barcelona gse
graduate school of economics

Experiencing Simulated Outcome

Robin Hogarth
Emre Soyer

June 2010

Barcelona Economics Working Paper Series

Working Paper nº 470

Running head: EXPERIENCING SIMULATED OUTCOMES

Experiencing sequentially simulated outcomes: A guide to improve statistical inference

Robin M. Hogarth[1,2] & Emre Soyer[1]

[1]Universitat Pompeu Fabra, Department of Economics & Business, Barcelona

[2]ICREA, Barcelona

Abstract

Whereas much literature has documented difficulties in making probabilistic inferences, it has also emphasized the importance of task characteristics in determining judgmental accuracy. Noting that people exhibit remarkable efficiency in encoding frequency information *sequentially*, we construct tasks that exploit this ability by requiring people to experience the outcomes of sequentially simulated data. We report two experiments. The first involved seven well-known probabilistic inference tasks. Participants differed in statistical sophistication and answered with and without experience obtained through sequentially simulated outcomes in a design that permitted both between- and within-subject analyses. The second experiment involved interpreting the outcomes of a regression analysis when making inferences for investment decisions. In both experiments, even the statistically naïve make accurate probabilistic inferences after experiencing sequentially simulated outcomes and many prefer this presentation format. We conclude by discussing theoretical and practical implications.

Keywords: probabilistic reasoning; natural frequencies; experiential sampling; simulation.

JEL codes: C00; C11; C15; C91

Version : June 22, 2010

"*I have come to feel that the only learning which significantly influences behavior is self-discovered, self appropriated learning*" (Carl Rogers, 1961, p.276).

In the last five decades, many studies have documented difficulties people have in reasoning probabilistically (see, e.g., Cohen 1960; 1972; Edwards, 1968; Kahneman, Slovic, & Tversky, 1982; Hogarth, 1975; 1987). One way of characterizing this literature is to note that whereas people have well-developed capacities for dealing with data in the form of frequencies and averages (Peterson & Beach, 1967; Nisbett, Krantz, Jepson, & Kunda, 1983), they have greater difficulty in understanding concepts of variability, the role of random error in phenomena such as regression toward the mean, and implications of probabilistic reasoning that involve combinations of events (see, e.g., Lathrop, 1967; Bar-Hillel, 1973; Cohen & Chesnik, 1970; Cohen, Chesnik, & Haran, 1971; Kahneman & Tversky, 1973; Tversky & Kahneman, 1974; 1983). The human mind, it seems, is quite effective at aggregating information in an additive manner but the multiplicative demands of probability theory are hard to master (Juslin, Nilsson, & Winman, 2009).

In considering this state of affairs, it is important to recall that, although humans have always faced uncertainty, probability theory itself only dates from the mid-17th century (Daston, 1988), and is a discipline that requires considerable intellectual abstraction. It is thus not unreasonable to conjecture that probability theory solves some problems in ways that are foreign to the response tendencies that have been honed by human evolutionary forces. At the same time, however, the demands of our modern, technologically oriented society increasingly require the ability to understand the implications of statistical reasoning. In managing an investment portfolio, for example, it is essential to understand the distributional implications of potential

returns to assess tradeoffs between risks and returns. In the practice of medicine, failing to assess correctly the probabilistic implications of test results can have disastrous consequences. And yet, most of the time, people – including professionals – deal with statistical information in an intuitive manner that may or may not lead to appropriate inferences (Gigerenzer, Gaissmaier, Kurz-Milke, Schwartz, & Wolosihn, 2007).

An argument can always be made for better teaching of statistical reasoning. However, whereas we are not against "better teaching" (who is?), we doubt whether this is a solution. The reason is that whatever is taught needs to be reinforced through practice and feedback and it is not clear that there will always be opportunities for all types of problems. What is needed is a general approach – based on well-established theoretical and empirical grounds – that can be easily adapted to specific situations. In fact, the main implication of this paper will be to suggest such an approach.

In an extensive review of issues of risk perception and communication in the medical domain, Gigerenzer et al. (2007) note that the ways in which statistical information is presented have large, and often predictable effects on the inferences people draw. For example, people are impacted far more by descriptions of risk reduction – due, say, to some intervention or treatment – when this is expressed in relative as opposed to absolute terms, e.g., as 50% instead of from 2 in 1,000 to 1 in 1,000. Similarly, physicians show remarkable improvements in probabilistic reasoning – using Bayesian updating when interpreting test results (e.g., mammograms) – when data are presented in *natural* frequency format as opposed to the more typical probabilistic statements (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1966; Hoffrage & Gigerenzer, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Brase, 2008). Indeed, frequency representations have also been observed to improve inferences in the famous "Linda" problem

(Tversky & Kahneman, 1983; Fiedler, 1988; Hertwig & Gigerenzer, 1999), "sample size" tasks (Sedlmeier, 1998)[1], and the "Monty Hall" problem (Krauss & Wang, 2003). In summarizing these and other studies, Gigerenzer et al. (2007) wisely state

> ….statistical literacy is largely a function of the outside world and …can be fostered by education and, *even more simply, by representing numbers in ways that are transparent to the human mind* (p. 54, italics added).

Unfortunately, it is not always clear how to define *a priori* what is "transparent to the human mind" in the presentation of statistical data. For example, graphs may or may not be helpful in different circumstances (Soyer & Hogarth, 2010), and providing probabilistic information in frequency formats does not guarantee that people will make appropriate inferences (Griffin & Buehler, 1999; Mellers & McGraw, 1999; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002).

Nonetheless, the work of Gigerenzer and his colleagues suggests a research strategy: first, identify the cognitive mechanisms that people perform well naturally (i.e., without specific training); and second, structure statistical inference tasks in a form that exploits these mechanisms.

In this paper, we follow this strategy. First, we identify one important mechanism that humans have been demonstrated to possess in terms of handling data; specifically, the ability to encode automatically in memory frequency information about events they experience across time.[2] Second, by using simulations we provide statistical representations of problems that allow

[1] The "hospital problem" (Tversky & Kahneman, 1974) that we present below is an example of a "sample size" task.

[2] As an everyday example of this mechanism, imagine that you are asked how many times you have been to the cinema (or undertaken another similar activity) in the last three months. Most people have little difficulty in providing a rapid and fairly accurate answer to this question. And yet, they do not consciously record the frequency of their visits to the cinema or make a mental note (of this or other frequencies) in case someone asks the question just posed.

people to exploit this mechanism, that is, by providing opportunities to experience *sequentially generated* frequency data.

Whereas using simulated data in this general manner may seem "obvious" to statistical practitioners, there is a remarkable lack of systematic scientific evidence concerning its use as suggested here. In fact, we know of only two studies that have explicitly investigated effects on probabilistic inference of experiencing sequential frequency data. In the first, Christensen-Szalanski and Beach (1982) investigated whether sequentially observing 100 instances of either base-rate or base-rate and diagnostic information would impact subsequent assessments of Bayesian posterior probabilities. They found no effect for base-rate information alone, but a favorable impact for base-rate and diagnostic information. In the second study, Betsch, Biel, Eddelbüttel, and Mock (1998) showed that, when people explicitly sampled frequency information, they were more appropriately sensitive to base rates in a Bayesian updating task than if provided with probabilistic information. In a related investigation, Sedlmeier (1999, Chs. 10, 11) has used what he terms a "flexible urn model" in the shape of a computer simulation model that does allow participants to observe data dynamically and where their probabilistic inferences are quite accurate. (We consider Sedlmeier's work again in the General Discussion.)

The evidence on the natural encoding of frequency information is both uncontroversial and overwhelming. It has been summarized by, amongst others, Hasher and Zacks (1979; 1984) and Zacks and Hasher (2002). As their studies show, humans have a remarkable capacity for the accurate encoding of frequency information. Moreover, this cognitive activity demands little by way of attention, does not require intention, is invariant to learning, age, and many individual differences, and also involves recognizing the frequencies of subcategories of experienced events. That it is a basic cognitive mechanism that was probably developed through evolutionary

pressures is reinforced by the findings that several non-human species show similar capacities, e.g., in understanding frequency distributions associated with different sources of food. Moreover, consistent with this evidence is that accumulated by Fiedler and his colleagues concerning human sampling of data for inferential purposes (Fiedler, 2000; Fiedler, Brinkmann, Bestsch, & Wild, 2000; Fiedler & Juslin, 2006). These authors show that people are quite accurate in encoding the data they have observed; the blame for systematic inferential errors, they argue, lies with failures to exercise the meta-cognitive judgment necessary to offset biases in the sampling process.

Our suggestion, therefore, is to transform statistical reasoning problems into a form that exploits the ability to encode and interpret *sequentially* observed frequency information. That is, instead of asking people to solve statistical reasoning problems analytically, we propose having them *experience* frequency data sequentially thereby allowing their natural encoding capabilities to inform their answers. Before proceeding, however, we make three important remarks.

First, our work builds on the pioneering contributions of Gigerenzer and his colleagues who perceptively drew attention to the difference between representing probabilistic problems by natural frequencies as opposed to probabilities. Our innovation consists of extending their argument to what we consider its logical conclusion. Specifically, Gigerenzer and Hoffrage (1995, p. 686) defined natural frequencies as "actually experienced in a series of events" noting that "From animals to neural networks, systems seem to learn about contingencies through sequential encoding and updating of event frequencies…". They further expanded on the meaning of *natural sampling* as involving the "sequential acquisition of information by updating event frequencies *without* artificially fixing the marginal frequencies" (p. 686). On the other hand, as far as we can tell, participants in their experiments never actually experienced data

sequentially, that is, "as a series of events." Instead, they observed totals. That is, Gigerenzer and his colleagues presented data in the form of *summarized* natural frequencies.

This is an important point. Experience is typically not in the form of summed frequencies presented in some tabular format. Instead, frequencies (as described by Hasher and Zacks above) are characterized by being *experienced sequentially* – one-at-a-time – across some period of time and/or space. The foraging animal, for example, does not consult a table of data in a natural frequency format when deciding which of two potential sites has produced more food. Instead, across time it has accumulated experience – either directly or by observation – of how often the two sources have yielded food. To test, therefore, whether frequency data lead to accurate probabilistic inferences, one must allow the organism to *experience* the data in sequential format as opposed to simply providing a count.

Second, numerous studies conducted with animals have shown appropriate sensitivity to environmental probabilities and, for the most part, what could be considered rational behavior (see, e.g., Real, 1991; 1996; Weber, Shafir, & Blais, 2004). However, it is important to emphasize that to test animals' response tendencies in these studies, it was necessary to have them first observe sequentially generated frequency data.[3]

Third, to provide somebody with a sequential frequency representation of a probabilistic problem requires the ability to construct an appropriate simulation model. This, it could be argued, implies a greater "input" on behalf of the person structuring the problem with the consequence that results should not be compared with attempts to answer problems in the usual probabilistic formats. We have three answers to this objection. One is that knowing *how* to present probabilistic information in natural frequency formats (see, e.g., Gigerenzer & Hoffrage,

[3] Alex Kacelnik, personal communication, April 2010.

1995) also requires more knowledge than just presenting respondents with the usual probabilistic format. The second is that allowing respondents to experience "raw" data seems the simplest test of human ability to handle statistical problems. And third, as noted above, the presentation of sequential frequency information is the norm in studies of rational behavior in animals. Thus, if it is legitimate to allow animals to use this form of data presentation, why deny humans?

In this paper, we test the effects on statistical inference of experiencing sequentially simulated outcomes in two experiments. In Experiment 1, we present participants with seven well-known problems from the literature. These concern (1) Bayesian updating, (2) the "birthday" problem, (3) the conjunction problem, (4) the Linda problem (Tversky & Kahneman, 1983), (5) the hospital problem (Tversky & Kahneman, 1974), (6) regression toward the mean, and (7) the Monty Hall problem. Two groups of participants, familiar with statistical reasoning, answer these questions both with and without the aid of experience in the form of sequentially simulated outcomes in a design that permits both between- and within-subject comparisons. Participants are further required to provide a third answer to determine their remuneration. Finally, a further group of statistically naïve participants answers all questions without and then with the aid of experience. The results of Experiment 1 demonstrate that being exposed to sequentially experienced data leads to more accurate statistical inferences and judgments that benefit from experience are preferred over analytic responses when providing final answers. Moreover, these important results apply across the range of statistical sophistication that we investigate.

