i$ (DPE$_i$) is given by:

$$DPE_i = \frac{N_{SED}}{N_{Tot}} DR_i + \left(1 - \frac{N_{SED}}{N_{Tot}}\right) (s \text{NPE}(n)_i + (1-s) \text{RFH}(n)_i), \quad (2)$$

where
\[ DR_i = \int_{\gamma_i}^{\gamma_{i+1}} \text{Beta}(\alpha, \beta), \]  
(3)

and

\[ NPE(n)_i = \int_{\gamma_i}^{\gamma_{i+1}} N\left(\frac{\alpha}{\alpha + \beta}, \sqrt{\frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}} \frac{1}{n}\right), \]  
(4)

where \( N(\mu, \sigma) \) is a standard normal distribution with mean \( \mu \) and standard deviation \( \sigma \). Thus, \( NPE(n)_i \) is the density of the sampling distribution of \( \text{Beta}(\alpha, \beta) \) with sample size \( n \) on the interval \( [\gamma_i, \gamma_{i+1}] \). Finally,

\[ RFH(n)_i = 1 - \left( \int_{\gamma_i}^{\gamma_{i+1}} \text{Beta}(\alpha, \beta) \right)^n - \sum_{t=0}^{i-1} RFH(n)_i. \]  
(5)

Thus \( RFH(n)_i \) is the probability that the lowest value in the SSD is found in interval \( i \). The resulting distribution of point estimates (DPE) will thus be a linear combination of the three types of responses (i.e., a weighted mixture of the DR, NPE and RFH distributions).
The four empirical studies in the present thesis explore the general topic of which cognitive processes that are used by people to make intuitive statistical judgments. More specifically the four studies address three main questions. First, and with respect to the distinction between eager and lazy cognitive processes, to what extent are intuitive statistical judgments better described by a lazy than an eager cognitive process (Study I)? That is, do people spontaneously extract abstract representations of properties of numerical variables online during encoding or are calculations postponed until the time of a judgment and then carried out on a small sample from memory? Second, when applied to point estimates (Study II) and inference of a generating distribution (Study III), can a lazy cognitive process describe the judgments people make? Third, what boundary conditions govern the use of eager and lazy processes with respect to statistical judgments (Study IV)?

Study I: Calculate or Wait: Is Man an Eager or a Lazy Intuitive Statistician?

The main purpose of Study I was to investigate which of the two cognitive processes outlined in the introduction that best describes how people make intuitive statistical judgments. In Study I participants experience numerical values from a target variable and their knowledge of the target variable is later tapped using three tasks. In a proportion-production task, similar to methods used to elicit subjective probability distributions from experts (see e.g., Hora, Hora, & Dodd, 1992; Ludke, Stauss, & Gustafson, 1977; Winkler, 1970), they estimate the proportions of values from the target variable that fall into predefined, equally wide, intervals. In a visual-identification task, they are presented with several graphical illustrations of possible distributions of the target variable and choose the graph that best represents its distribution. Finally, in a descriptive task, participants estimate the central tendency (mean and median) and variability (mean absolute deviation) of the target variable.
Predictions

Based on theoretical arguments and previous empirical findings, all discussed above, it was hypothesized that people in general construct their representations of distribution properties post hoc by sampling from memory rather than induce them online. As in other learning tasks, however, it was expected that people are able to generate abstract representations if encouraged to do so during training. Below follow predictions derived from this general hypothesis and limiting conditions that affect whether a lazy or an eager cognitive process is likely to determine performance.

Format dependence

Abstract representations are independent of perceptual modality and response mode and can often be applied flexibly to a problem. This implies that the same representation can be used to inform judgments regardless of response format. Thus, if a respondent masters two response formats and the judgment in both is derived from the same abstract representation, we should expect similar performance in both formats. By contrast, if representations are reproduced post hoc by retrieval of observations there should be profound format dependence (Juslin et al., 2007). While some sample properties, like proportions, are unbiased estimates of the population property others, like variance, are not. Accordingly, participants are expected to perform well if the task invites the use of an unbiased sample property. In contrast, if the task invites the use of a biased sample property, less accurate performance is expected. With respect to the two formats, production and identification, used in Study I it was predicted that when the judgment is based on post hoc sampling from memory there should be a profound format dependence, with better performance with the production format than with the identification format. This format dependence should be much smaller in situations where the participants can benefit from an abstract representation.

Intentional learning

If people have an ability to generate abstract representations of distribution shape when actively encouraged to under training, this leads to distinct predictions when distribution shape is learned under incidental and intentional learning. In conditions of incidental learning, people should be confined to post hoc sampling from memory and therefore be victims of format dependence. In conditions with intentional learning, people should be able to induce abstract representations of distribution shape, improving the performance with the identification format, and thereby decreasing the format dependence.
Format order effects

It was also predicted that characteristic order effects for the two formats would appear under incidental learning. Participants encountering the identification format first are fully exposed to the unreliability of small samples and should therefore exhibit poor performance in this format, a performance that is later improved in the production format. In contrast, if the production format yields quite accurate representation of distribution shape it should allow participants to produce an abstraction that is informative for a later identification task. In other words, first making productions that imply, for example, a bimodal distribution, should allow participants to gain insight from this format that makes them less likely to mistake the distribution shape for a unimodal one with the subsequent identification format. Thus performing the production task before the identification task should make the format effect diminish.

Method

Both experiments in Study I adopt the same basic design. During a learning phase participants experience a target variable \((T)\) consisting of 60 numerical values. The values of \(T\) are shown repeatedly to participant and each value is associated with a text label. During a test phase participants are asked to perform three tasks. In a proportion-production task, they estimate the proportion of values from \(T\) that fall into ten predefined, equally wide, intervals. This task is similar to methods used to elicit subjective probability distributions from experts (Hora et al., 1992; Ludke et al., 1977; Winkler, 1970). In a visual-identification task, participants are presented with several graphical illustrations of possible distributions of \(T\) and choose the graph that best represents \(T\)'s distribution. Finally, in a descriptive task, participants estimate the central tendency (mean and median) and variability (mean absolute deviation) of \(T\).

In Experiment 1 the shape of the underlying distribution and the intentionality of learning were manipulated. Participants either observed a target variable that followed a symmetric unimodal \((T\sim\text{Beta}(3.4, 3.4))\) or a symmetric bimodal \((T\sim\text{Beta}(.33, .33))\) distribution (see Figure 1 for examples of unimodal and bimodal distributions). Further, participants were either explicitly told about the subsequent tasks (intentional learning) or not (incidental learning).

Experiment 2 aimed to replicate findings from Experiment 1 with regards to the shape of the underlying distribution and participants accordingly observed a target variable that was unimodal or bimodal. Also, the findings in Experiment 1 indicated that the salience of the end-points of the target variable might influence knowledge of the statistical properties of \(T\). Therefore, participants observed \(T\) either with salient or non-salient end-points.
Dependent measures

Performance in the production task and the visual identification task in both experiments was assessed by a mean absolute error ratio (MR) given by

\[ MR = \frac{MAE_R - MAE_S}{MAE_R}, \]  

(6)

where \( MAE_R \) is a measure of the difficulty of the respective task, given by the mean absolute error expected by random performance, and \( MAE_S \) is the performance of the participant. The MR scales performance against difficulty of the task to a \([-\infty, 1]\) scale where 1 represents perfect performance and 0 random performance. The descriptive task used signed and absolute deviation from the respective normative descriptive values as measures of performance.

Results

The variance in the two formats was not homogeneous in either of the two experiments. The MR-score was therefore standardized in the two formats, across each format separately, to achieve equated variances (and means), thereby enabling the use of ANOVA to analyze interactions. Main effects were investigated using non-parametric tests. The figures illustrate the effects with means and 95% confidence intervals for unstandardized variables.

Experiment 1 revealed three significant interactions (all other \( ps > .16 \)) illustrated in Figure 3. First, Figure 3A illustrates the significant format by distribution interaction (\( F(1,40) = 6.37, MSE = .55, p = .016 \)), with larger format dependence for the unimodal than the bimodal distribution. Second, Figure 3B illustrates a significant format by intentionality interaction (\( F(1,40) = 6.38, MSE = .55, p = .016 \)). As predicted if people induce abstract representations of the distribution with intentional learning, the format dependence is smaller than with incidental learning. Finally, Figure 3C shows the significant format by order interaction (\( F(1,40) = 4.17, MSE = .55, p = .047 \)). As predicted, performance in the production format was unaffected by order whereas performance in the identification format improved when it was preceded by the production format.

The main effects of distribution, format, and intentionality were investigated separately. An overall advantage in performance for the unimodal over the bimodal distribution was predicted. As is clear from Figure 3A there was no such overall advantage (Mann-Whitney: \( U = 249, Z = 0.79, p = .43 \)). Collapsed across both distributions the mean MR was .71 (\( SD = .15 \)) for the production format and .31 (\( SD = .42 \)) for the identification format. This confirmed the format dependence, with superior performance in the production
format (Wilcoxon; \( T = 78, Z = 5.23, p < .001 \)). Comparing performance with intentional versus incidental learning revealed no significant difference (Mann-Whitney: \( U = 256, Z = 0.64, p = .52 \)).

Figure 3. Mean absolute ratio for estimates in the bimodal and unimodal conditions (Panel A), incidental and intentional learning conditions (Panel B) and the production and identification formats (Panel C) as a function of task type (Panel A and B) or order (Panel C) in Experiment 1. Vertical bars denote 95%-confidence intervals.

The estimates of descriptive measures were analyzed both for accuracy (absolute deviation) and bias (signed deviation) using ANOVAs. With respect to accuracy, participants in the unimodal condition (\( M = 77.2, SD = 72.7 \)) gave more accurate estimates (\( F(1, 45) = 10.9, MSE = 11785, p = .002 \)) than did those in the bimodal condition (\( M = 150.5, SD = 119.6 \)). With respect to bias, variability was underestimated (\( M = -77.4, SD = 124.8 \)) whereas central tendency was slightly overestimated (\( M = 21.4, SD = 163.2 \)), a difference that was significant (\( F(1, 45) = 15.5, MSE = 15101.1, p < .001 \)).

The results of Experiment 2 revealed two significant interactions. First, as is clear from Figure 4A and the significant distribution by order interaction (\( F(1,40) = 4.71, MSE = .96, p = .032 \)), the order effects were different for the
unimodal and the bimodal distributions. With the unimodal distribution, the accuracy advantage of encountering the production format prior to the identification format was very modest but with the bimodal distribution, there was a clear advantage of performing the production format first. Second, and illustrated in Figure 4B, the order effects observed in Experiment 1 were replicated in Experiment 2 \((F(1,40) = 13.5, MSE = .57, p < .001)\).

In Experiment 2, the overall difference in performance between the unimodal \((M = .63, SD = .26)\) and bimodal distribution \((M = .39, SD = .41)\) approached significance \((Mann-Whitney: U = 195.5, Z = 1.89, p = .057)\). Comparing performance between the two distributions for the two formats separately replicated the findings from Experiment 1 with a significant difference in the production format \((t(46) = 3.04, p = .003)\) whereas the difference in the identification format approached significance \((t(46) = 1.96, p = .056)\).

Experiment 2 replicated the format dependence observed in Experiment 1, with better overall performance with the production format \((M = .72, SD = .15)\) than with the identification format \((M = .30, SD = .67)\) \((Wilcoxon; T = 205, Z = 3.80, p < .001)\). The ranges of the target variable used in Experiment 2 were created to be salient and non-salient. Comparing performance with obvious versus non-obvious range end points revealed no significant difference \((Mann-Whitney: U = 211.5, Z = 1.57, p = .11)\).

Experiment 2 replicated the main findings with respect to accuracy and bias of estimates of variability and central tendency found in Experiment 1. In addition, there was a significant main effect of range salience \((F(1, 44) = 4.4, MSE = 16745.3, p = .04)\) where the non-obvious condition was associat-
ed with a larger underestimation ($M = -84.6, SD = 135.3$) than the obvious condition ($M = -28.1, SD = 138.5$).

**Discussion**

Study I aimed at addressing the question of whether knowledge of a numerical variable is abstracted online during encoding, similar to a running mean or a posterior distribution, or if data are stored in a raw format during encoding and calculations are made first at the time of judgment. Put differently, are people eager intuitive statisticians that generalize the training data before the time of a query or do they, like lazy intuitive statisticians, wait to perform calculations until after a query?

The two experiments tested three main predictions. First, a lazy process predicts a format dependence effect with better performance when an unbiased estimator can be used than when it cannot. Both experiments documented strong format dependence effects. Second, if people have an ability to generate abstract representations of distribution shape, when actively encouraged to do so, an effect of intentional learning was predicted. There was no main effect of intentionality in Experiment 1. However, a format by intentionality interaction indicated a smaller format effect with intentional learning than with incidental learning. Finally, the predicted order effects were present in Experiment 1 and replicated in Experiment 2 with significant format by order interactions. In both experiments performance with the production format was unaffected by order while performance with the identification format benefited substantially from being preceded by the production format. The results of Study I invite two major conclusions. First, the process spontaneously engaged by the participants seems to be a post hoc sampling from memory. The naïve intuitive statistician thus seems to be lazy. Second, when instructed to, or strongly invited by the format, people have the ability to induce abstract representations.

In addition to the three main predictions, Study I also investigated the prediction that people will have a general response bias towards unimodality suggested by previous research (Flannagan et al., 1986; Fried & Holyoak, 1984). There was no overall advantage in performance when the underlying distribution was unimodal but in both experiments there was an effect of the shape of the underlying distribution in the production format (the effect was marginally significant in the identification format in Experiment 2) with better performance when the distribution is unimodal as opposed to bimodal.

Further, Study I replicated the finding from previous research that people are quite accurate at giving estimates of central tendency (Peterson & Beach, 1967; Pollard, 1984) but that variance is often underestimated (Kareev et al., 2002). The latter was extended to hold for explicit estimates of numerical (as compared to the more commonly used implicit estimates and perceptual variables) variables and the degree of the bias was shown to be related to the
distribution of the experienced variable. That explicit estimates for numerical variables are comparable to implicit estimates for perceptual stimuli indicates a common underlying process for estimates of variability. Although issues with scale invariance make direct comparisons difficult it might be a promising venue for future research to investigate a possible common mechanism.

Study II: Naïve Point Estimation

Study II took the same point of departure as Study I, that properties of numerical variables can be extracted to abstract summary representations online during encoding or calculated on a sample drawn post hoc at the time of judgment. However, instead of focusing on estimations of descriptive properties Study II investigated how people use knowledge of descriptive properties to make point estimates of an unknown observation. A point estimate includes using data to give a single best estimate of an unknown observation. There is no single normative strategy that can guide point estimates but, at least on the face of it, using larger data sets should yield more accurate estimates. From previous research it is, however, unclear how people use experienced instances of a numerical variable to form point estimates.

The results from Study I suggested that people are inclined to be lazy intuitive statisticians. In a similar vein, Study II introduced the possibility that point estimates can either be based on abstract summary representations, which are limited by long-term memory capacity, or on properties from small samples activated in short-term memory at the time of judgment. Study II, therefore, formulated a model for naïve point estimation, outlined above, based on the NSM (Juslin et al., 2007). The aims of Study II, accomplished by using the model to predict distributions of point estimates, were to a) further investigate if people are eager or lazy intuitive statisticians and b) test the model for naïve point estimation.

Predictions

The two possible ways in which point estimates can be informed by prior knowledge suggest diverging predictions, outlined next. First, if people can accrue abstract summary representations, their knowledge of the underlying distribution should improve with experience, whereas naïve sampling suggests that their judgments should continue to be constrained by short-term memory capacity, also in the face of extensive experience.

Further, whereas the ability to abstract a summary representation should allow people to detect a bimodal distribution shape, the sampling model suggests that the small samples that people have at their disposal will be too small to detect the shape of a bimodal distribution and potentially these will
even be misleading. With a bimodal distribution there should, therefore, be a tendency to place point estimates close to the mean of the distribution also after extensive experience and even if a pay-off schedule explicitly encourages maximizing behavior. A sampling model simultaneously predicts that even with small samples, judgments based on unbiased sample properties have the potential to yield accurate results. When probed for the proportions of the distribution that fall in predefined intervals, there should be a potential for accurate production of the distribution shape, because sample proportion is an unbiased estimator. Thus, when the underlying distribution is bimodal it was predicted that participants would place point estimates close to the mean of the distribution. At the same time, when using proportions they were expected to reproduce the bimodality of the underlying distribution that indicates a low probability for an object to be found close to the mean. Even though using a mean strategy in the point prediction task, similar to least-squares minimization, is not inherently wrong, it indicates an interesting set of believes where estimates are placed in a region where the probability of being correct, as explicitly indicated by participants in the proportion task, is the lowest.

Method

Study II consisted of four experiments adopting the same basic design, similar to the one used in Study I. During a learning phase participants experienced a target variable (T) consisting of 60 (Exp. 2, 50) numerical values. The values of T were shown repeatedly to participants and each value was associated with a text label. In a subsequent test phase participants performed a point-prediction task where they were shown the text labels associated with the values of T and were asked to make their best guess (give a point estimate) of the corresponding value. Participants made point estimates for old exemplars, seen during learning, and new exemplars not previously encountered. In addition, participants performed the proportion-production task described in Study I.

In Experiment 1 a unimodal (T~Beta(3.4, 3.4)) and a bimodal (T~Beta(.33, .33)) distribution defined the two conditions. The experiment tested the prediction that people are inclined to make point estimates according to a mean-strategy (i.e., close to the distribution mean), even when T is distinctly bimodal, while at the same time disclosing an ability to correctly reproduce the bimodal shape when probed with the proportion-production task.

In Experiment 2 the two conditions were defined by a positively skewed (T~Beta(.4, .6)) and a negatively skewed (T~Beta(6, .4)) distribution. The experiment was designed to verify that the tendency to make point estimates in the mid-interval, seen in Experiment 1, was truly a response to the experienced distribution.
Experiment 3 tested if the tendency to use a mean-strategy persisted even with a loss function that explicitly and strongly rewarded guesses close to the correct value and investigated if the effect could be manipulated by changing the format into one that invites learning proportions. In both the learning and test phase $T (T \sim Beta(\cdot33, \cdot33))$ was either continuous, as in Experiments 1 and 2, or discretized into ordinal categories using labels (“Small”, ”Quite small”, ”Medium”, ”Quite large”, ”Large”), which should encourage the learning of proportions. This, in turn, should make it easier to perceive the distribution shape as bimodal and thus invite more guessing in the most frequently occurring extreme categories of the interval. In the monetary incentive condition participants were informed that they would be paid a bonus proportional to the degree to which their answers in the test phase were close to the actual values. The instructions made it obvious to participants that an all-or-none incentive structure was used in the experiment.