The task used in Experiment 2 is inherently more complicated than those in Experiment 1. Specifically, we investigate effects of experiencing sequentially simulated outcomes on understanding the probabilistic implications of a regression model in the context of an economic

investment. Once again we compare responses of participants varying in statistical sophistication, and once again, we find that experience leads to accurate inferences for most of our participants.

Experiment 1

Design

We varied two between-subject factors in an incomplete 2 x 2 design in which all participants gave three answers to each of seven questions (thereby also allowing within-subject comparisons). One between-subject factor was level of statistical sophistication. We compared responses of advanced undergraduate students who had taken classes in statistical reasoning and probability theory with those of a group of older, university-educated adults with less formal statistical knowledge. The second between-subject factor was whether participants first answered the questions before experiencing sequentially simulated outcomes, and then again after having done so, as opposed to the reverse, that is, first after having experienced sequentially simulated outcomes, and then without having done so. This second factor, however, was incomplete in that it was only varied for the advanced undergraduate students.

The experimental design is illustrated in Table 1. As shown there, one group of advanced undergraduates first answered all seven questions without having experienced sequentially simulated outcomes. In contrast, the second group of undergraduates first answered after experiencing sequentially simulated outcomes. After each answer provided in the second task (with and without simulated outcomes, as appropriate), both groups were required to state a final answer that, if correct, would earn them € 1.00 (for each correct answer). We refer to these two groups as “Sophisticated A” and “Sophisticated B,” respectively.

The group of university-educated adults only answered the questions in one order: before and then after having experienced sequentially simulated outcomes. They also gave a final answer but were not remunerated financially for either their accuracy or participation. We refer to this group as "Naïve."

The particular questions used in the experiment were chosen because they represent a range of well-known problems that people typically answer incorrectly.

(Insert Table 1 about here)

Participants

Sixty-two undergraduate students were recruited from two classes at Universitat Pompeu Fabra and assigned at random to the Sophisticated A and Sophisticated B groups (31 to each). The students were in the 3rd/4th year of undergraduate studies in business and/or economics and had all taken courses in probability and statistics. When asked to indicate their level of comfort in probabilistic and statistical reasoning on a 5-point scale going from "does not know or remember anything" (1) to "expert and can teach others" (5) with a mid-point at "remembers some of the things and did well in related courses" (3), the mean self-reported rating for both groups was 3 (SD, 0.4). The average age of the Sophisticated groups was 22 and 52% were female. The mean remuneration participants received – for the correctness of their final answers – was € 4.51.

The Naïve group consisted of 20 university-educated adults recruited through personal contacts of one of the authors. Their mean age was 39 (range from 24 to 59) and 50% were female. In terms of statistical sophistication, the mean self-reported rating (using the same scale as the undergraduates above) was 2 (or "knows or remembers little" with a standard deviation of 0.6).

Procedure

Participants made individual appointments to meet with the experimenter and were alone with him when answering questions.[4] Experimental sessions lasted, on average, 42 minutes per participant.

When responses were made without experiencing simulated outcomes, participants simply responded to the questions in written form. Participants were free to make calculations using paper and pencil and/or use calculators.

Prior to asking a participant to provide answers with the aid of simulated experience, the experimenter first explained the concept of simulation using the example of a coin toss. After the explanation, the participant was invited to participate in and experience the simulation of a coin toss by using a mouse to click on the screen of a personal computer. Each click resulted in one simulated outcome with the result that the participant could experience the string of outcomes produced ("heads" = 0, "tails" = 1) by successive trials.[5] Any questions about simulation were then clarified by the experimenter. This exercise took approximately two minutes for each participant and is included in the time spent on the whole experiment.

When responses were made after experiencing simulated outcomes (after the coin toss example), the participants responded to each of the seven questions using simulations on a personal computer that followed the same *modus operandi* as the simulated coin toss. After the participants had read each question, the experimenter informed them what the program was

[4] For this experiment, we deemed it important to give individual attention to each participant to handle technical issues concerning the simulation technology.

[5] All the simulations used in this work were programmed in MS Excel.

doing in each simulation, e.g., for the birthday problem it was explained that each click resulted in seeing the outcome associated with a group of 25 randomly chosen individuals. Then participants started sampling by clicking with the mouse. The manner of sampling outcomes was left to each participant's discretion, i.e., number of trials, time taken per trial, and even different numbers of samples. At any time during the sampling, participants were free to stop and summarize the data they had experienced up to that point, i.e., to visualize the size of the sample created and to assess outcomes. Occasionally, participants asked questions about the simulation mechanism and the experimenter answered in a standard manner to avoiding giving cues as to "correct" responses to the probabilistic inference questions.[6]

For those participants who answered first after experiencing simulated outcomes (Sophisticated B), the simulation methodology and questions were presented first (as just explained above); once this was completed, they answered all seven questions without experiencing simulated outcomes. All participants, in all conditions, answered the different problems in individually randomized orders.

Table 2 provides details of the seven problems respondents were required to answer. Answers to the seven problems are provided in Appendix A.

(Insert Table 2 about here)

Finally, because the experiment was conducted in a multi-lingual environment, care was taken to ensure that the participants were fluent in the languages used, being Spanish (mainly), English, or Turkish.

[6] In each case, the experimenter responded saying: "The program simulates correctly the current situation/problem and provides you with an outcome each time you click."

Results of Experiment 1

Before discussing the main results, we first make a few comments about the participants' simulation experiences. First, the mean sizes of samples (i.e., number of simulated outcomes) per problem were almost the same for the two Sophisticated groups, 66 (SD, 46) and 65 (SD, 52) for A and B, respectively, but lower in the Naïve group which had a mean of 49 (SD, 37). The Sophisticated-Naïve difference is significant ($t = 3.67$, $p<.001$). Second, when simulating, nearly 80% of participants in Groups A and B took a small sample first and then gradually increased the sample size and updated their first impressions almost always toward the estimates they obtained from the larger sample. For the Naïve group this figure was 40%. Third, only four Sophisticated participants simulated multiple samples within a question. Finally, the groups differed in how long they took to answer the problems: members of Sophisticated A spent on average 19.6 minutes (SD, 4.3) on the first task of solving the problems analytically and 23.1 minutes (SD, 3.9) on the second task, whereas members of Sophisticated B spent on average 25.5 minutes (SD, 4.7) on the first task of experiencing the outcomes through simulation and 15.4 minutes (SD, 3.9) on the second task.

Table 3 provides an overview of the percentage of correct responses to the seven problems broken down by experimental conditions and groups. Figures 1 through 7 provide full information on responses made in all conditions by all groups to the seven problems. To simplify presentation, we refer to answers made without having experienced simulated outcomes by the term "Analytic." "Experience" refers to answers made after experiencing simulated outcomes.

(Insert Table 3 and Figures 1 through 7 about here)

Some general trends can be observed from Table 3. First, across all problems and groups, the percentage of correct answers after experience exceeds that of the analytic responses, and

typically by a large margin. Second, with one exception, the percentage of correct final answers lies between their experience and analytic counterparts (the exception is the Conjunction problem). This suggests that whereas participants were capable of interpreting their experience, they still wanted to give some weight to their analytic responses. Third, there are order effects. More participants in Sophisticated B (who answered after using experience first) gave accurate analytic responses than those in Sophisticated A. Experience clearly affected analytic responses. Finally, whereas the analytic responses of the Naïve group are generally less accurate than their Sophisticated counterparts, their post-experience and final answers are quite comparable. Statistical tests supporting all the above statements are provided in Appendix B. We now comment on each problem by referring to the appropriate figures.

Figure 1 reports the result of the Bayesian updating task. As in all the other figures related to Experiment 1, we display nine graphs. The three graphs in the top row report the data for the analytic responses for the three groups (from left to right, Sophisticated A, Sophisticated B, and Naïve, respectively). The middle and bottom rows show the analogous data for the experience and final responses.

The specific version of the Bayesian updating problem was taken from Gigerenzer et al. (2007). This was employed in a continuing education program in which 160 gynecologists were instructed how to use natural frequencies for solving Bayesian updating problems. The results of that session were quite successful. Whereas only 21% of the 160 gynecologists provided the correct answer before training, the percentage rose to 87% after training.

The comparison with our results can be seen by looking down the left-most column of graphs in Figure 1. Only 5 out of 29 (17%) answer correctly initially (similar to Gigerenzer's 21%). However, after experience 28 out of 29 (97%) answer correctly although this figure drops

to 23 out of 29 (79%) for the final answer. In short, our results are comparable to those achieved with the natural frequency method. Moreover, it is probably as easy (or easier) to provide the instructions for experiencing simulated outcomes than to teach the calculations required to use natural frequencies.

Figure 2 displays results for the birthday problem. Here we note that analytic responses are skewed for all three groups toward incorrect, low values. Experience makes a dramatic difference. In this specific case it is obtained through simulating binary outcomes where "1" means there are at least two people with the same birthday in a group of 25 people and "0" otherwise. Whereas the actual percentage correct is less than in other problems, the answers of a clear majority are close to correct. This pattern is mainly maintained in the final response by the Sophisticated groups but here the Naïve group exhibits a quite wide dispersion of responses, sometimes preferring a "middle" solution between the outcomes of the two tasks.

The results of the conjunction problem in Figure 3 are clear. The analytic responses are somewhat dispersed. But experience makes a big difference that is largely maintained by all groups in their final responses.

For the Linda problem – Figure 4 – we consider only whether participants recognized that the event "bank teller and active in the feminist movement" could not be more likely than "bank teller." Parenthetically, for this problem participants experienced a vector of "1"'s and "0"'s for each of the outcomes simulated. These numbers indicated whether each simulated Linda character did or did not have the attributes that were to be ordered by probability.[7] The analytic

[7] The text of our problem refers to Jessica as opposed to Linda to avoid the possibility that the Sophisticated participants might have heard of the "Linda problem."

and experience–based responses are generally opposites for all groups (incorrect and correct, respectively). The majority responses for the final answer, however, are correct.

In Figure 5, experience leads to almost 100% correct responses for the hospital problem and the majority of final answers are also correct. For this problem there is a striking order effect. In the Sophisticated B group, there is a majority of correct analytic responses. In this case, prior experience was probably particularly relevant because no calculations were needed to answer the analytic question.

Figure 6 reports the results of the regression toward the mean problem. The modal responses of all groups to the analytic question are centered on the incorrect answer of "equal," thereby suggesting that the respondents did not understand the principle behind the question. The effect of experience is to shift answers to being more correct. However, at the final stage the Naïve group is not convinced.

Experience has a big impact for the Monty Hall problem – Figure 7. Almost everybody chooses the correct answer of "change" after experience. However, a minority regress to the incorrect answer at the final stage.

Discussion of Experiment 1

The stimuli in Experiment 1 were chosen precisely because previous research has shown that responses to their presentation in a standard probabilistic format typically imply incorrect inferences. And yet, when we presented the problems to people in a form that allowed them to experience sequentially simulated outcomes, responses for all questions were remarkably accurate. To this we add three points.

First, training people to participate in the simulations by using the coin toss example was quite easy and took little time, on average some 2 minutes per person. Participants related easily to the task of experiencing the outcomes of simulations.

Second, despite the fact that our participants varied on levels of statistical sophistication, the accuracy of all participants' responses benefited from experience.

Third, we allowed our participants to choose third, and final, answers thereby requiring them to express either a preference for answers achieved with/without the aid of simulated experience or some combination of the two. Whereas some participants did revert toward answers made without experience, a large majority gave more weight to those achieved through experience.

It is important to emphasize that we did not give participants any indications as to how large their samples of experienced outcomes should be. What we found was that the Sophisticated participants sampled more than the Naïve and some problems involved systematically more sampling than others. For example, the Bayesian and birthday problems both involved the largest numbers of trials (means of approximately 80 to 90 for the Sophisticated groups) whereas the Linda problem stimulated far fewer trials (around 30 for all groups). However, in this problem, participants had to simulate multiple outcomes for each individual sampled thereby experiencing vectors of "1's" or "0's" and not just single "1's" or "0's." Thus, the task was more cognitively taxing.