Finally, Experiment 4 used the same two target variables as Experiment 1 and the same incentive structure as Experiment 3. In addition, half of the participants had their training divided into segments with tests occurring in between, to capture the training effect, while half of the participants were tested at the end of training only. The experiment was designed to replicate the findings in Experiment 3 with regards to the use of a loss function and to investigate the effect of experience on point estimates.

**Dependent measures**
In all four experiments the naïve point estimation model with two free parameters ($s$ and $N_{SED}$) was fitted to group data for the distribution of point estimates for new exemplars and point estimates for old exemplars using least-squares non-linear fitting. The model was fitted in two versions, one where point estimates are constrained by short-term memory capacity (STMC) and one where point estimates are constrained by long-term memory capacity (LTMC). The two versions are defined by values of the $n$ parameter in Eq. 2. In the STMC version $n$ was set to 4, a value motivated by the documented capacity of short-term memory (Cowan, 2000) and previous applications of the NSM (Hansson, Juslin, & Winman, 2008; Juslin et al., 2007). In the LTMC version $n$ was set to equal $N_{SED}$, that is the number of items in the SED. The dependent measures of model fit were Root Mean Squared Deviation ($RMSD$) and Pearson correlation ($r$).

**Results**
The results of Experiment 1 are summarized in Figure 5. Figures 5A and 5B present average assessed distribution shape based on the proportion-production task in 10 intervals on the continuum (the grey bars), as compared to the objective distribution shapes (the white bars). The empirical distributions of point estimates for old exemplars seen in training are pre-
sented in Figures 5C and 5D and the empirical distributions of point estimates for new exemplars first encountered in the test phase are presented in Figures 5E and 5F, in both cases classified in the same 10 intervals on the continuum (grey bars). (The black bars in Panels C to F for model predictions are discussed later.)

Figure 5 suggests three conclusions. First, with relative frequency estimates, the participants had fairly accurate knowledge of the underlying distributions, reproducing the unimodality in the unimodal condition and the bimodality in the bimodal condition (see the white bars in Figures 5A & 5B). Second, while the point estimates for old exemplars (the grey bars, Panels C & D) reproduced the underlying distributions (the white bars Panel A and B), in both conditions there is a relative shift for the new exemplars towards more predictions in the intervals 450 and 550 close to the mean of the distribution (the grey bars Panel E and F). Third, there is a bias to make low point estimates that is especially pronounced for new exemplars.

The bias to make low point estimates was interpreted as an inference resembling the recognition heuristic (RH) discussed by Goldstein and Gigerenzer (2002), which can be expressed as an inference rule: Large \((X)\) if recognized \((X)\). McCloy, Beaman, Frosch, and Goddard (2010) showed that this inference rule extends empirically to the closely related, but not formally equivalent: Small \((X)\) if not recognized \((X)\), although to a lesser degree. This finding led to the incorporation of a recognition mechanism in the model for naïve point estimation (see Figure 2 and Eq. 2-5).

Figure 5 illustrates the ability of the STMC-model (the black bars Panel C-F) to reproduce three important qualitative patterns in the data. First, the different distributions for old exemplars in the two distributions; second, the distributions of new responses that are a result of the RFH-responses, with the distinctive spike of a relatively high proportion of responses in the lowest interval for point estimates of new exemplars in the bimodal condition (Panel F) and the positive skew in the distribution of point estimates of new exemplars in the unimodal condition (Panel E); third, the model reproduces the high proportion of responses in the mid intervals for new exemplars in the bimodal condition, responses that are generated by use of NPE. The STMC-model also achieved better fit than the LTMC-version.

As in Experiment 1, participants in Experiment 2 correctly reproduced the bimodal distribution shape for the bimodal distributions. Crucially, the mean point estimates for new exemplars was significantly higher in the condition with a negative than with a positive skew suggesting sensitivity to the sample central tendency. As in Experiment 1 there was a better fit for the STMC-version than the LTMC-version of the model for naïve point estimation in both the positive skew and negative skew condition.
As in the previous experiments the results of Experiment 3 showed that participants had accurate knowledge of the distributions when the knowledge is elicited with frequency estimates. Further, the shift of point estimates to-
wards mean responses occurred mainly in the continuous condition. Consistent with the previous results Experiment 3 revealed a better fit for the STMC than the LTMC-version of the model. The results of Experiment 3 also showed that the tendency to place point estimates for new exemplars close to the distribution mean was resistant to monetary incentives.

In Experiment 4 the variability of the point estimates was not significantly affected by learning. Further, the tendency in the previous experiments, to place a disproportionate proportion of point estimates for new exemplars close to the distribution mean was replicated. This even with an explicit all-or-none monetary incentive structure. The STMC-version of the naïve point estimation model provided better fit than the LTMC-version and fit improved as training progressed. As predicted the relative use of NPE-responses over RFH-responses increased with learning. However, the proportion of responses directly retrieved from memory for point estimates of old exemplars increased from the first to the second test but then slightly decreased, indicating little additional learning from the second to the third test.

In the modeling results described above, the two models were fitted to group data. The number of responses generated by each individual participant and for new and old responses separately made it difficult, however, to fit the models to individual data with poor RMSD as a result. Nonetheless, comparing RMSD over the four experiments for the two models fitted to individual data revealed lower median RMSD for the STMC-version than the LTMC-version.

Discussion

The aims of Study II were to further investigate if people are eager or lazy intuitive statisticians and to test a model for the process of point estimation. The model proposes that limitations in short-term memory capacity and a reliance on small samples of information, constraints previously related to judgments (Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Juslin et al., 2007; Kareev et al., 2002; Stewart et al., 2006), will also influence how people make point estimates of unknown quantities. The four experiments compared the predictions by two possible accounts of how people use their knowledge of statistical distributions to make point estimates of unknown quantities.

Study II obtained three main findings: First, across all four experiments the results consistently favored the naïve sampling account of point estimation, suggesting that people base their point estimates on properties of a small sample constrained by short-term memory drawn from exemplars stored in memory during training, rather than having access to representations that benefit from most or all of the experience acquired during training.
Second, an unexpected recognition-based inference was documented whereby people, in the absence of other information, guessed that an unrecognized exemplar is more likely to have a low rather than a high value. After considering this additional mechanism, the model of naïve point estimation could appropriately account for the data. While including the recognition-based inference in the model for point estimates made it possible to account for the pattern of data found in Study II there was no independent test of this suggested process. An important task for future research will therefore be to test whether the guesses on low values are actually the result of a recognition failure heuristic.

Third, the results documented limitations of the naïve sampling framework. In Experiment 1, it was clear that the participants found it much easier to learn a unimodal than a bimodal distribution. Although this finding, in general terms, is consistent with the claim that the use of small samples may contribute to a belief or expectation that most distributions are normally distributed, there is no mechanism in the naïve sampling model that captures such an effect. Likewise, there is no mechanism that captures the better memory for large than small values in this specific task that drives the recognition-based inference and which makes the distribution with negative skew in Experiment 2 easier to learn than the corresponding distribution with positive skew.

Study III: Where Did That Come From? – Identifying the Source of a Sample

Both Study I and II suggested that people spontaneously engage in a lazy cognitive process when making intuitive statistical judgments. In everyday life people will engage in several types of intuitive statistical judgments in addition to, for example, point predictions and estimates of central tendency. One class of statistical judgments, present in several real life situations, is found in tasks where people are required to infer the distribution or process that has generated a small sample of values. The generating distribution is the underlying distribution from which a sequence of random numbers has been drawn (cf. Brown & Steyvers, 2009). Study III investigated to what extent people are able to make such inferences. Because of the findings in the two previous studies this was done within a naïve sampling framework (Fiedler & Juslin, 2006a; Fiedler, 2000; Juslin et al., 2007) with the general hypothesis that people spontaneously are lazy intuitive statisticians. In addition to the two cognitive processes discussed in the introduction, Study III outlined a third process, a memory inference, which might be used by participants to solve the task of identifying which distribution a given sample has been drawn from. As with a lazy algorithm the memory inference suggests
that no operations are carried out on the data during exposure and that observations become encoded in LTM. At the decision stage, however, the memory inference differs from a lazy algorithm. Rather than computing any properties on the experienced data the memory inference uses identity between previously experienced data and data in the presented sample to make the judgment. Thus, a participant using the memory inference would infer that a sample \( s \) comes from a distribution \( S \), where some values of \( S \) have been experienced, if a sufficient proportion of values in \( s \) can also be found in the version of \( S \) that is stored in LTM. In addition, Study III extended previous research on inference by introducing continuous variables that were learned by experience on a trial-by-trial basis.

Study III aimed at addressing 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 underlying distribution, and, allows them to experience all values in the test sample? Second, do people engage in a cognitive process that utilizes statistical properties of both the experienced data and the presented test sample to solve the inference task (i.e., a lazy or an eager process) or are their judgments based on a memory inference? Finally, does the possible use of statistical properties involve a small-sample representation, similar to the NSM, or a large-sample representation where population properties are pre-computed online during exposure?