An interesting benchmark for the amount of sampling undertaken by our participants is the behavior observed by Hertwig, Barron, Weber, and Erev (2004) in a paradigm where participants learned the features of two alternative choice options by active sampling of experience (in a manner quite similar to ours, i.e., by clicking a key on a personal computer). In

Hertwig et al.'s study, the median number of observations sampled was 15, far less than the medians we observed of 52, 51, and 30 (for Sophisticated A and B, and Naïve, respectively). The reason why our samples are bigger is unclear although it is interesting that Lejarraga (2010) – using the same paradigm as Hertwig et al. (2004) – found that more analytically oriented participants sampled more than the less analytical, a result that parallels our finding that the Sophisticated groups experienced larger samples than the Naïve. [8]

Clearly, there are some normative principles that participants should follow in determining sample size. For example, if there are relatively few "1's" or "0's," the distribution may well be skewed in which case a larger sample should be experienced than if the number of "1's" or "0's" is more equal. In our data, there are no hints of such awareness.

The seven inferential problems we chose to use as stimuli were selected for two reasons. The first (noted already above) was that we wanted to test our ideas on problems that were well-known so that we could better assess improvements in the quality of statistical inferences achieved after participants had been exposed to experience. The second reason was that if our suggested "method" were to work well across a range of situations as opposed to within variations of the same problem (e.g., different Bayesian updating tasks), it would provide a stronger test of its efficacy. Indeed, as was noted, the methodology was successful across a range of problems.

As noted before, Participants in B often transformed their calculations to obtain the result they had experienced in the simulation, using this as a cue to the answer. This suggests that

[8] Lejarraga (2010) compared his participants using Pacini and Epstein's (1999) scales of rational ability and engagement.

simulated experience can play an important role in providing insights to improve the quality of analytical thinking.

However, despite these outcomes, one might still argue that in many cases an alternative method such as summarizing natural frequencies should still be preferred because it is simpler to implement. (For example, you don't have to construct a simulation model.) We therefore sought to examine the efficacy of experiencing simulated outcomes in a situation where it is less obvious how an alternative presentation could be achieved using a natural frequency presentation format. Experiment 2 was designed to do just this.

Experiment 2

Design

The design of Experiment 2 involved between-subject comparisons of two groups that were required to answer questions based either on the analytical description of a problem that used regression analysis or after gaining experience with a simulation tool. We label the groups as Analytic and Experience, respectively, except that there were two subgroups in the Experience condition. One involved statistically sophisticated, graduate students in economics whom we label Sophisticated, and who were similar to respondents in the Analytic group. The other was comprised of university-educated adults without advanced statistical knowledge whom we refer to as Naïve.[9] We therefore make comparisons between three subgroups: Analytic, Sophisticated, and Naïve.

(Insert Figures 8 and 9 about here)

[9] Specifically, these participants did not know what "regression analysis" is.

The problem set-up and method

Figure 8 provides the wording of the problem set-up for participants in the Analytic condition. As can be seen, the problem involves an investment situation, which requires allocating funds (40 credits) across three alternatives: "Investment 1", "Investment 2", and "no investment." The profitability of the two investment opportunities are described by a regression model. The specific questions were:

1. How would you allocate your 40 credits in order to expect an increase of 5 credits (obtain 45 credits)? How much of 40 credits in Investment 1, how much in Investment 2, how much in N (no-investment)?
2. Given your investment decision in (1), what would you say is the probability of your obtaining a final total credit amount that is below 40 (Y<40), i.e. less than what you have started with?
3. Given your investment decision in (1), what would you say is the probability of your obtaining a final total credit amount that is below 45 (Y<45)?
4. Given your investment decision in (1), what would you say is the probability that you will get a larger outcome with respect to a person who does not invest in Investments 1 and 2 (someone with N=40)?

The statistical rationales for the answers are provided in Appendix C.

Figure 9 depicts the simulation interface for the Experience group. When conducting the experiment, we sat down one-by-one with the participants in this group, explained briefly how the tool works, and then asked them to choose an investment plan so that they can expect to increase their 40 credits to 45 (the same as question 1 above). We allowed them to experience as many choice options as they wished. Once they made their decisions, we asked them to answer

questions 2, 3, and 4 above. Once again, we allowed them to utilize the simulation tool and they could experience the outcomes of their choices as many times as they desired. Moreover, we made sure that the participants could see all their choices and outcome histories and even calculate and compare averages of their past outcomes.

Participants

The Analytic group consisted of 26 graduate students in economics at Universitat Pompeu Fabra in Barcelona who had at least completed their first and second semesters. This ensures that all of them had taken at least one graduate course in econometrics and were knowledgeable about linear regression analysis and its interpretation. They did not have a time limit. Participation was voluntary and anonymous. Participants could use any tools they wanted and, upon completion of the survey, they slid the questionnaire into a sealed box in front of an office. The average age of this group was 25 and 30% were female. Of 35 surveys distributed, 26 were completed.

The Sophisticated participants within the Experience group consisted of 28 graduate students in economics drawn from the same population as the Analytic group. The Naïve participants were 18 members of the general public having university degrees but no knowledge of regression analysis. They were recruited from the contacts of one of the authors. Their mean age was 35 (range from 23 to 60) and 40% were female.

Before participating in the experiment, a chocolate bar was donated to each of the participants.

Results of Experiment 2

Table 4 documents the means (standard deviations) of different variables – the decisions taken, and answers to the required probabilistic inferences – for the different experimental conditions. The first two rows of the table (labeled I1 and I2) indicate the mean amounts invested in Investments 1 and 2, respectively, by the different subgroups. According to the regression results, these two investments differ in the expected level and variability of their returns – Investment 1 having both greater expected return and more variability than Investment 2. On average, therefore, it can be observed that the Analytic participants adopt less risky strategies than their Sophisticated counterparts but that all three subgroups select investment strategies that essentially meet the demands of the first question, i.e., to achieve an expected target of 45.

(Insert Table 4 about here)

Question 2 asks for the probability that the investment strategies will lead to outcomes of less than 40 (i.e., the amounts participants started with). The accuracy of each participant's response can be assessed by calculating the difference between the response itself and its normative counterpart (i.e., the correct response implied by the regression analysis). Using this measure, we note that whereas the Analytic group seriously underestimates the probability that Y is less than 40 (the average deviation from the correct answer is -16% with standard deviation 9%) this is not the case for the Experience group: 0% for Sophisticated (SD, 11%) and 5% for Naïve (SD, 9%). The difference between the Analytic and Experience conditions is significant ($t = 7.1$, $p < 0.001$).

Question 3 asks for the probability that the investment strategies will lead to outcomes of less than 45 (i.e., the investment target). On average, answers to this question are all quite accurate. In fact, these responses are consistent with answers to the first question that lead to

expectations of, on average, about 45, that is, with a symmetric predictive distribution there is as much chance of exceeding as falling short of the target.

Question 4 asks for the probability that the chosen investment strategy will lead to outcomes superior to a strategy of no investment. The Analytic group overestimates this probability (the average deviation from the correct answer is 24% with standard deviation 10%) while this is again not the case for the two subgroups in the Experience condition: 1% for Sophisticated (SD, 16%) and 3% for Naïve (SD, 8%). Again, the difference between the Analytic and Experience conditions is significant ($t = 7.7$, $p < 0.001$).[10]

Discussion of Experiment 2

Unlike the specific probabilistic inference tasks of Experiment 1, Experiment 2 required participants to choose an investment plan and make probabilistic inferences based on their own idiosyncratic decisions. Also unlike several tasks of Experiment 1, it is unclear how one could have provided alternative representations of the questions asked in the form of natural frequencies. However, like the representations of all tasks in Experiment 1, participants in the Experience group experienced data in the form of sequentially generated outcomes.

Experiment 2 only permitted between-subject comparisons. In brief, we found – holding analytical ability constant – that Sophisticated participants gave more accurate probabilistic inferences when allowed to experience simulated outcomes than those who were required to solve problems analytically. Second, there was little or no difference in accuracy of probabilistic inferences between the groups of Sophisticated and Naïve participants who experienced

[10] For questions 3 and 4, there are no significant differences between the (response-correct) measures of the two subgroups in the Experience condition.

simulated outcomes. These results are important. They suggest that the ability to encode frequencies in the form of sequentially experienced frequency data can be harnessed to improve probabilistic inferences across a wide range of tasks.

We note also that the questions posed in Experiment 2 are important for decision makers considering investment plans. In such situations, individuals would primarily base their decisions on the chances of being worse off with respect to their starting point, to their goal and to other individuals who do not make these particular investments. In a recent survey (Soyer & Hogarth, 2010); we posed a simpler (univariate) version of this problem to economic scholars from prestigious universities. These respondents made the same kinds of mistake as the Analytic group in Experiment 2.

Finally, we note that for both the Experience subgroups, we collected data on numbers of simulations for each choice. Before deciding on a final investment plan, the Sophisticated simulated an average of 7 different strategies some 19 times each. The Naïve simulated an average of 5 strategies about 8 times each. Thus, as in Experiment 1, we find that more statistically sophisticated participants choose to experience more outcomes than the less sophisticated.

General Discussion

The main theoretical concept underlying our work is simple. People can successfully perform complex intellectual tasks if these are presented in a format that exploits their natural abilities for processing information. In Experiment 1, we investigated seven probabilistic inference problems that have a long history of eliciting erroneous responses. The human ability we identified was the

capacity to encode the outcomes of sequentially generated outcomes experienced across time. Thus, when we presented problems in a format that allowed participants to use this natural ability, we observed vastly more accurate probabilistic inferences than those elicited after presentation of the standard probabilistic format. Moreover, this result held for both between- and within-subject comparisons and across participants varying in statistical sophistication. In Experiment 2, we obtained similar results using what are arguably more complex problems involving inferences from a regression equation modeling investment decisions. Taken together, our work suggests a strategy for a general approach to help people make appropriate probabilistic inferences.

It is important to stress that our work builds on the illuminating contribution of Gigerenzer and his colleagues (notably Gigerenzer & Hoffrage, 1995) who showed how the use of natural frequencies – as opposed to probabilities – leads to simpler and more accurate calculations in probabilistic inference (notably for Bayesian updating). We reasoned, however, that Gigerenzer and his colleagues did not take their own arguments about human abilities to handle frequency data to their logical conclusion. Instead of presenting people with problems framed in terms of aggregated frequencies (that still require some calculations), we advocate letting people experience the raw data as generated from the underlying process or, if not possible, from a simulation model. Indeed, this is essentially the same technique that is used to provide non-human animals with information in investigations of their reasoning skills except, of course, that the animals do not typically intervene and determine the number of trials. Moreover, the animals are seen to be quite skilled (Weber, Shafir, & Blais, 2004).

At one level, our work can be viewed through the perspective that has recently been popularized by the expression “choice architecture” (Thaler & Sunstein, 2008). This is a

recognition that, since the variety of tasks the human mind confronts is much larger than the variety of responses that humans can make, much can be gained by designing tasks in ways that allow humans to make appropriate choices. Of course, this is not a new principle. In psychology it can be traced to the work of Brunswik (1952) and was further elaborated by Simon (1978) and Tversky and Kahneman (1981) (see also Hogarth, 1982). However, it is one thing to elaborate such a principle at a general level; it is quite another to demonstrate how it works in specific situations and to define boundary conditions.

A number of questions can be raised about the boundary conditions of our proposal. We consider five issues: (1) How much and what kind of experience do people need to make appropriate responses? (2) Do people trust simulation mechanisms? Why or why not? (3) How general is the simulation technique or, in other words, can models be easily constructed for all types of situations? (4) How does experience in the form of simulated outcomes solve the problem of understanding probabilities of unique events? (5) How does simulated experience relate to the distinction sometimes made between intuitive and analytic processes? We now consider each of these questions.

(1) In our experiments, we deliberately let participants determine the amount of information – in terms of number of trials – that they wanted to experience. This procedure raises two issues. First, how much experience – that is number of trials – do people need to reach conclusions with which they feel comfortable? Second, does being actively involved in the sampling process make a difference compared to simply observing outcomes?

Our data did show a relation between statistical sophistication and sample size with the more sophisticated requiring larger samples. Thus, we suspect that individual differences could play a role in the answers to both questions. We also believe that the two questions are important

and demand further research. For example, it would be relatively easy to conduct experiments varying both sample size and active intervention in as opposed to passive observation of the sampling process and to elicit not only probabilistic inferences but measures of confidence in such assessments. We suspect that active participation is an important factor – possibly interacting with sample size – but reserve judgment in that, in animal studies, the organisms typically do not intervene in the sampling process.

(2) The degree of transparency of the sampling mechanism is clearly important. This may have several dimensions. One is the level of the participant's familiarity with the data generating process. For instance, it is probably easy for the participant to understand the coin toss example with which we introduced the simulation technique in Experiment 1, and particularly since the evidence would typically confirm prior beliefs that there should be roughly as many "heads" as "tails." On the other hand, simulating birthdates of different groups of 25 people in the birthday problem might seem odd as well as the fact that the experiential evidence typically runs counter to prior intuitions. At the same time, when people have little insight into the structure of a problem – as occurs in both the hospital and Monty Hall problems – living the experience of many outcomes can be quite illuminating.

However, if the participant already understands the structure of the problem – as happens in the conjunction problem – and recognizes that her capacity for calculation is deficient, she might welcome the simulation tool. In fact, Lejarraga (2010) essentially tested this hypothesis by letting people decide whether they wanted to choose between gambles based on description (i.e., where probabilities of different branches leading to outcomes were indicated) or experience (after simulating outcomes). The same pairs of gambles were presented to three different groups of participants but varied in the complexity (number of branches) used to describe them. As

problem complexity increased, groups displayed a greater tendency to make their choices after experiencing outcomes as opposed to trusting their analytical abilities to figure out the implications of the presentation by description.

Finally, it is easy to dismiss simulated experience as simply being the outcome of a "black box." However, we believe a more appropriate metaphor is that of a "grey box" where individuals experience outcomes generated by a computer as opposed to those arising from the naturally occurring environment. But much research is needed to determine what affects the different shades of grey and thus the conditions under which people do or do not feel comfortable in relying on outcomes of simulated experience.

(3) Our third question centers on limits to the generality of the simulation technique itself. At a conceptual level, and given sufficient ingenuity on the part of the investigator, there is almost no technical limit to the probabilistic situations that can be constructed. Whether they are meaningful, however, is another issue that can be viewed from two perspectives: the reality being modeled and the experience of the user. For the latter, the critical issue is that already discussed above, namely the shade of grey of the box. For the former, it should be clear that the models are only as good as the fit of their assumptions to reality. As we see it, the goal of simulated experience is not necessarily to reach a precise probabilistic answer to a problem but more a means of gaining insight into effects of assumptions made about the structure of the problem as well as reaching an *approximate* answer. Thus, it would be illuminating to employ techniques of sensitivity analysis and to experience, say in a Bayesian updating task, how different assumptions concerning prior probabilities or base-rates result in different sequences of outcomes.

(4) Our fourth issue speaks to the meaning of probability. The main distinction is whether the concept is something that applies to unique events (e.g., the probability that a particular person has a certain disease) or classes of events (e.g., that people that belong to a particular group have the disease). This distinction has been given different names in the literature, for example, *epistemic* as opposed to *aleatory*, or singular versus distributional (Reeves & Lockhart, 1993). Although from the subjectivist or Bayesian perspective a probability simply measures a degree of belief such that the distinction is irrelevant, there is much evidence that people's intuitions of the probability concept are more clearly aligned with the distributional perspective (see Gigerenzer & Hoffrage, 1995). For example, people relate more easily to a statement that a fair coin tossed 100 times is expected to show heads roughly 50% of the time than the statement that the probability of heads on a single toss of the coin is 0.5. For the former, there is some informational "certainty" in the 50%. For the latter, 0.5 is a statement of total uncertainty. The experience of simulated outcomes clearly taps into people's distributional intuitions about the meaning of probability and this, in part, may explain why they find it illuminating.

(5) If experience is so powerful, why, it can be asked, did our participants not all state that their final answers were the same as those reached after experiencing simulated outcomes? Indeed, by failing to do so, participants in the Sophisticated group in Experiment 1 actually lost money. One reason has already been alluded to above, namely, participants may not have always trusted experience in the form of simulated outcomes. Another and related reason could be what might be called a clash of intuitions.

Although we referred to answers given by participants without experience in Experiment 1 as being "analytic," it should be clear that many of these responses were driven by intuitive reactions. Indeed, the problems are interesting precisely because past studies have shown that

people's intuitive reactions are typically contrary to analytic principles. The Linda problem is a prime example. When unaided by formal analysis or simulated experience, people are strongly drawn by intuition to believe that it is more likely that Linda is "a bank teller and active in the feminist movement" than that she is "a bank teller." Now, reactions to experience in the form of sequential frequency data could also be classified as intuitive (Hogarth, 2001). Thus, in many cases, our respondents faced a conflict between two intuitions, one being their reactions to the analytic presentation format, the other being their feelings after the experience they sampled. As the evidence shows, the latter form of intuition did not always overcome the former.

That there should be a link between sequentially encoded frequency information and intuition has been emphasized by Sedlmeier (2005; 2007) who, in addition, modeled this as a process of associative learning (Sedlmeier, 1999). Moreover, in exploring different ways to train people to reason probabilistically, his 1999 "flexible urn" concept is perhaps closest to our suggestions in that it involves both perceiving simulated data dynamically and some active involvement with a computer interface. However, most of his work – and suggestions – have focused on different ways of presenting information in the form of *aggregate* natural frequencies as opposed to *sequentially* observed frequency data (see, e.g., Sedlmeier, 2000; Sedlmeier & Gigerenzer, 2001).

Our work also speaks to the issue of whether and when to trust intuition or analysis in making a judgment (Hogarth, 2001; 2005; Kahneman & Klein, 2009). If we classify the analytic responses as being "analysis and intuition" and the experience judgments as "intuition," it is clear that intuition alone is better in that the latter produced the highest proportions of correct responses. However, it would be erroneous to draw any general conclusions from our study largely because our stimuli in Experiment 1 were specifically chosen for their history of

inappropriate responses. What we have shown is that intuitive processes based on experiencing simulated frequency data result in quite accurate probabilistic inferences across a range of problems (i.e., in both Experiments 1 and 2).

Our investigation raises an important practical issue: What advice might one give, say, to a physician who should be using Bayesian updating to assess how likely a patient is to have a specific disease following a positive test result? Should you just give her the correct Bayesian answer? The answer is probably no because unless she fully understands how the number is calculated, she is unlikely to believe it.

The classic advice would be to teach the physician Bayes' theorem but, unless this is replicated on many occasions, it is unlikely that she will be able to reproduce accurate answers in the future. A better approach would be instruction using the natural frequency approach but, once again, how well this would be recalled on future occasions is unclear. (However, see Sedlmeier & Gigerenzer, 2001.) We believe that simulated experience provides a level of understanding that would help the physician understand why the standard Bayesian and natural frequency approaches are correct. In fact, our position is to advocate using simulated experience as a means to reinforce understanding the natural frequency approach. In this way, the physician could reach conclusions that do not involve any conflict between intuition (based on experiencing simulated outcomes) and analysis (based on natural frequency calculations). In time, this would allow the physician to use the natural frequency approach directly (and particularly if there is no available simulation technology). As an additional point, we see potential in the idea that simulated experience could provide a useful way of communicating statistical information. For example, physicians might use simulated experience to provide patients with a better understanding of the probabilities of different outcomes. That is, letting patients experience simulated outcomes based

on analyses and past data could lead to more accurate calibration of expectations and consequently better decisions.

In related work, we have shown that knowledgeable economists have difficulties in making correct inferences given the standard presentation modes in the economics literature (Soyer & Hogarth, 2010). In this case, economists' inferences are blind to different levels of uncertainty as they tend to rely disproportionately on the statistical significance of regression coefficients. The results of Experiment 2 provide insight into how simulated experience might be used to aid decision makers in interpreting statistical outcomes. It could help in taking different levels of risk into account and in identifying variables that are not only statistically significant but economically important. The distinction between statistical significance and economic importance is discussed in Ziliak and McCloskey (2008).

These last points speak to the importance of using simulated experience for teaching probability and statistics at all levels – from grade school through university and beyond. Nowadays, it is relatively simple to build simulation models for all kinds of applications and problems and with the widespread availability of personal computers – linked by the internet – there is no reason why the simple idea championed in this paper could not have wide application. Indeed, the Statistics Online Computational Resource (SOCR) website – www.socr.ucla.edu – provides a repository of elegant simulations and applets for many probabilistic problems, including several featured in Experiment 1. Moreover, Dinov, Sanchez, and Christou (2008) have shown that using the website while teaching statistics enhances students' understanding and retention of concepts.

One could also envisage a computer tool in the form of an expert system that could aid people with little statistical sophistication to build their own simulation models and thereby gain

insight into a variety of inferential problems. Indeed, one could even imagine such programs being developed for cell phones such that they could be almost as common as calculators.

At the head of this article is a quote from Carl Rogers (1961). At first sight, this might appear odd for research developed from a cognitive view of psychology. However, what Rogers was emphasizing – and where we concur – is that the understanding that really changes behavior is that which comes through self-directed and experienced learning. For this and other reasons already enumerated, we maintain that simulated experience can be an effective route to gain insight into the nature of probabilistic reasoning and thereby guide behavior to meet the demands of today's technological society.

References

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance, 9,* 396-406.

Betsch, T., Biel, G.M., Eddelbüttel, C., & Mock, A. (1998). Natural sampling and base-rate neglect. *European Journal of Social Psychology, 28,* 269-273.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review, 15,* 284-289.

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago, IL: The University of Chicago Press.

Christensen-Szalanski, J. J. J. & Beach, L.R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29,* 270-278.

Cohen, J. (1960). *Chance, skill, and luck.* Harmondsworth, Middlesex, England: Penguin Books.

Cohen, J. (1972). *Psychological probability: Or the art of doubt.* London, England: George Allen and Unwin.

Cohen, J. & Chesnik, E. I. (1970). The doctrine of psychological chances. *British Journal of Psychology, 61,* 323-324.

Cohen, J., Chesnik, E. I. & Haran, D. (1971). Evaluation of compound probabilities in sequential choice. *Nature, 232,* 414-416.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Dinov, I. D., Sanchez, J., Christou, N. (2008). Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses. *Computers and Education, 50*, 284-300.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York, NY: Wiley.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*, 123-129.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107(4)*, 659-676.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129 (3)*, 399-418.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review*, *102*, 684-704.

Gigerenzer, G., Gaissmaier, W., Kurz-Milke, E., Schwartz, L. M., & Wolosihn, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8 (2)*, 53-96.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology*, *38*, 48-78.

Hasher, L., & Zacks, R. T. (1979), Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*, 356-358.

Hasher, L., & Zacks, R. T. (1984), Automatic processing of fundamental information: The case of frequency occurrence. *American Psychologist, 39,* 1372-1388.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15 (8),* 534-539.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275-305.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73,* 538-540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition,* 84, 343-352.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290,* 2261-2262.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association, 70,* 271-289.

Hogarth, R. M. (Ed.) (1982). *Question framing and response consistency: New directions for methodology of social and behavioral science,* No. 11. San Francisco, CA: Jossey-Bass.

Hogarth, R. M. (1987). *Judgement and choice: The psychology of decision*. (2nd ed.). Chichester, England: John Wiley & Sons.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL: The University of Chicago Press.

Hogarth, R. M. (2005). Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. In T. Betsch & S. Haberstroh (eds.), *The routines of decision making* (pp. 67-82). Mahwah, NJ: Erlbaum.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116 (4),* 856-874.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64 (6),* 515-526.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237-251.

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

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132 (1),* 3-22.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology*, *73*, 498-502.

Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making*, *23*, 100-116.

Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*, 417-424.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*. *90(4),* 339-363.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68 (1),* 29-46.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*, 980-986.

Real, L. A. (1996). Paradox, performance, and the architecture of decision-making in animals. *American Zoologist*, *36*, 518-529.

Reeves, T., & Lockhart, R. S. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General*, *122 (2)*, 207-226.

Rogers, C. R. (1961). *On becoming a person*. Boston, MA: Houghton Mifflin Company.