In three experiments participants learned the distribution of a single variable (Experiment 1 and 2) or two variables (Experiment 3) and were later asked to identify which of two samples (\( s_1 \) and \( s_2 \)) that had been drawn from the experienced distribution. Both a lazy and an eager process assume that deciding which of the two samples that has been drawn from the experienced distribution is made by comparing the statistical properties of the two samples with the properties of the experienced distribution. More formally, let \( E \) be the set of values that participants experiences and let \( SP(x, y) \) be a function that calculates the deviation between the statistical properties of the two sets of values \( x \) and \( y \). In the formulation of the lazy and eager processes in Study III, \( SP(x, y) \) compares the means \( (\mu_x \text{ and } \mu_y) \) and standard deviations \((\sigma_x \text{ and } \sigma_y)\) of the two sets and is given by,

\[
SP(x, y) = \alpha_1 \frac{|\mu_x - \mu_y|}{\mu_x} + \alpha_2 \frac{|\sigma_x - \sigma_y|}{\sigma_x},
\]

where \( \alpha_1 + \alpha_2 = 1 \). The division by \( \mu_x \) and \( \sigma_x \) is to assure that both terms will be on the equivalent similarity scales. The inference that \( s_1 \) rather than \( s_2 \) has been drawn from the same distribution as \( E \) will be made if \( SP(E, s_1) < SP(E, s_2) \).
The difference between the eager and lazy model is found in how the statistical properties of \( E \) are estimated. The eager model suggests that they are extracted online during encoding. As such, the statistical properties of \( E \) are estimated on a large-sample representation of the underlying distribution. In contrast, the lazy model suggests that the properties of \( E \) are estimated on a sample of values drawn from STM at the time of judgment. Thus, if participants utilize a lazy cognitive process to solve the inference task their estimate of the properties of \( E \) is made on a small-sample representation of the underlying 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 samples used in Experiment 2 this experiment investigated the role of memory processes in solving the inference task. Finally, Experiment 3 investigated how the ability to make inferences about a generating distribution is influenced by the presence of multiple distributions.

Predictions

Study III derived specific predictions based on the assumption that one or the other of the three suggested processes is utilized to solve the inference task. The predictions were derived both analytically and using computer simulation. Computer simulations were also used to verify that the predictions would hold for a range of distributions, sample sizes, and parameter values.

First, because a memory inference relies on matching values from memory with those presented in the samples it will be sensitive to whether or not the presented values have been previously experienced, thereby predicting a distinct old-new difference. No such old-new difference is expected with either an eager or a lazy process. In addition, a memory inference model predicts 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 a 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 underlying distribution and the size of the sample. Samples drawn from populations with low variance (e.g., unimodal rather than bimodal, see Figure 1) and consisting of a large number of observations will have smaller standard errors than others. Because of the formulation of the SP-function (Eq. 7), a general prediction for a lazy model will be that performance will increase as the standard errors of the memory sample (MS) activated in STM and the presented test samples (TS) decrease. This occurs because when standard errors are small the difference between the sample means and sample stand-
ard deviation in the MS and TS sample will be small, if they are drawn from the same distribution. In the experiments of Study III, the standard errors of the MS and the TS are affected by two manipulations; the underlying distribution of the experienced values and the sample size of the TS.

Both Study I and II have indicated that judgments are more accurate if the underlying distribution is unimodal as opposed to bimodal. Samples from a unimodal distribution generally have smaller standard errors than those from a bimodal distribution. Therefore, the lazy model predicts a higher proportion of correct responses when the underlying distribution is unimodal as opposed to bimodal. This should occur even if the two distributions are equally well learnt. Because the eager model differs from the lazy model only with respect to how the properties of $E$ are estimated, it also predicts a unimodal-bimodal difference. However, while a change in distribution only affects the precision of TS if an eager model is used, it affects both the precision of TS and MS with a lazy model. Thus, the effect of distribution should be larger for a lazy than an eager model. In contrast, there should be no such differences between the two distributions for a memory inference.

Both the lazy and the eager model predicts that participants 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). This should occur because the standard errors of the TS become smaller as sample size increases. 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 previous predictions do not differentiate between a lazy and an eager model. Under certain task conditions the two models will, however, predict qualitatively different patterns of performance. While the lazy model predicts that performance will increase monotonically when the standard errors of the TS and MS decrease monotonically, the eager model predicts that under certain conditions the change in proportion correct will not be a monotonic function of the standard errors. This occurs because a change in the variability of the underlying distribution has less impact on an eager model than a lazy model and can therefore be compensated by an increase in sample size. However, the difference in variability between the manipulated distributions needs to be sufficiently large for the different predictions to arise. In Study III the distributions and sample sizes were chosen to give different patterns of results if participants used an eager or a lazy process. More specifically, if participants used a lazy process it was predicted that performance should be better with a unimodal distribution and small samples than with a bimodal distribution and large samples. In contrast, the use of an eager process predicted the opposite.
Method

All three experiments of Study III used a common method. During learning participants experienced a target variable \( T \) of 100 numerical values. The values of \( T \) were shown once to participants on a trial-by-trial basis. In Experiment 1 and 2, two sets of 100 unique values defined the unimodal \((T\sim\text{Beta}(5, 5))\) and bimodal \((T\sim\text{Beta}(.5, .5))\) conditions respectively. During test participants performed three tasks.

On each trial in the sample-identification task participants were presented with two samples of values and were to decide which of the two that had been drawn from the distribution of values observed during learning. Both samples on each trial had an equal number of values. One of the samples (target sample) consisted of values randomly drawn from the same distribution of values seen during learning while the other (distractor sample) included values drawn from another distribution. In Experiment 1 the two samples presented on each trial contained 5 or 10 values while Experiment 2 made a minor change and used 4 and 8 values. In addition, in Experiment 2 the values in the target sample had either been seen during learning (old sample) or were values from the same distribution but not previously seen (new sample).

The production task was equivalent to the production task used in Study I and II. In the descriptive task participants gave estimates of central tendency and variability. In addition, Experiment 1 and 2 included two measures, long term memory for numbers and numeracy (Lipkus & Peters, 2009; Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006; Peters, Slovic, Västfjäll, & Mertz, 2008), in order to measure a general proficiency for dealing with numbers.

Experiment 3 saw an alteration of the learning phase of the two previous experiments. In the experiment participants experienced both the unimodal and the bimodal distribution during learning. On each trial one value from each distribution was presented with separate text labels that identified the distributions. In addition, in the production task participants estimated the proportions of both the unimodal and bimodal distribution. Finally, on half of the trials in the sample task the target sample came from the unimodal distribution and on the other half it was drawn from the bimodal distribution. Thus, rather than manipulating which distribution participants had experienced, Experiment 3 manipulated which distribution was in focus on each trial of the sampling task. The two distributions seen during learning were labeled with the fictitious product names Alpha and Beta. Accordingly, on half of the trials in the sample task participants were asked to indicate which of the two samples that was representative for the Alpha product and on the other half which was representative for the Beta product.
Results

The results of Experiment 1, illustrated in Figure 6, revealed that participants performed better \( (F(1, 34) = 17.03, p < .001, MSE = .03) \) in the sample task when the underlying distribution was unimodal \( (M = .99, SD = .03) \) than bimodal \( (M = .82, SD = .17) \) and that performance was better \( (F(1, 34) = 25.0, p < .001, MSE = .003) \) with large \( (M = .94, SD = .13) \) than with small \( (M = .87, SD = .18) \) sample size. However, the interaction, illustrated in Figure 6, \( (F(1, 34) = 13.89, p < .001) \) indicates that sample size had an effect primarily in the bimodal condition.

![Figure 6. Proportion correct in the sample identification task in Experiment 1 as a function of sample size and distribution type. Vertical bars denote 95%-confidence intervals.](image)

It is possible that the effect illustrated in Figure 6 is due to participants in the bimodal condition having less accurate knowledge of the underlying distribution than participants in the unimodal condition do. Controlling for participants’ knowledge of the underlying distribution, by entering accuracy in the production task and accuracy of estimates of distribution properties as covariates into the analysis, showed that both the main effect of distribution \( (F(1, 32) = 10.63, p = .002, MSE = .03) \) and the distribution by sample size interaction \( (F(1, 32) = 7.4, p = .01, MSE = .003) \) remained after controlling for the covariates while the main effect of sample size did not.

In Experiment 2 the main effect of sample size was replicated \( (F(1, 29) = 7.61, MSE = .01, p = .01) \), with better performance with large \( (M = .81, SD = .23) \) than with small \( (M = .76, SD = .22) \) samples whereas the main effect of distribution was marginally significant \( (F(1, 29) = 3.22, MSE = .18, p = .08) \) with better performance in the unimodal \( (M = .86, SD = .14) \) than the bimodal...
al \((M = .72, SD = .26)\) condition. Even though the general pattern of data was similar to Experiment 1, the sample size by distribution two way interaction did not reach significance, and neither did any of the other interactions. Further, there was no new-old effect in the sample task \((F(1, 29) = .04, MSE = .02, p = .83; Old: M = .79, SD = .23; New: .79, SD = .25)\). As in Experiment 1, these effects remained after controlling for participants’ knowledge of the underlying distribution.

The lazy and the eager models predicted different orders with respect to proportion correct over the four sample size by distribution conditions. The lazy model predicted that proportion correct would decrease monotonically in the following order; unimodal-large, unimodal-small, bimodal-large, bimodal-small. In contrast the eager model predicted a change of place for the unimodal-small and bimodal-large conditions. That is, while the lazy model predicted better performance in the unimodal-small than the bimodal-large condition, the eager model predicted the opposite. Proportion correct in each of the four conditions is summarized in Table 1. As is evident, the order is consistent with the prediction of a lazy model. Pooling the data from the two experiments revealed a significant linear trend \((F(1, 132) = 28.5, p < .001, \beta = -.74)\), with polynomial contrasts, for the order predicted by the lazy model. There were no significant higher order contrasts. In addition, pooled over the two experiments the difference between the unimodal-small and the bimodal-large condition was significant \((t(66) = 2.26, p = .03)\) in the direction predicted by the lazy model.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Unimodal-Large</th>
<th>Unimodal-Small</th>
<th>Bimodal-Large</th>
<th>Bimodal-Small</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>.99 (.02)</td>
<td>.98 (.05)</td>
<td>.87 (.15)</td>
<td>.76 (.19)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.87 (.17)</td>
<td>.85 (.13)</td>
<td>.76 (.27)</td>
<td>.68 (.26)</td>
</tr>
<tr>
<td>Exp. 1 and Exp. 2</td>
<td>.94 (.13)</td>
<td>.92 (.11)</td>
<td>.82 (.22)</td>
<td>.73 (.22)</td>
</tr>
</tbody>
</table>

There was no relationship between the individual measures of proficiency for dealing with numbers and performance in the sample task in either of the two first experiments.

In Experiment 3 participants learned two distributions simultaneously. In the sample task the sample size and the focal distribution was manipulated. The analysis of the sample task revealed that neither the main effect of sample size \((F(1, 17) = .37, MSE = .03, p = .55; \text{Small}: M = .83, SD = .26; \text{Large}: M = .85, SD = .22)\), the main effect of focal distribution \((F(1, 17) = .50, MSE = .04, p = .49; \text{Unimodal}: M = .82, SD = .27; \text{Bimodal}: M = .86, SD = .21)\) nor the sample size by distribution interaction \((F < 1)\) reached significance.
Comparing performance in the sample task of Experiment 3 with the corresponding task in Experiment 1 (Figure 7), which was possible because the same stimulus material was used, indicated that performance was only marginally affected by the introduction of a second distribution. As the figure shows, there was a main effect of condition ($F(2, 51) = 7.2, p = .002, MSE = .04$). Post hoc analysis showed that performance in the unimodal condition of Experiment 1 was better than both of the other two conditions, which did not differ significantly from each other. Second, the main effect of sample size ($F(1, 51) = 8.9, MSE = .008, p = .004$) showed better performance with a sample size of 10 values ($M = .91, SD = .16$) as opposed to 5 values ($M = .86, SD = .18$). The sample size by condition two-way interaction approached significance ($F(2, 51) = 2.97, MSE = .008, p = .06$). This was due to mainly the bimodal condition in Experiment 1 being sensitive to sample size. Further, participants learned both distributions to an almost equally degree with primarily accuracy in the production task being lower in Experiment 3 as compared to the two conditions of Experiment 1.

![Figure 7. Proportion correct in the sample identification task for the conditions in Experiment 1 (bimodal and unimodal) and Experiment 3 (mixed) as a function of sample size. Vertical bars denote 95%-confidence intervals.](image)

**Discussion**

The findings from Studies I and II indicated that using the NSM framework to study the cognitive processes of intuitive statistical judgments could be a promising endeavor. In Study III the focus of interest was directed towards the cognitive process that governs how a generating distribution is inferred from a sample. This because even though previous research has investigated how samples from memory are utilized in decisions (e.g., Busemeyer &
Townsend, 1993; Stewart et al., 2006) and which information in samples that people use (e.g., Bar-Hillel, 1979; Chesney & Obrecht, 2011; Kareev et al., 2002; Obrecht & Chesney, 2013) little attention has been given to the underlying processes.

Study III extended previous research by addressing three main questions in three experiments. First, can people solve an inference task that uses a continuous variable, allows them to experience values from the underlying distribution, and allows them to experience all values in the sample? Second, are the statistical properties of the distribution and the sample used to solve the task or are judgments based on a memory inference. Finally, does the possible use of statistical properties involve a post-hoc sampling process similar to the NSM or a process comparing sample properties with population properties that are pre-computed online during exposure?

The results of Experiment 1 suggested two major conclusions. First, people seem to be quite apt at inferring whether a sample of numerical values has been drawn from the same distribution as a set of previously experienced values. Second, the cognitive process used to make the inference seems to be one that uses the statistical properties of the test sample. Further, the results supported the idea that the inference was generated by a lazy rather than an eager cognitive process.

Experiment 2 was designed to rule out the possibility that participants are engaged in a memory inference rather than a lazy or an eager process when solving the sampling task and to replicate the main findings from Experiment 1. As in Experiment 1 the results indicated a remarkable ability of participants to identify which of the two samples that was drawn from the same distribution as the experienced values. Even in a situation where the test samples contained only four values, none of which had been shown previously, participants still performed well above chance. Further, there was no effect of the old-new manipulation, indicating that people use statistical properties of the test sample and the experienced values rather than rely on a memory inference. In addition, the marginally significant effect of distribution in Experiment 2 lends further support to the idea that the cognitive process of inference in the task is a lazy one.

While the lazy and eager models predicted similar effects of distribution and sample size, they made different predictions for the ordering of conditions with respect to proportion correct. In both Experiment 1 and 2, polynomial contrasts and the critical difference between the unimodal-small and the bimodal-large conditions suggested that the order predicted by the lazy model was more likely.

Experiment 1 documented a sample size by distribution interact. The same pattern of results was found in Experiment 2 but there the interaction did not reach significance. Not being predicted by any of the three models, the interaction was unexpected. The interaction could be explained by an asymmetry in difference in standard errors between the large and small sam-
ple of the two distributions. If the difference in standard errors between the unimodal-large and unimodal-small conditions is smaller than that between the bimodal-large and bimodal-small conditions this might give rise to a pattern of results similar to those observed. It is also possible that the interaction seen in Experiment 1 was the result of a ceiling effect in the unimodal condition. The experiments in Study III were not designed to distinguish between the two possibilities and future research will be needed to address the question.

In Experiment 3 the results of the two previous experiments were extended by investigating if 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 to an almost equal degree as did the 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. These results taken together suggest that the participants were able to keep the values from the two distributions separated throughout the learning phase and access properties specific for each distribution in the test phase.

There was no effect of sample size found in Experiment 3, which may be due to one of two possibilities. First, the somewhat lower overall performance in Experiment 3 might have pushed down performance to a region where there is no longer a difference between the two sample sizes. Because of the comparable results between Experiment 3 and the bimodal condition in Experiment 1, where the sample size effect was evident, this is however unlikely. Further, the sample size effect was observed in Experiment 2 where the overall performance is similar to that seen in Experiment 3. A second possibility is that participants were able to evaluate both samples against both of the underlying distributions when making inferences. It is possible that such a strategy would reduce the effect of sample size observed in the other two experiments. Experiment 3 also lacked an effect of distribution. However, rather than manipulating the experienced distribution, Experiment 3 manipulated the focal distribution in the sampling task. Because Study III did not formulate which processes that might be used in the two distribution case it is difficult to make strong predictions of the influence of focal distribution. What process that supports inference in the two distribution case, and the influence it might have on the sample size effect and focal distribution effect, should be an interesting question for future research.
Study IV: Are all Data Created Equal? - Exploring Boundary Conditions for a Lazy Intuitive Statistician

The three previous studies have evaluated the distinction between eager and lazy cognitive processes as the foundation for intuitive statistical judgments. In general the results seem to support the notion of people as lazy intuitive statisticians. However, even though there is reasonable evidence to support the claim that people generally construct statistical judgments with a lazy cognitive algorithm, there are several boundary conditions and assumptions that are not addressed in Study I-III. More specifically, a lazy algorithm is expected to afford computational flexibility in complex situations. However, in order to achieve efficiency it requires a sufficiently large portion of undistorted data to be accessible from memory at the time of judgment. Little research has addressed if this is the case for statistical judgments. Further, a lazy algorithm requires a minimum of assumptions about the properties of the experienced data to function well. If people enter laboratory tasks with strong a priori assumptions about the data, this might indicate that a lazy algorithm is not the default cognitive strategy. However, few studies have investigated if people do enter laboratory tasks with assumptions about properties of the data.

Study IV included two experiments aimed at addressing two major questions. First, is the sampling process from LTM at the time of judgment unbiased with respect to when the data is experienced? Second, do people enter the laboratory task with expectations about the statistical properties of the data? Experiment 1 was designed to investigate whether common memory effects such as primacy and/or recency are present in statistical judgments. Such a finding would indicate that participants’ sampling of the data from LTM is conditional on when it was presented during learning. The design of Experiment 2 allowed, in addition to replicating the results of Experiment 1, for an evaluation of the extent to which participants enter the laboratory task with expectations about the properties of the data. Both eager and lazy cognitive algorithms are assumed to calculate statistical properties on the experienced data. However, as discussed in Study III it is possible that some tasks might be solved successfully without any calculations if memory for specific values could be used directly. To extend the findings of Study III, both experiments in Study IV therefore included a manipulation to evaluate whether the observed responses were due to memory for specific values or if participants indeed made inferences of statistical properties of the data.