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

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science*, *28*, 227-262.

Sedlmeier, P. (2005). From associations to intuitive judgment and decision making: Implicitly learning from experience. In T. Betsch & S. Haberstroh (eds.), *The routines of decision making* (pp. 83-99). Mahwah, NJ: Lawrence Erlbaum Associates.

Sedlmeier, P. (2007). Statistical reasoning: Valid intuitions put to use. In M. C. Lovett & P. Shah (eds.), *Thinking with data* (pp. 389-419). Mahwah, NJ: Lawrence Erlbaum Associates.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130 (3)*, 380-400.

Simon, H. A. (1978). Rationality as process and product of thought. *American Economic Review, 68(2),* 1–16.

Soyer, E., & Hogarth R. M. (2010). *Econometrics and decision making: Effects of presentation mode***.** Barcelona: Universitat Pompeu Fabra working paper number 1204.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

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

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293-315.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in human and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111 (2),* 430-445.

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

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor, MI: University of Michigan Press.

Author Note

Robin M. Hogarth, ICREA and Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, Spain; Emre Soyer, Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, Spain.

The authors would like to thank seminar attendants at Universitat Pompeu Fabra for their illuminating comments and questions as well as participants at the "Second One-Day Workshop on Intuition" held in Bonn (May, 2010). The research has been supported by the Spanish Ministerio de Ciencia y Innovación, grant numbers SEJ2006-14098 and EC02009-09834.

Correspondence concerning this article should be addressed to: Robin M. Hogarth, Department of Economics & Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain or to the e-mail address: robin.hogarth@upf.edu

Table 1. Design for Experiment 1

Group	**1st Task**		**2nd Task**		**3rd Task***	**Remuneration**
Sophisticated A	Answer Analytically		*Coin toss example*	Answer with Experience	Final Answer	1 Euro / Correct Final Answer
Sophisticated B	*Coin toss example*	Answer with Experience	Answer Analytically		Final Answer	1 Euro / Correct Final Answer
Naïve	Answer Analytically		*Coin toss example*	Answer with Experience	Final Answer	None

(*) Final answers were given to each problem right after the 2nd task for that problem was completed.

Table 2. The seven probabilistic inference problems

1. Bayesian updating

Assume you conduct breast cancer screening using mammography in a certain region. You know the following information about the women in this region:

The probability that a woman has breast cancer is 1% (prevalence)
If a woman has breast cancer, the probability that she tests positive is 90% (sensitivity)
If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9% (false-positive rate)

A woman – chosen at random – gets breast screening and the test results show that she has cancer. What is the probability that she has cancer?

a) The probability that she has breast cancer is about 81%.
b) Out of 10 women with a positive mammogram, about 9 have breast cancer.
c) Out of 10 women with a positive mammogram, about 1 has breast cancer.
d) The probability that she has breast cancer is about 1%.

2. Birthday problem

In a group that has 25 people in it, what is the probability that 2 or more people have the same birthday?

3. Conjunction problem

A project has 7 parts. The success of the project depends on the success of these parts. In order to be successful, all its parts need to be successful.

Assume that each part is independent from the others and each has a 75% success rate.

What is the probability that the project will be successful?

Table 2. Cont'd

4. Linda problem (Tversky & Kahneman, 1983)

Jessica is 31 years old, single, candid, and very promising. She graduated in philosophy. As a student, she was anxious about discrimination issues and social justice, and also took part in anti-nuclear demonstrations.

Assign a rank to the following statements from most probable to least probable:

a) Jessica works in a bookstore and takes Yoga classes.
b) Jessica is active in the feminist movement.
c) Jessica is a psychiatric social worker.
d) Jessica is a member of the League of Women Voters.
e) Jessica is a bank teller.
f) Jessica is an insurance salesperson.
g) Jessica is a bank teller and is active in the feminist movement.

5. The hospital problem (Tversky & Kahneman, 1974)

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day. In the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are girls. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were girls.

Which hospital do you think recorded more such days?

a) the larger hospital?
b) the smaller hospital?
c) about the same for both hospitals?

Table 2. Cont'd

6. Regression toward the mean

A class of students enters in a TOEFL exam (it is a standardized test of English language). One of the students gets a better result than 90% of the class.

The same class, including the person who had done better than 90% of his class, enters another TOEFL exam. Past data suggest that the correlation between the scores of the different exams is about 0.8.

Which statement is correct?

a) It is more likely that the student in question now gets a better ranking.
b) It is more likely that the student in question now gets a worse ranking.
c) The chances that he gets a better ranking or a worse one are approximately equal.

7. Monty Hall problem

There are three doors A, B and C. We randomly selected one of them and put a Ferrari behind it. Behind the remaining two doors there is nothing.

You will select a door and we will open it. You will win the game if there is Ferrari behind it.

Now select a door. (The participant makes a selection, say A).

Before we open the door you selected, we open B and show you that there is nothing behind it. Now two doors remain: A and C. Behind one of them is a Ferrari. Given this situation, please state if you would like to

a) Stay with your original selection
b) Change to the other door

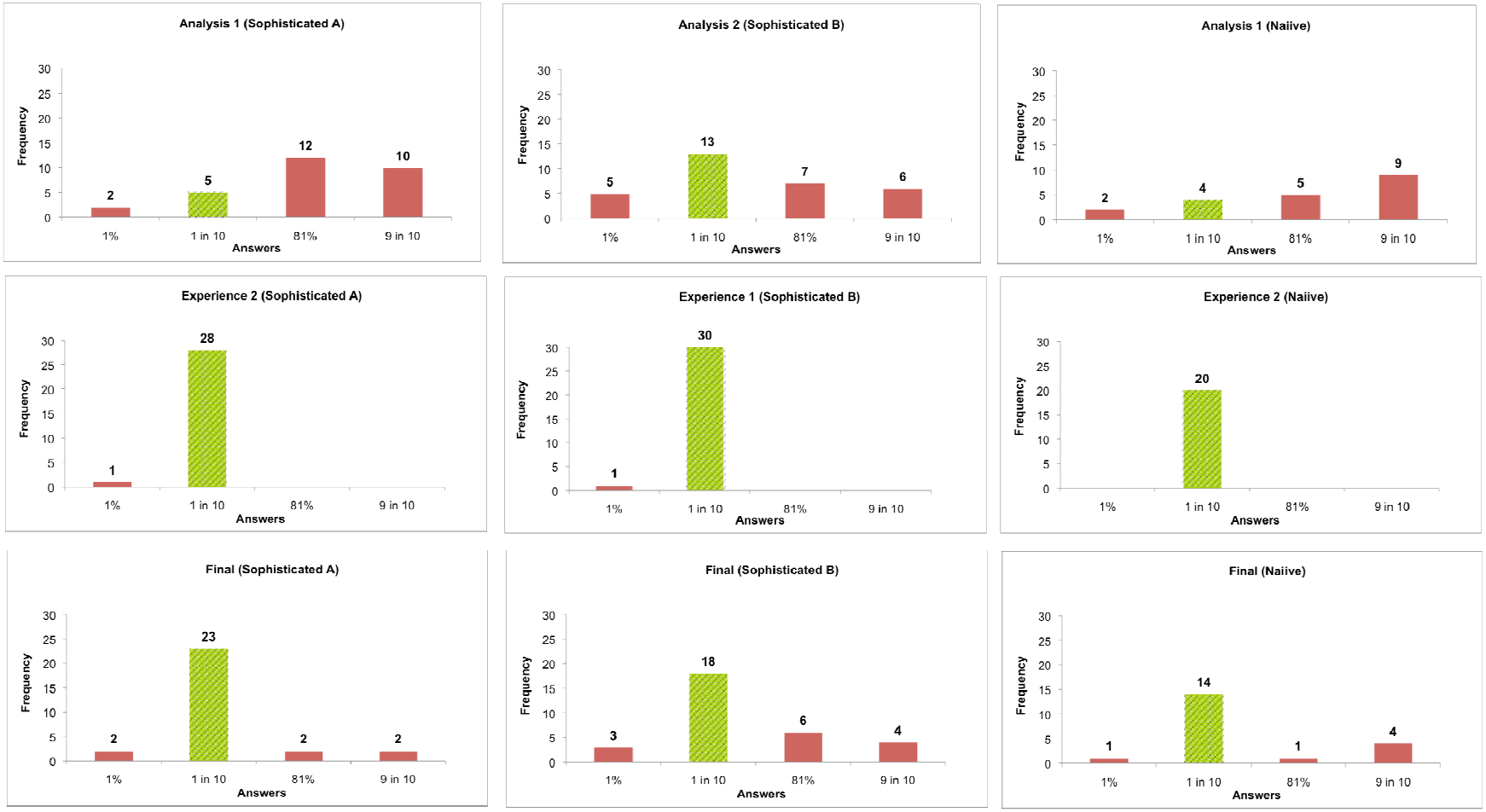
Table 3. Percentages of correct answers to inferential problems by experimental conditions

	Sophisticated		Naive	Mean
	A	B		
1. Bayesian updating				
Analytic	17	42	20	27
Experience	97	97	100	98
Final	79	58	70	69
2. Birthday problem				
Analytic	3	13	0	6
Experience	55	61	65	60
Final	35	61	30	44
3. Conjunction problem				
Analytic	55	52	25	47
Experience	74	77	75	75
Final	77	77	75	77
4. Linda problem				
Analytic	10	32	10	18
Experience	97	97	90	95
Final	65	71	60	66
5. Hospital problem				
Analytic	39	61	25	44
Experience	97	97	100	98
Final	81	68	65	72
6. Regression toward mean				
Analytic	32	45	25	35
Experience	68	90	70	77
Final	55	65	35	54
7. Monty Hall				
Analytic	31	48	15	34
Experience	93	97	95	95
Final	69	58	55	61
n =	31 (29)	31	20	

Table 4. Means <std. devs> for conditions in Experiment 2

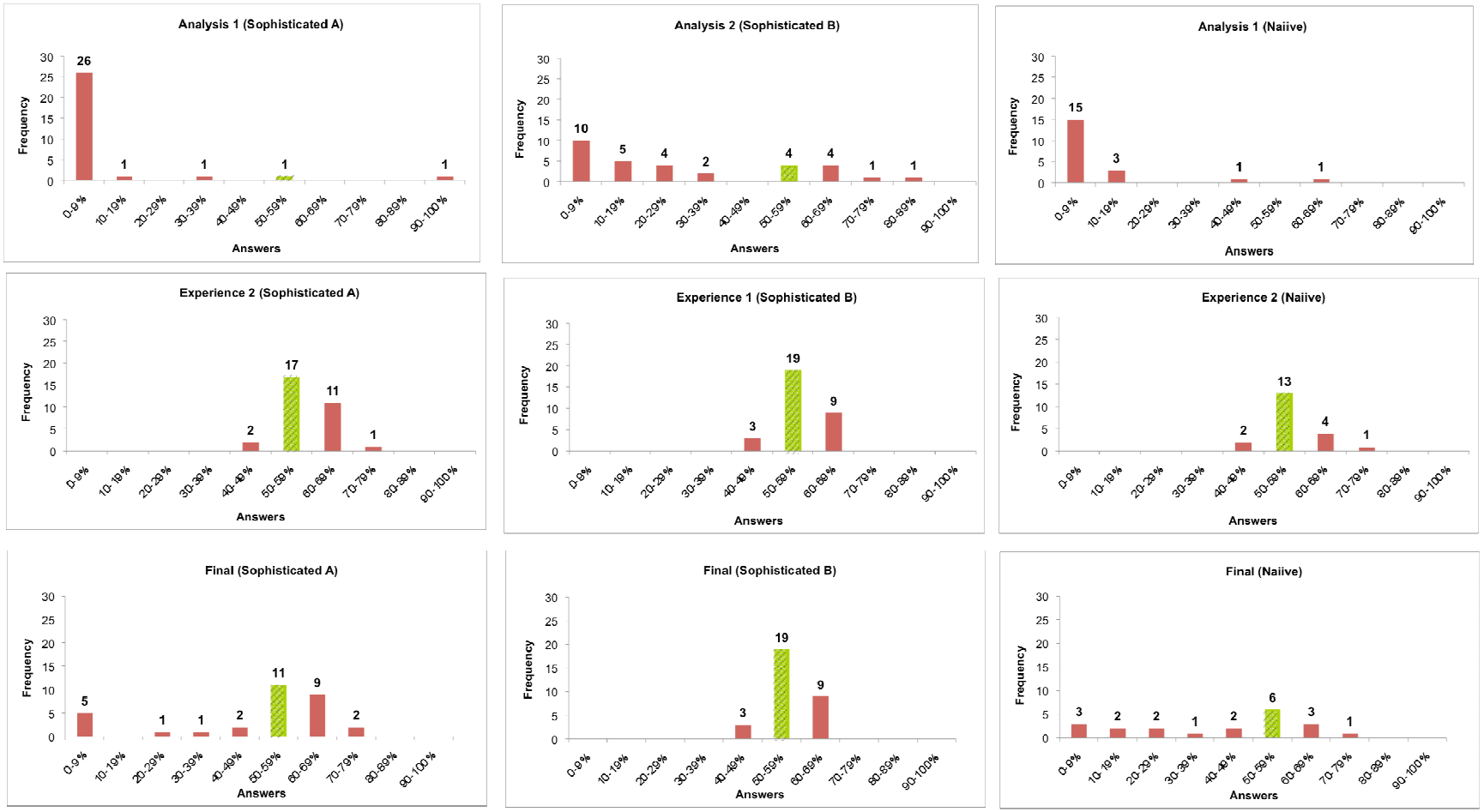
Condition:	Analytic	Experience	
		Sophisticated	Naive
(n=	26	28	18)
Decisions			
I1	**3.5**	**5.7**	**6.7**
	<4.6>	<3.9>	<5.9>
I2	**12.3**	**7.8**	**9.8**
	<7.0>	<4.5>	<7.7>
Expected outcome			
Y	**45.5**	**45.2**	**46.3**
	<0.9>	<1.6>	<2.1>
Prob (Y<40)			
Question 2: Response - *Correct*	**-16%**	**0%**	**5%**
	<9%>	<11%>	<9%>
Prob (Y<45)			
Question 3: Response - *Correct*	**2%**	**1%**	**6%**
	<2%>	<12%>	<9%>
Prob($Y\|I_1,I_2$) > Prob(Y|no investment)			
Question 4: Response - *Correct*	**24%**	**1%**	**3%**
	<10%>	<16%>	<8%>