Method

Both experiments in Study IV used a common method. During exposure participants experienced a target variable of 120 numerical values. The values were shown once to participants on a trial-by-trial basis. Participants’
task was to observe the values carefully in order to answer questions about them at a later point in the experiment.

Two sets of 120 values, both uniformly distributed in the interval $[1, 1000]$, were used. One set was presented during exposure while the other was withheld until the test phase. The three conditions, unimodal-bimodal (U-B), bimodal-unimodal (B-U), and uniform (UN), were defined by the order in which the data was presented. In the U-B condition the first 60 values followed a unimodal distribution while the second 60 values followed a bimodal distribution. This order was reversed in the B-U condition. In the UN condition both the first and the second 60 values were uniformly distributed. The transition from the first to the second 60 values was not announced or implied in any condition. Thus, the logic of the experiment is quite simple. Participants in all conditions received the same uniformly distributed data, but in different order. If participants are victims of for example a recency effect, they will perceive data in the B-U condition to be unimodal but in the U-B condition the impression will be a bimodal distribution.

During test participants performed four tasks. In the identification task, participants chose which of 7 histograms that best described the data seen during exposure. In the production task participants assessed how many of the values from the experienced data that fell into ten equally wide intervals. In the sampling task participants were presented with 8 matrices (5x6) containing 30 values each and were asked to choose 10 values from each matrix that were the most representative of the data seen during exposure. Four of the matrices contained values from the observed data and four contained values from the data set withheld from participants during exposure. Finally, in the descriptive task participants estimated the central tendency (mean and median) and variability (mean absolute deviation; MAD) of the experienced data.

Experiment 2 made one minor change to the method described above. In this experiment the exposure phase was interrupted after 60 values with a test (production task) after which participants experienced the remaining 60 values.

**Dependent measures**

Performance in the identification, production and sampling tasks were evaluated with a shape sensitive index (Shape Index; SI) calculated for each of the three tasks separately. With the range $[1, 1000]$ divided into 10 equally wide intervals $[1, 100]$, $[101, 200]$ ... $[901, 1000]$ and intervals numbered according to: $1...j...10$, SI is calculated as:

$$SI = \sum_{i=1}^{2} (p_i - x_j) + \sum_{i=4}^{7} (x_i - p_i) + \sum_{i=9}^{10} (p_i - x_j),$$

(8)
where $x_i$ is the participant’s judgment of the proportion of the distribution in the $i$:th interval and $p_i$ is the normative proportion (i.e., .1). In the production task $x_i$ is the frequency given explicitly by participants divided by 120. In the identification task $x_i$ was given by the frequency counts of the graph chosen by participants, divided by 120. Finally, in the sample task values were categorized to the ten intervals and the proportion of values in each interval gave $x_i$. Thus, SI is both a measure of the degree to which estimates deviate from the underlying uniform distribution and sensitive to the shape of the estimated distribution. An SI of 0 will indicate that judgments are uniform while SI > 0 and SI < 0 will indicate an estimated distribution that is unimodal and bimodal respectively.

Results

Figure 8 illustrates the main results with respect to estimates of distribution shape in Experiment 1. As is illustrated, SI was higher ($F(2,90) = 5.11, p = .008$) with the identification task ($M = .12, SD = .34$) than with the production ($M = .07, SD = .2$) and the sample ($M = .01, SD = 2.59$) tasks and post hoc tests showed that the identification-sample difference was the only one that reached significance.

![Figure 8](image)

*Figure 8.* Performance in Experiment 1 in the three tasks given by the Shape Index. Dashed line indicates unbiased (uniform) estimates. Vertical bars denote 95%-confidence intervals.

Figure 8 also suggests an apparent bias towards unimodality both in the identification and production tasks while there is no such bias in the sample
task. Single sample $t$-tests showed that this bias was statistically significant (deviation from zero) in both these tasks ($p < .05$).

The accurate performance in the sampling task may be due to specific memory of the experienced values rather than by any knowledge of the properties of the experienced values. A SI was therefore calculated for the new and old values separately. Analyzing these data revealed no old-new effect ($F(1,45) = 1.61, p = .21$; Old: $M = .02, SD = .27$; New: $M = -.01, SD = .27$).

Estimates of central tendency and variability were evaluated for accuracy and bias using absolute and signed deviations, respectively, from the normative value. With respect to accuracy, estimates of central tendency ($M = 78.6, SD = 46.7$) were more accurate ($F(2,44) = 10.9, p = .002$) than those of variability ($M = 119.6, SD = 81.7$). Analyzing for a possible bias revealed that whereas central tendency was slightly overestimated, variability was underestimated to a large extent.

Experiment 2 included both an intermediate and a final test. Comparing performance in the final test over the three tasks showed a significant main effect of condition ($F(2,39) = 4.0, p = .027$) with a larger deviation in the uniform ($M = .12, SD = .23$) than in the bimodal-unimodal ($M = .02, SD = .2$) and unimodal-bimodal ($M = -.05, SD = .23$) conditions. Post hoc tests indicated that the uniform/unimodal-bimodal difference was the only pairwise difference that reached significance. As in Experiment 1 there were no old-new differences in the sampling task ($F(1,39) = 0.05, p = .83$; Old: $M = .025, SD = .23$; New: $M = .03, SD = .22$).

![Figure 9. Performance in Experiment 2 in the Intermediate and Final test (production task) in the three conditions. Dashed line indicates unbiased (uniform) estimates. Vertical bars denote 95%-confidence intervals.](image-url)
More importantly, comparing performance in the intermediate test with that of the final test (both production tasks) revealed a strong condition by test time interaction ($F(2,39) = 69.3, p < .001$). As is evident from Figure 9, the interaction is due to participants in all three conditions reproducing a distribution consistent with the experienced data, both at the intermediate and final test.

Finally, analyzing estimates of central tendency and variability for accuracy and bias showed that estimates of central tendency ($M = 43.8$, $SD = 50.0$) were more accurate ($F(1,37) = 21.0, p < .001$) than those of variability ($M = 109.5$, $SD = 83.1$) and that whereas there was no bias for estimates of central tendency, variability was underestimated.

Discussion

Study I-III generated reasonable support for the idea that the preferred cognitive process when people make intuitive statistical judgments is described by a lazy cognitive algorithm. However, there were several boundary conditions and assumptions that had been left unaddressed. Study IV aimed at investigating two major questions directed towards these issues. First, is the sampling process from LTM at the time of judgment unbiased with respect to when the data is experienced? Second, do people enter the laboratory task with expectations about the statistical properties of the data?

In general the results from both experiments showed that the participants were very responsive to the properties of the data. At the same time they were remarkably resistant to the memory effects of primacy and recency. Most remarkably, perhaps, was the lack of effects for ratings of variability that undergoes a dramatic change from the first to the second half of the presentation sequence.

Experiment 1 investigated if common memory effects influenced statistical judgments. There were no order-effects present in the data, which seems to indicate that participants had access to all of the data at the time they made their judgment. Put differently, if participants’ preferred cognitive strategy to produce a statistical judgment is a lazy process, they calculate properties on samples drawn from memory without conditionalizing on when the data was encountered. A caveat to this conclusion is that the participants might not have been responsive to the underlying distribution at all but rather gave uniform estimates as the result of some default ignorant strategy. Experiment 1 also revealed a slight tendency for participants to report unimodal distributions. In both Studies I and II it was argued that this is a consequence of a lazy cognitive process. However, the results might also arise if participants enter the task with an assumption that the data will be unimodally distributed. Experiment 2 addressed both the issue of nonresponsivity and of prior assumptions about the underlying distribution.
Experiment 1 replicated the results of Study I with better performance in the production task than in the identification task. These results were also extended by showing that a task format (the sampling task) that is even closer to the suggested lazy cognitive algorithm will see participants perform even better. The lack of old-new differences in the sample task further indicates that the performance advantage is not only due to memory of specific values.

The results of Experiment 2 suggested both that the participants were sensitive to the underlying distribution and that they were efficient in incorporating new data as it was presented. The lack of order effects seen in Experiment 1 was thus not likely a result of nonresponsivity. It is possible that the intermediate test made participants more aware of the purpose of the data presentation and that they thereby would produce a more accurate representation of the underlying distribution. However, performance on the final test was very similar to that of Experiment 1 indicating only a small effect by the introduction of an intermediate test.

Experiment 2 also investigated if participants enter the task with expectations about the properties of the presented data. Previous research has indicated that such possible expectations are likely to be unimodal (Flannagan et al., 1986). As is evident in Figure 9, if participants have any expectations about the shape of the underlying distribution, it influences their judgments very little. In terms of priors, the participants seem to have very weak priors concerning the presented data and there is little evidence suggesting that such a prior would be unimodal.