Figure 1. Histograms of answers given to the Bayesian updating problem



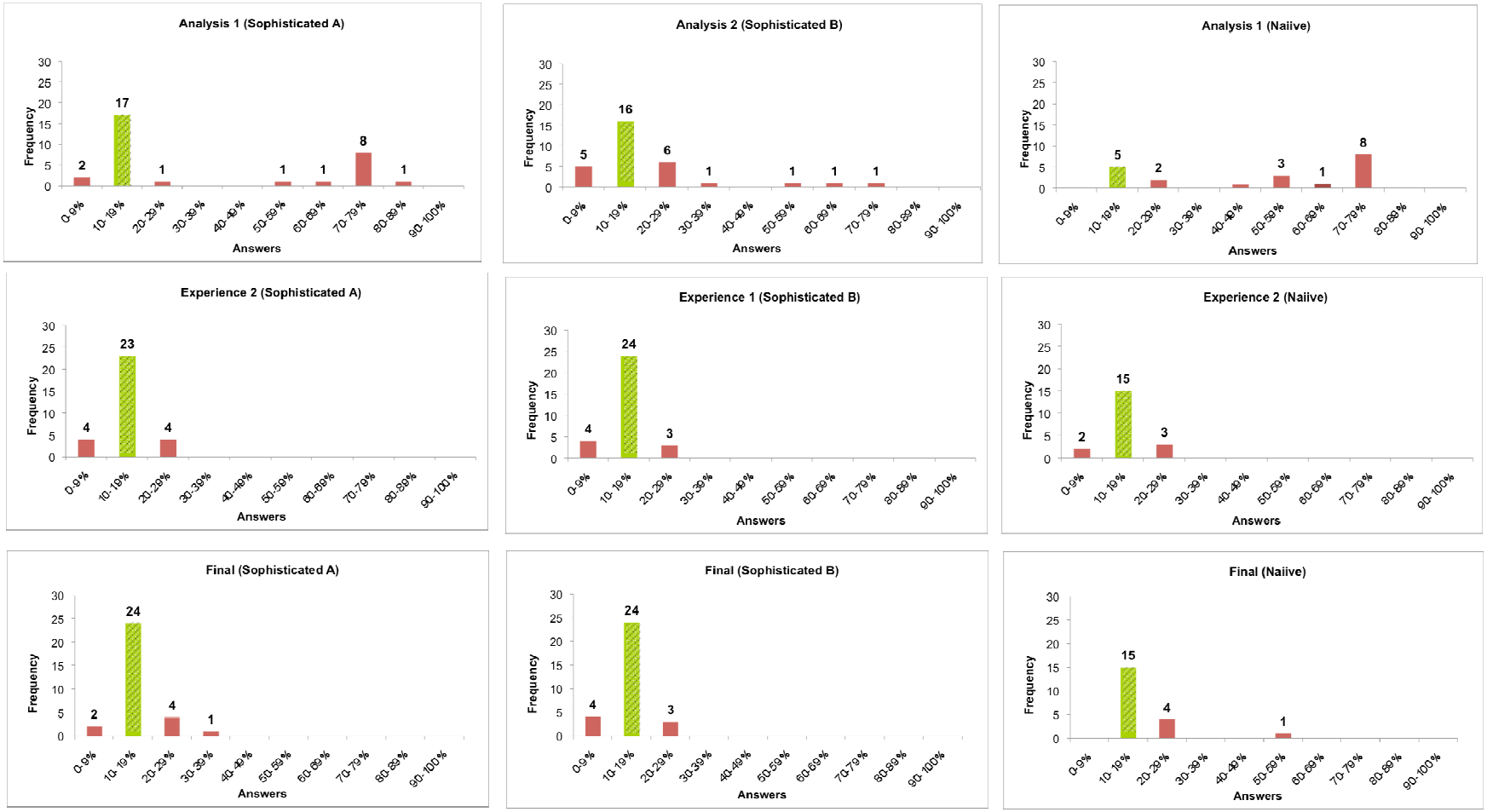
Sample sizes for Sophisticated A, Sophisticated B and Naïve are 29, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 2. Histograms of answers given to the birthday problem



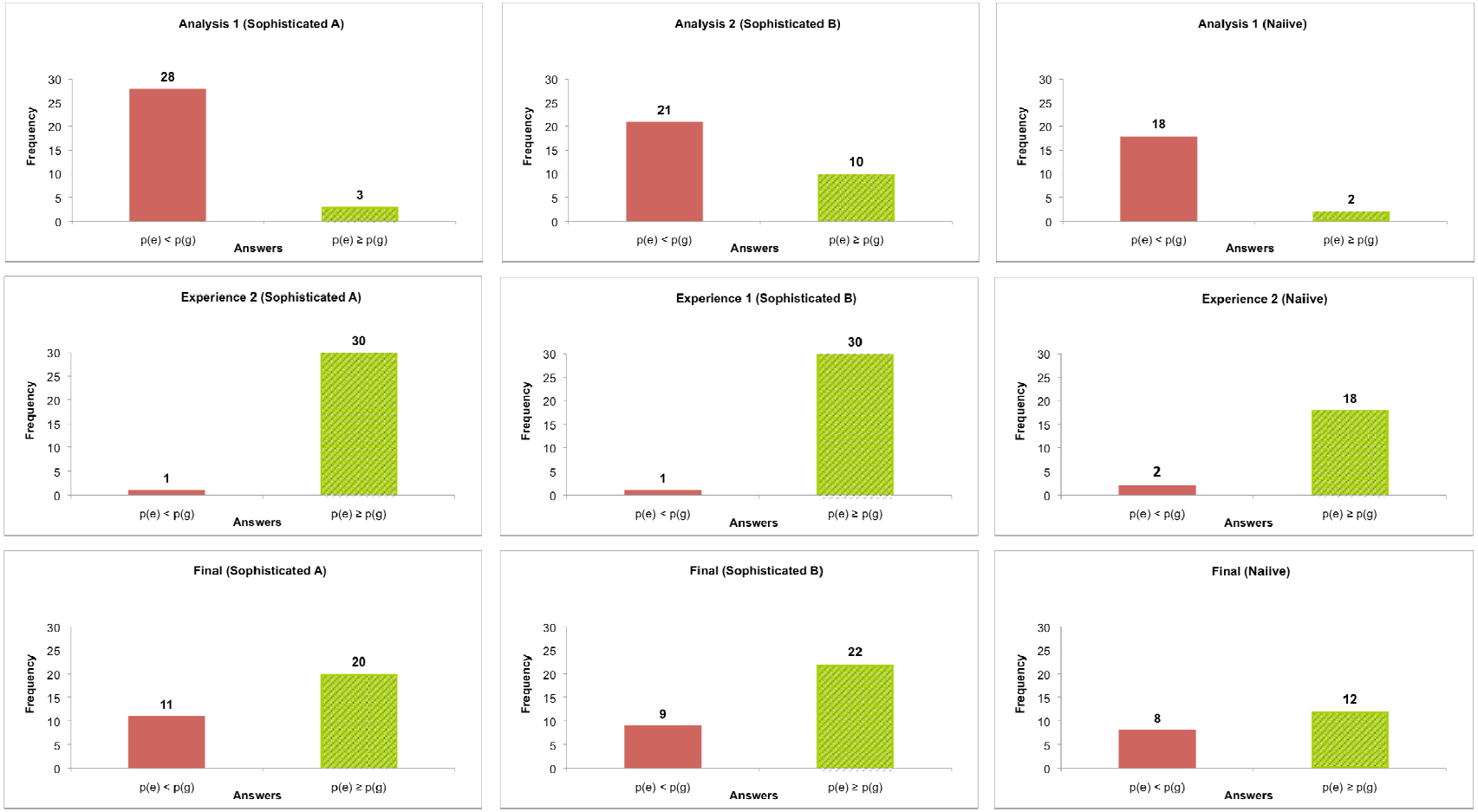
Sample sizes for Sophisticated A, Sophisticated B and Naïve are 31, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 3. Histograms of answers given to the conjunction problem



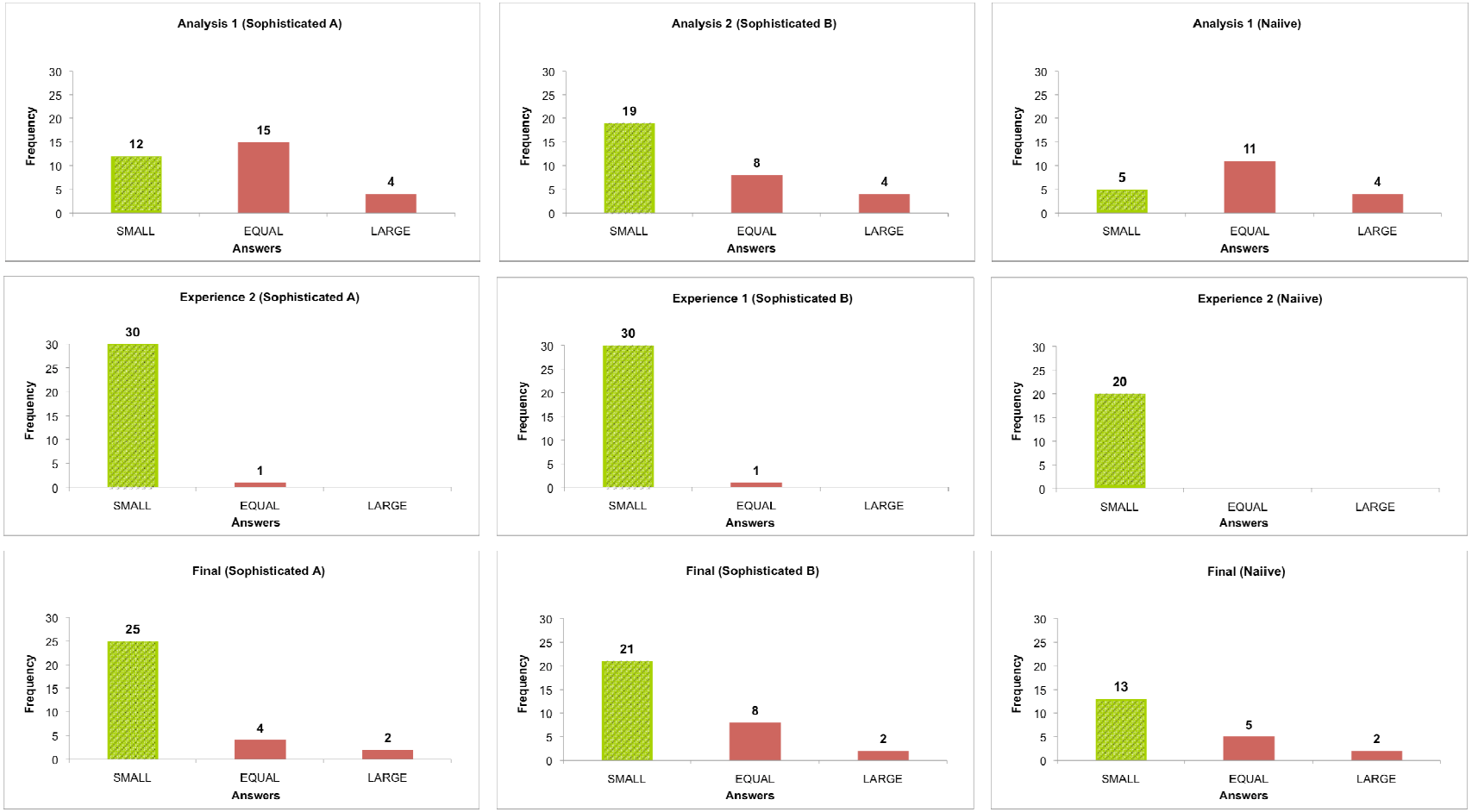
Sample sizes for Sophisticated A, Sophisticated B and Naïve are 31, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 4. Histograms of answers given to the Linda problem



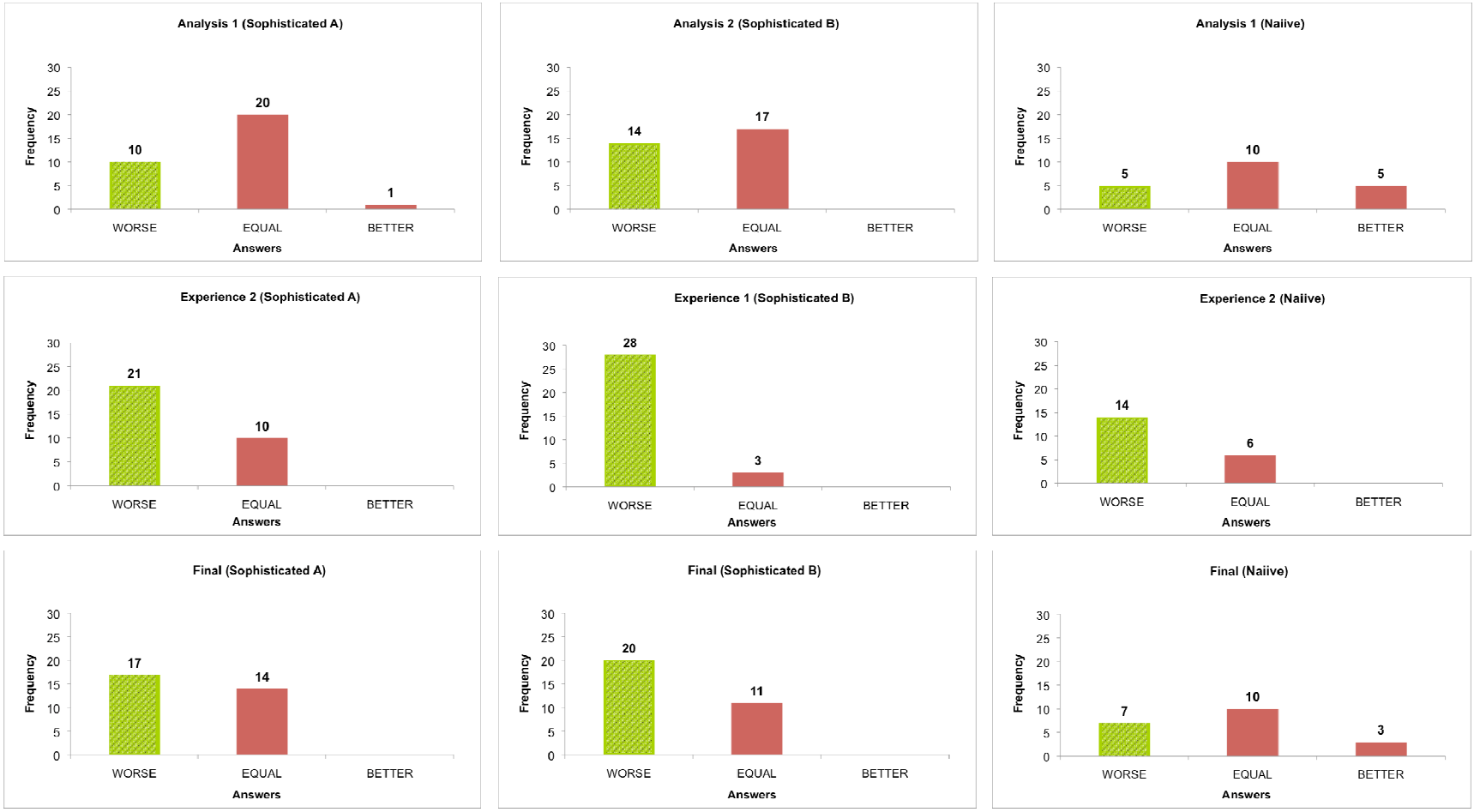
Sample sizes for Sophisticated A, Sophisticated B and Naïve are 31, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 5. Histograms of answers given to the hospital problem



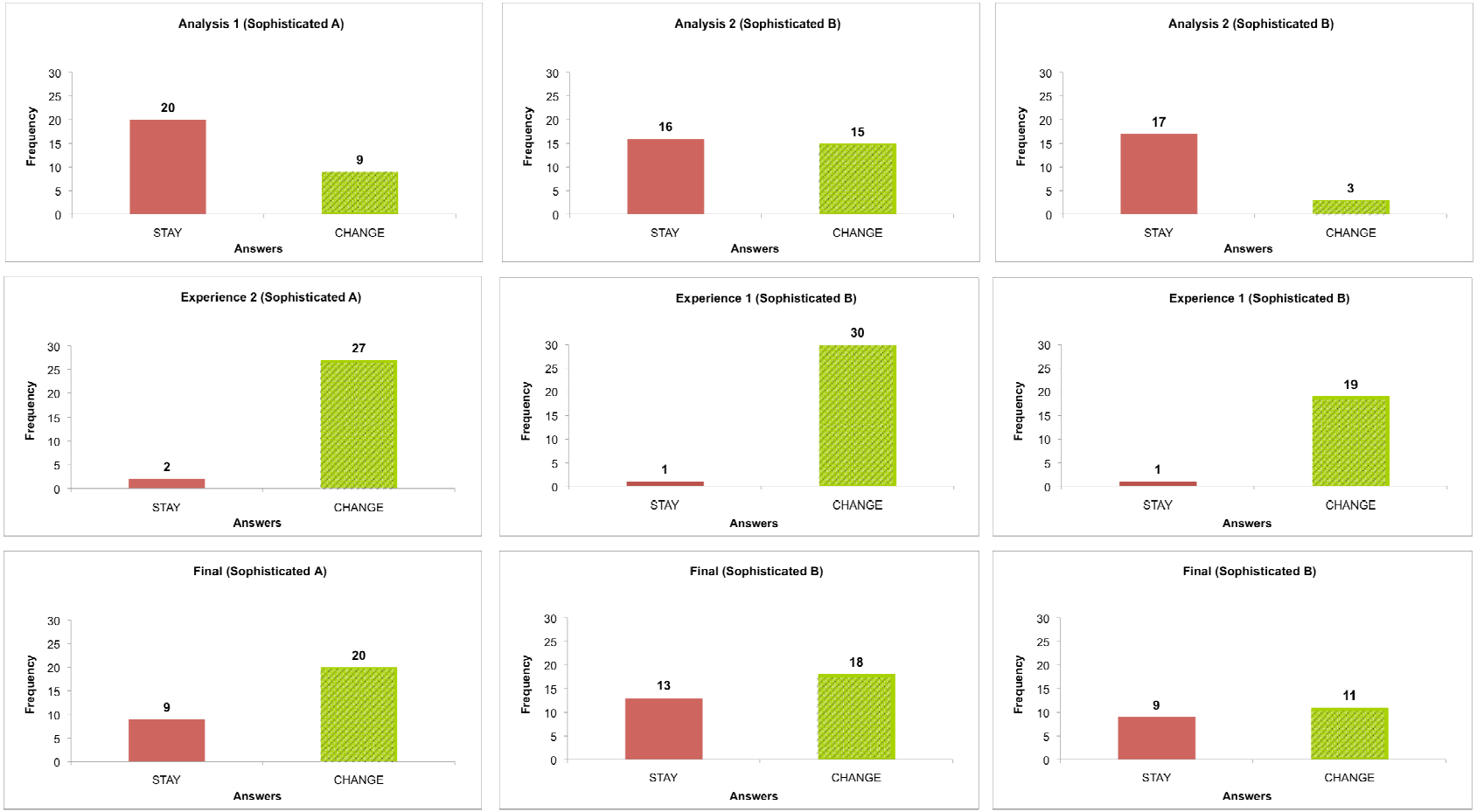
Sample sizes for Sophisticated A, Sophisticated B and Naïve are 31, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 6. Histograms of answers given to the regression toward the mean problem



Sample sizes for Sophisticated A, Sophisticated B and Naïve are 31, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 7. Histograms of answers given to the Monty Hall problem



Sample sizes for Sophisticated A, Sophisticated B and Naïve are 29, 31 and 20 respectively.
The numbers on the columns represent the number of answers.
The green (dashed) column represents the correct answer.

Figure 8: Experiment 2 Analytic group set-up

Thank you for participating in this experiment. It is anonymous, please do not write you name.

Here you will be asked to make an investment decision. You are given 40 credits. You can allocate these 40 credits in 3 ways:

I_1 : You can invest some in "Investment 1"
I_2 : You can invest some in "Investment 2"
N : You can choose not to invest some of it.

You can choose how much to put in each of these 3 options, provided that your choices add up to 40. The relationship between the investments and their effect on the outcome is given by the following linear equation:

$$\Delta Y_i = \alpha + \beta_1 \mathrm{I}_{1,i} + \beta_2 \mathrm{I}_{2,i} + e_i$$

Where "ΔY" is the **change** in resulting credits, "I_1" is the amount invested in investment 1, "I_2" is the amount invested in investment 2, "β_1" and "β_2" are the effects of investments on the change in credits and "*e*" is the random perturbation.

The return to each investment is estimated through historical data. Past 1000 investments were taken into account for each investment and an OLS regression was conducted to compute the relationship between each investment and its return

The sample statistics for the data are as follows:

Variable	Mean	Std. Dev.
ΔY	8.4	7.9
I1	11.1	5.8
I2	9.6	5.2

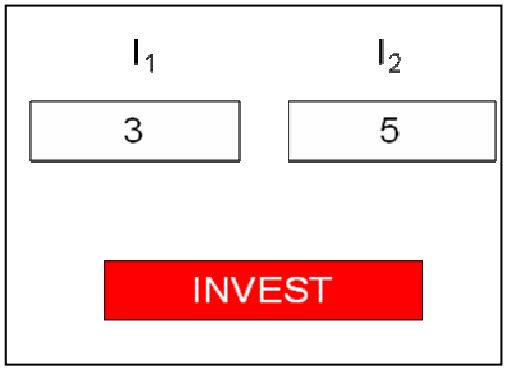
The OLS estimation results are as follows:

	Dependent Variable: ΔY	
I_1	0.5	(0.20)**
I_2	0.3	(0.05)**
Constant	-0.1	(0.15)
R^2	0.21	
N	1 000	

Standard errors in parentheses
** Significant at 95% confidence level
N is the number of observations

This means that both the investments are estimated to have positive and significant effects on the change in one's returns. Specifically, in the average, "Investment 1" is expected to generate a 50% increase and "Investment 2" is expected to generate a 30% increase over the invested amount.

Figure 9: Simulation interface used in the frequency condition



I_1	I_2	Final Y
0	5	42
0	5	37
0	5	49
0	5	39
3	5	44
3	5	49
3	5	40

Appendix A. Answers to the seven probabilistic inference problems in Experiment 1

1. Bayesian updating

p (C) = 1%
p (+ | C) = 90%
p (-- | C) = 9%

p (C | +) =
{p (C) * p (+ | C)}/{ p (C) * p (+ | C) + (1- p (C)) * p (-- | C) } ≅ 10%

Thus, the answer is:

c) Out of 10 women with a positive mammogram, about 1 has breast cancer.

2. Birthday problem

There are 365 days in a year.
The approximate probability of a birthday MATCH between any two people is 1/365. The probability of a NO MATCH is thus 364/365.
The probability of 2 NO MATCHES in a row is $(364/365)^2 = 0.9972$.
The probability of n NO MATCHES in a row is $(364/365)^n$
There are 300 different combinations of 2 people in a group of 25.
The probability of 300 NO MATCHES in a row is $(364/365)^{300} = 44\%$
The probability that there is at least one MATCH = 1 – 44% = 56%

Answer is approximately 56%.

3. Conjunction problem

p ($part_i$) = 75% , i = 1, 2, 3, …, 7
p (success) = p ($part_1$) * p ($part_2$) * … * p ($part_7$) = [p ($part_i$)]^7 = 13.3%
Approx. 13.3%

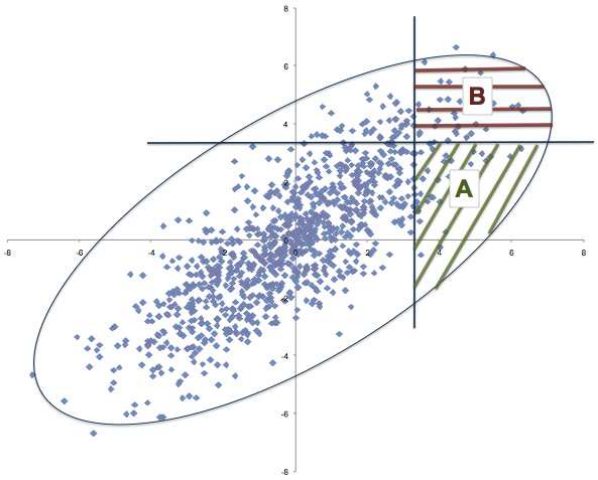
4. Linda problem (Tversky & Kahneman, 1983)

p(e) ≥ p(g) by conjunction rule.

5. The hospital problem (Tversky & Kahneman, 1974)

b) the smaller hospital……
…because smaller sample sizes exhibit more variability.

6. Regression toward the mean



A > B

b) It is more likely that the student in question now gets a worse ranking.

7. Monty Hall problem

The a priori probability that the prize is behind door *i* (D_i ; *i* = 1, 2, 3) is:

$p(D_i) = 1 / 3$

Assuming that the participant has selected door 1 (D_1), the probability that Monty opens door 2 (O_2) is

- if the prize were behind D_1; $p(O_2 \mid D_1) = 1 / 2$
- If the prize were behind D_2; $p(O_2 \mid D_2) = 0$
- if the prize were behind D3; $p(O_2 \mid D_3) = 1$

So, the probability that Monty opens door 2 is:

$$p(O_2) = \sum_{i=1}^{3} p(D_i) \cdot p(O_2 \mid D_i) = 1/6 + 0 + 1/3 = 1/2$$

Using Bayes Theorem, we have:

$$p(D_3 \mid O_2) = \frac{p(D_3) \cdot p(O_2 \mid D_3)}{p(O_2)} = \frac{1/3 \times 1}{1/2} = \frac{2}{3}$$

and

$$p(D_1 \mid O_2) = \frac{p(D_1) \cdot p(O_2 \mid D_1)}{p(O_2)} = \frac{1/3 \times 1/2}{1/2} = \frac{1/6}{1/2} = \frac{1}{3}$$

Therefore, the probability of winning is higher (2 / 3) if one changes the door, which implies that the optimal strategy is to change the initial choice, so:

b) Change to the other door

Appendix B. Statistical tests on differences between proportions of correct answers in Experiment 1

Table B1. Difference between the proportions of correct answers in Experience and Analytic

	Sophisticated A		Sophisticated B		Naïve	
	Δ	t	Δ	t	Δ	t
Bayesian updating	0.79	10.1*	0.55	5.8*	0.80	8.9*
Birthday problem	0.52	5.4*	0.48	4.6*	0.65	6.1*
Conjunction problem	0.20	1.6	0.23	2.2*	0.50	3.6*
Linda problem	0.87	14.1*	0.65	7.2*	0.80	8.4*
Hospital problem	0.58	6.2*	0.36	3.8*	0.75	7.7*
Regression toward the mean	0.36	3.0*	0.45	4.3*	0.45	3.1*
Monty Hall problem	0.62	6.3*	0.49	5.1*	0.80	8.6*

(*) indicates significantly positive difference at 95% confidence level

Table B2. Difference between the proportions of correct Analytic answers in Sophisticated B and A

	Δ	t
Bayesian updating	0.25	2.2*
Birthday problem	0.09	1.4
Conjunction problem	-0.03	-0.3
Linda problem	0.23	2.3*
Hospital problem	0.23	1.8*
Regression toward the mean	0.13	1.1
Monty Hall problem	0.19	1.4

(*) indicates significantly positive difference at 95% confidence level

Table B3. Difference between the proportions of correct answers in Sophisticated A and Naïve

	Analytic		Experience		Final	
	Δ	t	Δ	t	Δ	t
Bayesian updating	-0.03	-0.2	-0.03	-0.2	0.09	0.7
Birthday problem	0.03	0.2	-0.10	-0.7	0.06	0.4
Conjunction problem	0.30	2.3*	-0.01	-0.1	0.02	0.2
Linda problem	0.00	0.0	0.07	0.9	0.05	0.3
Hospital problem	0.13	1.05	-0.03	-1.0	0.16	1.2
Regression toward the mean	0.07	0.6	-0.02	-0.2	0.19	1.4
Monty Hall problem	0.16	1.4	-0.02	-0.3	0.14	1.0

(*) indicates significantly positive difference at 95% confidence level

Appendix C: Rationale for answers to the four questions in Experiment 2

Question 1

This question was posed to elicit an answer from the participants. We wanted them to make an investment decision with a particular expectation about the results it would lead to. The answers given suggested that the participants in all groups identified average effects quite accurately.

Question 2

This question reflects the desire to obtain a positive outcome given any investment decision. The most popular answer for this question in the Analytic group was I_1=0 and I_2=16.7. We therefore base the calculations in this section on these particular values. Answers associated with other choices can be calculated analogously.

The answer to Question 2 depends on the standard deviation of the estimated residuals (SDER). In a linear regression analysis, $SDER^2$ corresponds to the variance of the dependent variable that is unexplained by the independent variables and is captured by the statistic $(1\text{-}R^2)$. In the set-up, this is given as 21%. One can compute the SDER using the $(1\text{-}R^2)$ statistic and the variance of ΔY:

$$se(\hat{e}) = \sqrt{(Var(\Delta Y)(1-R^2)} = \sqrt{(7.9^2)(0.21)} \cong 7 \qquad \text{(A1)}$$

Given I_1=0 and I_2=16.7 the answer to Question 2 becomes:

$$\Pr(Y_i < 0 \mid X_{1,i} = 0,\ X_{2,i} = 16.7) = \Pr(\hat{C} + \hat{\beta}X_i + \hat{e}_i < 0 \mid X_{1,i} = 0,\ X_{2,i} = 16.7) =$$

$$= \Pr(\hat{e}_i < 0 - \hat{C} - \hat{\beta}X_i \mid X_{1,i} = 0,\ X_{2,i} = 16.7) = \Pr\left(\frac{\hat{e}_i}{se(\hat{e}_i)} < \frac{0 - \hat{C} - \hat{\beta}X_i}{se(\hat{e}_i)} \mid X_{1,i} = 0,\ X_{2,i} = 16.7\right) =$$

$$= \Phi\left(\frac{0 + 0.1 - 0.3*16.7}{7}\right) = \Phi(-0.7) \cong 0.24 \qquad \text{(A2)}$$

Question 3

Here, one needs to make similar calculations as for the answer to Question 2. Given I_1=0 and I_2=16.7 the answer to Question 3 becomes:

$$\Pr(Y_i < 5 \mid X_{1,i} = 0,\ X_{2,i} = 16.7) = \Pr(\hat{C} + \hat{\beta}X_i + \hat{e}_i < 5 \mid X_{1,i} = 0,\ X_{2,i} = 16.7) =$$

$$= \Pr(\hat{e}_i < 5 - \hat{C} - \hat{\beta}X_i \mid X_1 = 0,\ X_2 = 16.7) = \Pr\left(\frac{\hat{e}_i}{se(\hat{e}_i)} < \frac{5 - \hat{C} - \hat{\beta}X_i}{se(\hat{e}_i)} \mid X_{1,i} = 0,\ X_{2,i} = 16.7\right) =$$

$$= \Phi\left(\frac{5 + 0.1 - 0.3 * 16.7}{7}\right) = \Phi(-0.01) \cong 0.50 \qquad \text{(A3)}$$

Question 4

This question reflects the desire to be better off with respect to an alternative of no-action in terms of Investment 1 and 2. Finding the answer requires making one additional calculation. Specifically, we need to know the standard deviation of the difference between two random variables, that is

$$(Y_i \mid X_{1,i} = x_1, X_{2,i} = x_2) - (Y_j \mid X_{1,j} = 0, X_{2,j} = 0), \text{ where } x_1 > 0 \text{ and/or } x_2 > 0 \qquad \text{(A4)}$$

We know that $(Y_i \mid X_{1,i} = x_1, X_{2,i} = x_2)$ is an identically, independently and normally distributed random error with an estimated standard deviation of again 7. Given that a different and independent shock occurs for different individuals and actions, the standard deviation of (A4) becomes:

$$\sqrt{Var\left[(Y_i \mid X_{1,i} = x_1,\ X_{2,i} = x_2) - (Y_j \mid X_{1,j} = 0,\ X_{2,j} = 0)\right]} =$$

$$= \sqrt{Var(Y_i \mid X_{1,i} = x_1,\ X_{2,i} = x_2) + Var(Y_j \mid X_{1,j} = 0,\ X_{2,j} = 0)} = \sqrt{(7^2 + 7^2)} \cong 9.9 \qquad \text{(A5)}$$

Given I_1=0 and I_2=16.7 the answer to Question 4 becomes:

$$\Pr(Y_i \mid X_{1,i} = 0,\ X_{2,i} = 16.7 > Y_j \mid X_{1,j} = 0,\ X_{2,j} = 0) =$$

$$= \Pr(\hat{C} + \hat{\beta}X_i + \hat{e}_i - \hat{C} - \hat{\beta}X_j - \hat{e}_j > 0 \mid X_{1,i} = 0,\ X_{2,i} = 16.7,\ X_{1,j} = 0,\ X_{2,j} = 0) =$$

$$= \Pr(\hat{e}_i - \hat{e}_j > 0 - \hat{\beta}X_i + \hat{\beta}X_j \mid X_{1,i} = 0,\ X_{2,i} = 16.7,\ X_{1,j} = 0,\ X_{2,j} = 0) =$$

$$= \Pr\left(\frac{\hat{e}_i - \hat{e}_j}{se(\hat{e}_i - \hat{e}_j)} > \frac{0 - \hat{\beta}X_i + \hat{\beta}X_j}{se(\hat{e}_i - \hat{e}_j)} \mid X_{1,i} = 0,\ X_{2,i} = 16.7,\ X_{1,j} = 0,\ X_{2,j} = 0\right) =$$

$$= 1 - \Phi\left(\frac{0 - 0.3 * 16.7}{9.9}\right) = 1 - \Phi(-0.51) \cong 0.69 \qquad \text{(A6)}$$