With respect to estimates of central tendency and variability, both experiments documented results consistent with those found in Study I, Study III, and other previous research (Kareev et al., 2002). While estimates of central tendency were fairly accurate (although slightly overestimated) variability was systematically and strongly underestimated. In addition to being consistent with previous findings it is also consistent with what could be expected from a lazy cognitive algorithm.
Intuitive statistical judgments are an integral part of people’s everyday life and have received a fair amount of attention from researchers over the years. In this thesis I have tried to extend what we know of people as intuitive statisticians. The overarching aim was to derive a process model for intuitive statistical judgments that could be modified to describe different types of judgments. Therefore, the thesis outlined and tested two models for statistical judgments summarized in the metaphors of the Eager Intuitive Statistician and the Lazy Intuitive Statistician. The main question was thus, are people’s intuitive statistical judgments best described by an eager or a lazy cognitive algorithm?

Study I proposed two opposing models of intuitive statistical judgments. It was suggested that people could either be eager intuitive statisticians that generalize from the data as it is presented or lazy intuitive statisticians that record individual data points as they are presented but postpone any calculations until a time when they are asked for an estimate. As outlined in the introduction this distinction is related to the several previous distinctions in cognitive science. For example, the distinction between eager and lazy learning algorithms, as used in in machine learning and artificial intelligence (Aha, 1997) and cognitive science (Juslin & Persson, 2002), between exemplar based memory and cue abstraction, as used in research on multiple cue probability learning (e.g., Juslin et al., 2008), between prototypes and exemplars used in category learning (Ashby & Maddox, 2005; Nosofsky & Johansen, 2000), between online and retrospective models (Dougherty et al., 2008) in frequency learning, and to the distinction between semantic and episodic memory (e.g., Battaglia & Pennartz, 2011; Tulving, 2002). From previous research it was not clear which of the two to expect. On the one hand people are surprisingly good at storing frequencies (Estes, 1976; Gigerenzer & Murray, 1987; Peterson & Beach, 1967; Zacks & Hasher, 2002) and they seem to have access to individual values of an experienced variable (Malmi & Samson, 1983) on the other hand people seem to expect structure in data (Brehmer, 1974; Flannagan et al., 1986; Fried & Holyoak, 1984) and use pre-computed statistics to inform judgments (Gigerenzer & Goldstein, 1996). Further, and on a more theoretical note, an eager statistician could benefit from storage gains when a variable is reduced to only a few parameters. In contrast, a lazy algorithm affords more flexibility in the complex world that people have to orient in (Juslin & Persson, 2002).
fact, relying on on-line computations may even become computationally intractable as the complexity of the environment increases (Dougherty et al., 2008) suggesting that while a lazy algorithm has the potential to scale to situations outside of the laboratory an eager algorithm will do so only with much more difficulty.

In general, the results from Study I supported the notion of people as spontaneously being lazy intuitive statisticians. However, the results also indicated that under certain specific circumstances people may have the ability to induce abstract representations of experienced data. While the primary goal of Study I was to investigate if statistical judgments are governed by an eager or a lazy cognitive process, Study II aimed to show how a more specific type of judgment, point estimates, are made. The distinction between the eager and lazy intuitive statistician invites two major possibilities of how point estimates are formed. If people are eager statisticians, point estimates may only be informed by the parameters abstracted online during encoding. If on the other hand they are lazy statisticians point estimates should be informed by properties of small samples activated in short-term memory at the time of judgment. The results of Study I indicated that the latter rather than the former was to be expected. In the first experiment of Study II, one additional, recognition based, inference was documented and included in a model of naïve point estimation. The model, including the recognition based process is illustrated in Figure 2 and formalized in Eq. 2-5 and was formulated in two versions. One version describes a distribution of point estimates that would emanate from an eager intuitive statistician which is long-term memory constrained (LTMC), the other would produce the distribution of point estimates for a lazy intuitive statistician being short-term memory constrained (STMC). In all of the four experiments of Study II the STMC-model provided better fit than the LTMC-model. Further, the tendency to place point estimates in accordance with the STMC-model was robust to monetary incentives and parameters of the model changed in the predicted direction during learning.

In Study III the focus was changed to another type of statistical inference, namely the identification of a generating distribution. The main goal of the study was to investigate 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 underlying distribution, and allows them to experience all values in the sample? The results from Study III indicated that people had a remarkable ability to solve such tasks. Indeed, even when given as little as four data points in a test sample participants were able to solve the task well above chance levels. Also, having participants learn two variables, with very different properties, simultaneously influenced performance only marginally. The second question addressed by Study III was whether people use the statistical properties of the underlying distribution and the samples to solve the inferences task or if their judgments are based
on a memory inference. Here, the results indicated that people do perform some kind of statistical calculations rather than rely exclusively on memory of specific values. Finally, Study III investigated if the possible use of statistical properties involved a lazy or an eager cognitive algorithm. In general the results favored the conclusion that the participants spontaneously engaged in a lazy cognitive algorithm to perform inductive inference from samples.

The three first studies found reasonable support for the idea that people in general engage in a lazy process to construct intuitive statistical judgments. The studies, however, left several boundary conditions and assumptions of a lazy process unaddressed. Study IV was aimed at investigating two important questions with respect to constraints on a lazy cognitive algorithm. First, in order to be efficient the lazy intuitive statistician requires that a sufficiently large portion of undistorted data is available for sampling from memory. Study IV investigated if this is the case by looking at possible memory effects (primacy/recency) on estimates of the data. A finding of primacy and/or recency effects would have indicated that participants’ access to the experienced data is conditional on when it was experienced, thereby limiting the pool of undistorted data needed for an efficient lazy algorithm. The results of the experiments in Study IV indicated that there were no memory distortion effects. If, as suggested by Study I-III, people are indeed lazy intuitive statisticians these results indicate that the sampling from memory at the time of judgment is unbiased. A caveat to this conclusion is that even though sampling at the time of judgment might not be contingent on the order in which data is presented it is possible that it is contingent on other factors. One possible candidate factor could be similarity, which have been shown to influence memory sampling in categorization learning (Ashby & Maddox, 2005; Nosofsky & Johansen, 2000) and which is important when approximating Bayesian inference with exemplar models and importance sampling (Shi et al., 2010). It is an interesting venue for future research to explore how the samples used by a lazy statistician are formed and what factors that influence the process.

Second, it lies at the heart of a lazy cognitive algorithm to make minimal assumptions about the statistical structure of the data before a request for a judgment (Aha, 1997). The design of the second experiment in Study IV allowed for the evaluation of such possible prior expectations on the data. Any a priori assumptions of the statistical structure of the data are expected to become evident when people have only experienced a small set of new data. For example, if the participants in Experiment 2 of Study IV had strong expectations that the to be presented data was unimodal it would be expected that a majority of them would report a unimodal distribution at the intermediate test. However, the results of Experiment 2 in Study IV showed that already after having seen 60 data points, participants had converged on the shape of the actual underlying distribution. Thus, the results indicated that
the participants entered the experiment with none, or at least weak, assumptions about the properties of the data. It was further shown that the participants were, to a large extent, data driven. It is, of course, possible that people do have expectations about the data but that these expectations include a uniform distribution rather than the unimodal distribution that has been suggested by previous research (Flannagan et al., 1986; Fried & Holyoak, 1984). However, as indicated by the fact that participants reported distributions that were similar to the experienced data already after being exposed to a small set of data points, any expectations that people do have are probably weak.

Three of the studies (I, III, and IV) investigated the accuracy of estimates of central tendency and variability. The overall results indicated that estimates of central tendency were fairly accurate and unbiased while estimates of variability were quite poor and strongly underestimated the variability of the experienced data. Previous research (Kareev et al., 2002) has indicated that variability is underestimated for perceptual variables and using implicit tasks. The present thesis extends these results by showing that the same pattern of results emerges for numerical variables and explicit estimates. Not only is this a novel finding, it is also it is also consistent with what could be expected from a lazy cognitive algorithm.

Previous research has often painted a somewhat gloomy picture of people’s ability to monitor their judgments and decisions. Research concerned with statistical judgments has even suggested that people often are victims of cognitive myopia (e.g., Fiedler, 2000, 2008) in the sense of not acknowledging the processes that shape available information and thereby the biases present in the data (e.g., Fiedler, 2000, 2008; Juslin et al., 2007; Kareev et al., 2002). Two findings in the present thesis suggest that people might, at least under some circumstances, use alternative strategies that include monitoring of their own judgments in a way that previous research would not expect. First, Study II documented a recognition-based inference that indicated that when participants did not recognize a novel object they assumed that it had a low value. If participants are indeed using the recognition based inference it is necessary for them, for the inference to work, to have the insight that in general they remember large values but not small. Second, Study III indicated that people could solve the sample task well above chance even after having experienced two distributions simultaneously. It was suggested that one possible strategy to solve the task was to evaluate the two presented samples against both experienced distributions. Deciding which of the two came from a bimodal distribution, which was the harder task, could be done by recognizing which sample that came from the unimodal distribution and choosing the other one. Again, such a strategy would require considerable metacognitive monitoring to be solved successfully. Neither the RFH-process from Study II, nor the just described process from Study IV, were tested independently. However, the results suggest that peo-
ple may use strategies including metacognitive monitoring, when making intuitive statistical judgments, to a larger extent than what could be expected from previous research.

There are some limitations to the studies presented in this thesis. It is argued that STM-capacity is a key limitation for intuitive statistical judgments and thus imposes constraints on the cognitive process underlying such judgments. None of the studies however include a measure of STM-capacity. Although the modeling done in Study II indicates better fit for a STMC-model than a LTMC-model it should be an important task for future research to empirically outline the link between STM-capacity and statistical judgments in greater detail. Further, it should be noted that previous studies investigating the NSM have verified that a relationship with STM-capacity does exist (Hansson et al., 2008; Juslin et al., 2007).

In Study II, the parameters of the model were fitted to both group and individual data. However, the low number of point estimates for each individual and each type of response led to a poor fit for individual data. This, in turn made comparisons between the parameter values and fit of the two models difficult. It would be an important issue for future research to evaluate the two models on individual level data. Further, in its current formulation the model contains three parameters; s, n, and N_SED. The first is, in contrast to latter ones, theoretically underspecified which make strong predictions about it difficult. It remains for future research to investigate the relationship between RFH- and NPE-responses to be able to formulate stronger predictions of the value of s under different conditions.

In all studies except Study II, the two accounts of intuitive statistical judgments, the eager and the lazy intuitive statistician, are investigated by evaluating qualitative predictions about an expected pattern of results. As such, the two process models are evaluated in an indirect manner and the conclusions in the thesis will depend on the validity of the predictions. Although the predictions are derived from results and theoretical accounts found in previous research (e.g., Juslin et al., 2007) and although the results throughout the four studies converge on a common interpretation, it would be warranted to conduct further research including more direct approaches before concluding that people are indeed lazy intuitive statisticians. A more direct approach would be to, similar to the strategy used in Study II, formulate and compare computational models for the two processes of the eager and the lazy intuitive statistician.

Psychologists often need to make trade-offs between the control available in laboratory studies and the generalizability found in field studies. In the words of Brehmer and Dörner we need to be “Escaping both the narrow straits of the laboratory and the deep blue sea of the field study” (Brehmer & Dörner, 1993, p. 171). The studies in the present thesis share four features that may limit the possibility to generalize the findings to situations outside of the laboratory. First, the target variable that participants is expected to
learn is presented during a short time frame, approximately one hour. Second, the time between learning and when participants’ knowledge of the target variable is elicited is brief. Although these task features mimic how someone might learn and use time-series data, they are of course idealizations. In most real-life situations the learning of a target variable is spaced out over a long period of time. In addition, the time between encountering instances of a target variable and a subsequent use of this information is likely to be longer than that used in the current studies. Conducting studies that separate exposure and test with days or weeks would be an interesting extension of the present work. However, the main aim of the present study was not to fully mimic real-life situations but rather to investigate what processes that govern intuitive statistical judgments in near optimal condition.

A third feature relates to the cover stories presented to participants. While they were created to be plausible, it was also important that they did not induce strong expectations by the participants. There have been several previous attempts to investigate what knowledge people have of higher order properties of numerical variables when the variables are assumed to have been experienced in the everyday life of participants (e.g., Griffiths & Tenenbaum, 2006; Nisbett & Kunda, 1985). However, because these investigations have had little control over what data participants have actually encountered, it is difficult to determine how accurate the participants’ knowledge actually was. In order to achieve the necessary control needed to investigate the underlying cognitive process of intuitive statistical judgments the present studies were therefore conducted using laboratory settings with plausible but non-informative cover stories. Thereby, of course, a trade-off was made between control and the possibility to make strong generalizations of the findings. The fourth task feature common to all four studies is the use of specific tasks to elicit judgments of distributions. Whereas explicit judgments of distributions are not very common in the everyday life of most people, they are often used to elicit knowledge from experts in various domains (O’Hagen et al., 2006). For example, a subjective probability distribution for a country’s inflation rate might be elicited from a group of experts (Engelberg et al., 2009). The findings in the present thesis may therefore help to improve such judgments by indicating which processes that inform and constrain them.

The thesis proposes and contrasts two simple accounts of human statistical judgment, illustrated with the metaphors of the eager and lazy intuitive statisticians. It is of course possible to formulate other accounts of the type of judgments used in the thesis. For example, in Study II the model proposed that people use the mean of the SSD as their estimate. They might, however, instead use something more complex like trimmed means or density estimation. However, such accounts would be more complex and would place additional storage and computational demands on the human cognitive system, demands that probably lie beyond the capacity limitations documented in
previous research (Cowan, 2000; Dougherty & Hunter, 2003; Gaissmaier et al., 2006; Hansson et al., 2008; Juslin et al., 2011, 2007; Kareev et al., 2002; Nilsson et al., 2009; Stewart et al., 2006). The results of the studies show that it is sufficient to assume that people base their statistical judgments on a small sample activated in short-term memory to explain the data.

In the extent formulation of the lazy intuitive statistician it is assumed that only one sample is drawn from STM at the time of judgment. However, reproducing the distribution in the production task would be difficult, especially with a bimodal underlying distribution, if only one sample is activated. Thus, considering the fairly good performance in the production task it is possible that participants treat the production task not as one single judgment but as several repeated judgments. This would imply that multiple samples are drawn sequentially from LTM to complete the task. It will be an important task for future research to establish if people can activate multiple samples and, in such case, how the information from the individual samples is integrated to reach a judgment.

Finally, the thesis was limited to investigating intuitive statistical judgments related to statistical theory. More specifically it investigated estimates of statistical properties and intuitive statistical inference. With this limitation, judgments related to probability theory were excluded. Building an overarching theory of intuitive statistical judgments would, of course, require extending the lazy and eager models to include also probability judgments. This will most certainly be an interesting challenge for future research.

To summarize: the present thesis makes two major contributions with respect to previous research. First, research investigating people as intuitive statisticians has primarily been interested in measuring how accurate estimates of various statistical properties are under a plethora of different circumstances (see e.g., Pollard, 1984, for a review). However, much less attention has been given to investigating the cognitive processes that are at the core of intuitive statistical judgments. Inspired by models for specific judgments (e.g., Fiedler & Juslin, 2006b; Juslin et al., 2007), important distinctions previously made in cognitive science (e.g., Aha, 1997; Dougherty et al., 2008; Juslin & Persson, 2002), and previous empirical results (e.g., Malmi & Samson, 1983), the present thesis outlined two models that were tested in four studies. As such, the present thesis is a first attempt at outlining a process model that could possibly be applied for a variety of statistical judgments.

Second, as discussed in the introduction, there are several research areas that, often implicitly, assume that people have access to statistical properties of experienced information. The present thesis suggested that such knowledge is generated by a lazy cognitive algorithm that postpones judgments until the time of a query. Because the lazy cognitive algorithm imposes constraints on, for example, the accuracy of statistical judgments it is rea-
sonable to assume that judgments relying on intuitive statistical judgments will also be constrained. Thus, future research in various areas will need to take the limitations of a lazy intuitive statistician into consideration.
Completing a PhD is a daunting endeavor. Successfully reaching the finish line requires the help and support of several important people. I have been fortunate enough to have a wonderful group of people around me, both at work and at home, during my endeavor.

First, I would like to thank my excellent supervisor Anders Winman. Anders, you have all the qualities a PhD-student could wish for in a supervisor. Not only do you have a deep knowledge of experimental psychology, you also invest enormous amounts of time, effort, and patience into everything we work on. I could not have hoped for a better guide to become a researcher.

I would also like to thank my second supervisor Peter Juslin. Peter, it has been a privilege to have you as a supervisor. Your insightful comments have sometimes made me leave meetings more confused than when I entered. They have, however, pushed me to think further than to the next p-value and broadened the scope of my research. Thank you for your time and engagement, which far exceeds what could ever be expected from a second supervisor.

At both the early and latter stages of writing this thesis Gustaf Gredebäck, Håkan Nilsson, and Hedvig Söderlund has contributed with very valuable comments to make it better. Thank you for your time and commitment.

Research is often a team sport and I would like to thank past and present members of the Judgment and Decision Making Group in Uppsala (aka. the Team). Ebba Elwin, Mona Guath, Göran Hansson, Maria Henriksson, Linnea Karlsson, Neda Kerimi, Philip Millroth, and Håkan Nilsson, thank you for valuable comments and discussions an all things research related. And thank you for discussions and events not at all research related.

Though not formally on the Team but nonetheless an important part of everyday life at the department are the fellow PhD-students. I would especially like to thank Robin Bergh and Jonas Engman. Robin, if you took the time to read the entire thesis before coming to the acknowledgements you would realize that yours is probably better. Also, thank you for many inspiring discussions on subjects too many to count and for getting me out on the occasional bike ride. Jonas, after our endless discussions on research productivity and academic writing I always leave with the feeling that I should be a little bit more like you. If I could have implemented all the ideas we’ve discussed this thesis would have been twice as long and twice as good. I, however, leave that challenge to you.
The people in the research group and at the department have of course been important for completing this project. Equally important for the project, and an integral part in keeping me sane, is my family. Thank you Gun-Britt and Nils-Eric (in memoriam) for raising me to be curious and persistent and thank you for supporting me when I suggested that going back to school was a good idea. Elisabeth, you have shown how almost anything is possible if you work hard. You are a true inspiration. Alfred, you are the joy of my life. It’s a privilege to see you grow up and I strive to keep up with you in terms of learning speed. Finally, Maria, you are the love of my life. Thank you for all your love and support, not only when it comes to work matters. I know my explanations of what I do at work bore you to death, but love you for listening with a smile every time.
References


A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences.