Journal of Behavioral Decision Making, Vol. 10, 33–51 (1997)

Intuitions About Sample Size: The Empirical Law of Large Numbers

PETER SEDLMEIER
University of Paderborn, Germany

GERD GIGERENZER
Max Planck Institute for Psychological Research, Germany

ABSTRACT

According to Jacob Bernoulli, even the 'stupidest man' knows that the larger one's sample of observations, the more confidence one can have in being close to the truth about the phenomenon observed. Two-and-a-half centuries later, psychologists empirically tested people's intuitions about sample size. One group of such studies found participants attentive to sample size; another found participants ignoring it. We suggest an explanation for a substantial part of these inconsistent findings. We propose the hypothesis that human intuition conforms to the 'empirical law of large numbers' and distinguish between two kinds of tasks — one that can be solved by this intuition (frequency distributions) and one for which it is not sufficient (sampling distributions). A review of the literature reveals that this distinction can explain a substantial part of the apparently inconsistent results. ©1997 by John Wiley & Sons, Ltd.

KEY WORDS sample size; law of large numbers; sampling distribution; frequency distribution

Jacob Bernoulli, who formulated the first version of the law of large numbers, asserted in a letter to Leibniz that 'even the stupidest man knows by some instinct of nature *per se* and by no previous instruction' that the greater the number of confirming observations, the surer the conjecture (Gigerenzer *et al*., 1989, p. 29). Two-and-a-half centuries later, psychologists began to study whether people actually take into account information about sample size in judgements of various kinds. The results turned out to be contradictory: One group of studies seemed to confirm, a second to disconfirm the 'instinct of nature' assumed by Bernoulli.

In this paper, we propose an explanation that accounts for a substantial part of the contradictory results reported in the literature.

INCONSISTENT RESULTS

From one group of studies, it has been argued that people are good 'intuitive statisticians' who properly take sample size into account; from another group of studies the opposite claim has been made.

CCC 0894–3257/97/010033–19

Received 19 May 1995
Accepted 29 May 1996

Sample size is taken into account

Piaget and Inhelder (1975) reported that from age 11 or 12 children show an understanding of the role of sample size in tasks involving simple chance devices. For instance, children had to judge whether, in a simplified Galton board with only two slots (a box divided in two equal parts with a funnel in the top middle), a large sample of balls would be more likely than a small sample to generate a uniform distribution across the slots (i.e. a proportion of $p = 0.5$). Children understood that the large sample was more likely to produce a uniform distribution. Piaget and Inhelder attribute children's attention to sample size to an intuitive understanding of the 'law of large numbers'. Indeed, for Piaget and Inhelder the grasp of the mathematical law of large numbers is the 'touchstone' for understanding the notions of 'chance' and 'probability' (p. 234). In their theoretical framework, this ability is contingent on combinatorial operations (e.g. combination and permutation), which the child acquires at the stage of formal operations.

Peterson and Beach's (1967) review of research on adults' statistical thinking, including the use of sample size, agrees with Piaget and Inhelder's results: 'Experiments that have compared human inferences with those of statistical man show that the normative model provides a good first approximation for a psychological theory of inference' (p. 42). In the same vein, Evans and Pollard (1985, pp.68–69) conclude: 'Overall subjects did quite well as intuitive statisticians in that their judgements tended, over the experiments as a whole, to move in the direction required by statistical theory as the levels of Mean Difference, Sample size and Variability were varied'. According to Nisbett (1993, pp. 8–9), people have a 'highly generalized, domain-independent, but not purely syntactic' rule system for the 'law of large numbers'. All these authors, and many more (see section 'Beyond Choice Tasks') conclude from their experiments that humans take sample size into account in a broad variety of tasks.

These conclusions confirm what Jacob Bernoulli asserted two-and-a-half centuries ago. However, there is another group of studies that supports a different view.

Sample size is not taken into account

In an influential paper, Kahneman and Tversky (1972) came to the conclusion that untutored people generally disregard sample size in situations where it should play a role. For instance, one problem stated that for a period of one year, two hospitals, the larger one having about 45 births per day and the smaller about 15 births per day, recorded the days on which more than 60% of the babies born were boys (given a gender ratio of 50:50). Participants were asked which hospital recorded more such days. The majority of participants did not understand that the smaller hospital was more likely to record more such days. According to Kahneman and Tversky (1972), people ignore sample size because they use a *representativeness heuristic*. The notion of representativeness, however, has been only loosely defined in this context as the 'similarity of [the proportion or mean] to the corresponding parameter of the population' (p. 437). Kahneman and Tversky's view has been accepted by many (e.g. Bar-Hillel, 1979; Fischhoff, Slovic, and Lichtenstein, 1979; Well, Pollatsek, and Boyce, 1990). As Reagan (1989) summarized it: 'The lesson from ''sample size research'' is that people are poorly disposed to appreciate the effect of sample size on sample statistics' (p. 57).

Little is known about why participants sometimes attend to sample size and sometimes not. Kahneman and Tversky (1982), for example, suggested that participants might attend to sample size if particular conditions such as 'transparent formulation' and 'more extreme sample outcomes' are fulfilled, but possibly for the wrong reasons (p. 131). Although several factors that might influence participants' solutions have been discussed, it seems fair to say that so far no good and precise

explanation has been found for why people sometimes take sample size into account and sometimes do not.

Why do people sometimes attend to sample size and sometimes not?

In this paper, we propose an explanation for a substantial part of the contradictory results. We shall argue (1) that common intuitions about sample size conform to the *empirical law of large numbers* (a 'prehistoric' version of the mathematical law of large numbers, see below), and (2) that this law works only for one group of sample-size problems (which concern frequency distributions) but not for a second type (which concern sampling distributions). If this conjecture is valid, one should find that frequency distribution problems have been typically used by those who reported that people attend to sample size, and sampling distribution problems by those who concluded that people largely ignore sample size.

The empirical law of large numbers is not to be confused with the (mathematical) law of large numbers. The mathematical law of large numbers is about a situation in which the sample size approaches *infinity*, whereas none of the studies reviewed here deals with this situation, but with finite sample sizes. Nevertheless, several researchers have described people's reasoning as following the 'law of large numbers' (if they attend to finite sample sizes) or as violating it (if they do not). Because this misconception is widespread, we clarify in the Appendix what the law of large numbers is, why it does not apply to this research on sample size, and which mathematical results do apply.

THE EMPIRICAL LAW OF LARGE NUMBERS

What is the 'empirical law of large numbers'? Before the first law of large numbers was formulated by Jacob Bernoulli, there existed a 'prehistoric' version of the law. As Daston (1988, p. 234) observed, 'Gerolamo Cardano, Edmund Halley, and the author of the last chapters of the Port Royal *Logique*, had appealed to the principle that there was an approximate fit between observed frequencies and 'true' probabilities which improved as the number of observations increased'. This intuition — that larger samples generally lead to more accurate estimates of population means — is commonly referred to as the 'empirical law of large numbers' (e.g. Freudenthal, 1972) or the 'law of averages' (e.g. Freedman *et al*., 1991). The empirical law of large numbers is a common-sensical intuition and not a mathematical theorem like the (mathematical) law of large numbers. When Bernoulli spoke of an 'instinct of nature', he was referring to the empirical law of large numbers as a general human intuition. Note that the empirical law of large numbers says nothing more than that a large sample is better than a small sample for estimating a population parameter.

The hypothesis that common intuitions about sample size can be expressed by the empirical law of large numbers has an important implication. The empirical law of large numbers pertains to the accuracy of estimates derived from *frequency distributions* (as in Piaget and Inhelder's tasks), but by itself is not sufficient to capture the relation between variability and size of samples in *sampling distributions* (as in Kahneman and Tversky's tasks, see below). The empirical law of large numbers therefore leads us to distinguish between two kinds of tasks: (1) *frequency distribution tasks*, in which participants judge how well a sample mean (a mean of α frequency distribution) estimates a population mean and (2) *sampling distribution tasks*, in which participants judge the variance of sampling distributions. These two kinds of tasks have rarely been distinguished in research on intuitions about sample size, leading us to derive the prediction that studies reporting attention to sample size used frequency distribution tasks, while those reporting disregard of sample size used sampling distribution tasks.

FREQUENCY DISTRIBUTIONS AND SAMPLING DISTRIBUTIONS

A frequency distribution is a distribution of values from *one* sample. The overall range of values is divided into categories and the number of cases in each category is recorded. An example including a quantitative variable is the frequency distribution of heights in a sample of Italian men, where the categories might be 160 cm, 161 cm, 162 cm, and so on; an example with a qualitative (binary) variable is the distribution of male and female births during one day at a certain hospital.

We will use the term 'sampling distribution' for a distribution of means from independent samples of fixed size, drawn from the same population. A sampling distribution is not about the frequency of *observations* in different categories but about the frequency (or probability) of *sample means* falling into different categories.[1] The height distribution of 100 randomly sampled Italian men is a frequency distribution; the distribution of height means in repeated random samples of 100 Italian men is a sampling distribution.

The difference between the variance of frequency and sampling distributions is particularly evident in the limiting case in which the sample includes the whole population. In such a case, a frequency distribution will be identical to the population distribution. A sampling distribution, however, will ultimately converge into a distribution concentrated at a single value: all sample means will be identical to the population mean, and the variance of the sampling distribution will be zero.

We will now examine whether the distinction between frequency and sampling distributions can account for a substantial part of the inconsistent results.

A PROPOSAL TO RESOLVE INCONSISTENT RESULTS

Two kinds of sampling distribution tasks have been used in the literature, one in which participants have to *make a choice* regarding specific parts of a sampling distribution and one in which participants have to *construct* a sampling distribution. We begin by analyzing the 'choice tasks'.

Choice tasks

We first consider all the studies of which we are aware that involve sampling distribution tasks and compare the results with participants' performance on analogous frequency distribution tasks. Because directly comparable frequency distribution tasks are rare in the literature, we will later analyze frequency distribution tasks that have no sampling distribution analogues.

An example of a problem that has been formulated both as what we call a sampling distribution task and a frequency distribution task is the 'maternity ward' problem (Kahneman and Tversky, 1972, p. 443; the wording of the two following versions is that of Evans and Dusoir, 1977, pp. 133–134):

Maternity ward problem

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

---

[1] Sample means include sample proportions. For instance, one way of calculating the proportion of male births in a sample is to assign boys the value '1' and girls the value '0' and calculate the mean of that frequency distribution (see Freedman *et al.*, 1991, pp. 275–277, for an example).

Sampling distribution version
For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

Frequency distribution version
Which hospital do you think is more likely to find on one day that more than 60% of babies born were boys?

In the sampling distribution version, repeated samples of average sizes $n = 15$ and $n = 45$, respectively are drawn for 365 days (one year). Participants can solve the task if they realize that the variance of the sampling distribution for the smaller hospital is greater than that for the larger hospital. Evans and Dusoir (1977) constructed the frequency distribution version (this is our term, not theirs) in order to 'simplify' the task with respect to ' "on one day" as opposed to "most days included in the year" ' (p. 134). This reduces the number of samples from 365 to one and the task is now to judge 'the probability of the outcome of a single specified trial which constitutes our simplified version' (p. 135). Evans and Dusoir consider the difference as merely one of 'complexity' (p. 135). We argue instead that the second version has all the features of a frequency distribution task and therefore can be solved by the *empirical law of large numbers*, which again states that a proportion from a larger sample is a more accurate estimator of the population proportion than one from a smaller sample. Because the proportion from the larger sample is more likely to be close to the true proportion (50%), a deviation from the true proportion by 10% or more ('more than 60%') would be more likely to be found in the smaller sample. The empirical law of large numbers, however, cannot be applied to the sampling distribution version, because it is not explicit about how the variance of the distribution of the proportions depends on sample size.

A second example of a problem that has been formulated both as a frequency and a sampling distribution task is the 'post office problem' (Well, Pollatsek, and Boyce, 1990, p. 297, 'tail version' and 'accuracy version'):

Post office problem
When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.

Sampling distribution version
Every day for one year, 25 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.
Which would you expect to be true? (circle one)

1. The number of days on which the average height was 6 feet or more was greater for post office A than post office B.
2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.
3. There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

Frequency distribution version
Yesterday, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which would you expect to be true? (circle one)

1. The average height at post office A was closer to the national average than was the average height at post office B.
2. The average height at post office B was closer to the national average than was the average height at post office A.
3. There is no reason to think that the average height was closer to the national average at one post office than at the other.

In the sampling distribution version, repeated samples of sizes $n = 25$ and $n = 100$, respectively, are drawn for 365 days (one year). Participants can solve the task if they realize that the variance of the sampling distribution for post office A is greater than that for post office B. In the frequency distribution version, only one sample per post office is taken. Well, Pollatsek, and Boyce (1990) draw the same distinction between frequency and sampling distributions as we do (which they call distribution of scores and distributions of averages, respectively) but conclude that the distinction does not explain their results. They used, as did Evans and Dusoir, several other versions of the problem, which we have included in our analysis in Exhibit 1. We will now analyze the entire body of evidence available on participants' performance in choice tasks.

Evidence

The vast majority of studies employing sampling distribution tasks have two characteristic features: (1) a choice task (as opposed to an estimation or confidence task) with (2) three choices involving two sample sizes. In most cases these three response alternatives were explicitly stated as 'larger sample', 'smaller sample', and 'no difference'; in others they were embedded as in the post office problem. We will consider two-alternative forced-choice tasks and studies using other dependent variables later because these cannot be directly compared to the three-alternative tasks described above. We have found 35 studies (in eight articles) that satisfy these two criteria. Most of the studies investigated whether some factor would facilitate the use of sample-size information, such as the ratio of sample sizes (e.g. '1,000 versus 5' births instead of '45 versus 15' births). We determined the unit of a 'study' as follows: If, within an article, one problem (such as the maternity ward problem) was given to two (or more) independent groups of participants, the results of each group were coded as a separate 'study'. If one group of participants worked on more than one frequency distribution task (or sampling distribution task), then we counted these problems as one study, and the result reported in Exhibit 1 is the weighed average across these problems. If participants had to work on both frequency and sampling distribution tasks, then the results were coded as two separate studies. Exhibit 1 divides the studio into sampling distribution tasks and frequency distribution tasks.

The results are shown in the form of a stem-and-leaf display. The display shows the percentage of participants who made the correct choices for frequency distribution tasks (left side) and sampling distribution tasks (right side). The 'stem' (the central, vertical part of Exhibit 1) represents the ten's place and the 'leaves' represent the one's place of the percentage of correct choices. For sampling distribution tasks, the percentage of correct choices ranges between 7% and 59% while for frequency distribution tasks, the range is from 56% to 87%. The medians are 33% and 76%, respectively, and there is almost no overlap between the two distributions of percentages. Note that the median of correct answers for the sampling distribution tasks is exactly what one would expect by chance, that is, if participants had randomly picked one of the three alternatives. Exhibit 1 shows that there is an explanation for the apparently contradictory result that people sometimes do take sample size into account and sometimes do not. The distinction between sampling distribution tasks and frequency distribution tasks account for most of the differences within this group of studies.

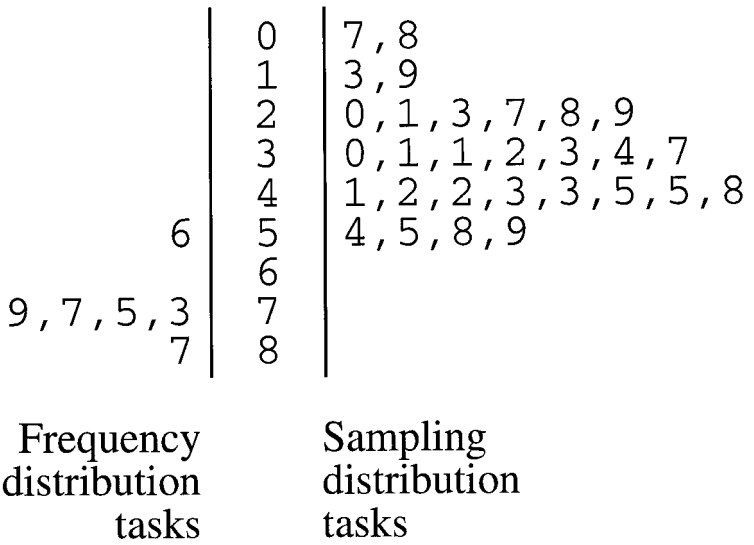

Exhibit 1. Stem-and-leaf display of percentages of participants taking sample size into account in multiple-choice studies. Results are shown separately for frequency distribution tasks (left leaves, $N = 6$ studies) and sampling distribution tasks (right leaves, $N = 29$ studies). The stem represents the ten's place, and the leaves represent the one's place. For instance, the top row of the diagram '| 0 | 7, 8' represents two studies (sampling distribution tasks) where 7% and 8% of participants took sample size into account. Studies are taken from the following sources: Bar-Hillel (1979, 1982); Kahneman and Tversky (1972); Murray, Iding, Farris, and Revlin (1987); Reagan (1989); Sedlmeier (1994); Swieringa, Gibbins, Larsson, and Sweeney (1976); and Well *et al*., (1990). See text for further explanation

Note that none of the studies reported in Exhibit 1 was designed for the analysis we have made except the study by Sedlmeier (1994). Interestingly, the results of the latter study (32% and 75% correct choices for the sampling and the frequency distribution tasks, respectively) match the medians for all studies (33% and 76%). A few studies were not included in Exhibit 1 because they differed from the others in two respects. First, in one article (Jones and Harris, 1982, Experiments I and II), three sample sizes had to be compared. There were two choice problems, one a frequency distribution task ('Question 1') and the other a sampling distribution task ('Question 2'). No difference in performance was found between the two tasks, which is inconsistent with our hypothesis. However, in a second frequency distribution task (a simple Galton board, see Exhibit 2), 94% of the participants took sample size into account. The main result was that if participants had 'hands-on' experience with the Galton board prior to working on the two other tasks, then the proportion of correct answers in these tasks approximately doubled.

Second, a few other studies used only two response categories. As Reagan (1989) demonstrated, if the 'same' response category is eliminated and participants are forced to choose either the 'smaller sample' or 'larger sample' response, then the proportion of correct choices increases considerably (by over 30 percentage points). For this reason, results from studies without the 'same' category cannot be directly compared to those in Exhibit 1. But frequency and sampling distribution tasks that both use only two response categories each can be directly compared. Evans and Dusoir (1977) report 55% and 70% correct answers for their sampling distribution tasks (as mentioned previously, this is our term, not theirs) and 85% and 85% for their frequency distribution tasks. This difference between frequency and sampling distribution tasks is smaller than the median difference in Exhibit 1, but points in the same direction ($\phi = 0.26,\ p = 0.02$, combined).

To summarize, the studies we have analyzed were heterogeneous in the sense that they employed a broad range of variables that did or did not influence the use of sample size. Despite this heterogeneity, the distinction between frequency and sampling distribution tasks was shown to be a strong predictor of participants' use of sample-size information.

Beyond choice tasks

Exhibit 1 contains only a few frequency distribution tasks because these tasks often involved dependent variables other than choice, such as confidence judgements about the accuracy of means of differently sized samples, quantitative estimates of population means given two samples of different size, and open-ended answers, to name a few. We have analyzed all studies known to us that use frequency distribution tasks in which (1) sample size was the only independent variable or, if there were several independent variables, they were systematically varied with sample size,[2] and (2) the assumptions underlying three well-known mathematical results that justify the superiority of larger samples (variance of means, Chebychev's inequality, and central limit theorem — see Appendix) were satisfied. The key assumption is that random variables — such as 'height' or 'gender' — are independently and identically distributed.[3] Exhibit 2 shows 17 articles representing 35 studies (not counting Piaget and Inhelder's single-case studies), where a 'study' is defined as in Exhibit 1. Exhibit 2 summarizes the kind of task, the measure, and the results in each study.

Do the results concerning the use of sample-size information in frequency distribution tasks reported in Exhibit 1 hold up in Exhibit 2? The 17 articles report a broad variety of dependent measures and experimental conditions. We do not see any way to compare these studies directly as in Exhibit 1, but it is useful to look at the general magnitude of the effect on performance due to the experimental variation of sample size. We calculated the effect sizes r, q, and η when the necessary information was given (Exhibit 2). The measure r expresses effect size as a Pearson correlation coefficient, the measure η can be treated as r for practical purposes, and q is the difference between Fisher z transformed correlations (Cohen, 1988; Rosenthal and Rosnow, 1991). Cohen's conventions for what constitutes a small, medium, or large effect size are identical for the three measures. The median effect size obtained in these studies (one per study) by varying sample size is 0.43, a medium to large effect by Cohen's standards. In only 4 out of 35 studies was sample size largely neglected (Evans and Dusoir, 1977, Experiment 1; Jones and Harris, 1982, Question 4 in Experiment 1; Jepson, Krantz and Nisbett, 1983, Study 1; Evans and Pollard, 1985, Experiment 2).

With few exceptions, the studies in Exhibit 2 show that participants generally take sample size into account in frequency distribution tasks. There is no simple, quantitative way to compare the amount of use of sample-size information in Exhibit 2 and Exhibit 1. In Exhibit 2 the effect size found in a

---

[2] Studies that varied sample size unsystematically with other independent variables, such as proportion and population size, and thus did not allow us to disentangle the effect of sample size from that of other variables, were not included. For instance, Evans and Dusoir (1977, Experiment 1) used two conditions in which sample size and proportion were simultaneously but not systematically varied, making it impossible to disentangle the effect of sample size from the effect of proportion. These two conditions are not included in Exhibit 2, but the condition in which proportion was kept constant and sample size varied is included. By the same token, studies in which participants were *taught* to use sample-size information are not included.

[3] Studies included in Exhibit 2 fulfill, at least approximately, the following criteria: (1) Samples (of different sizes) were drawn randomly from the same population. (2) The sources of information for samples to be compared were the same (e.g. problems where the small sample stemmed from 'personal experience' and the large sample came from 'statistical information' were not included in Exhibit 2). (3) The dependent measure did not differ for different samples (e.g. problems in which several letters of recommendations were compared to one job interview were not included in Exhibit 2). (4) The expected value of the random variables does not change over time. If it could not be clearly determined whether or not a particular criterion had been met, that criterion was treated as if fulfilled. If participants worked on more than one problem, then only the results from problems meeting the criteria were reported. All studies reported in Exhibit 1 fulfill these criteria.

Exhibit 2. The use of sample size in frequency distribution tasks (studies not included in Exhibit 1). Numbers in brackets represent the number of 'studies' per article.

Authors [numbers of studies]	Kind of task	Measure	General results (effect sizes)[a]
Piaget and Inhelder (1975) [several single-case studies]	Various chance devices	Open-ended answers	Sample size generally taken into account from age 11
Irwin, Smith and Mayfield (1956) [3]	Chance device (cards)	Confidence in estimation	Always (18 comparisons in Experiment 1 and 6 comparisons in Experiments 2) higher confidence for judgment using larger sample
DuCharme and Peterson (1969) [1]	Chance device (poker chips)	Confidence in estimation	Confidence ('credible intervals') increases monotonically with increasing sample size
Levin (1974a, Experiment 2) [2]	Mean IQs	Estimates of population IQ	Means from larger samples had more influence on population estimates ($\eta = 0.51$ and $\eta = 0.53$ for first study, $\eta = 0.55$ and $\eta = 0.38$ for second study)
Levin (1974b) [2]	Price information from grocery stores	Preference rating of stores	Larger samples (of price information) had more influence on preference. Effect sizes cannot be calculated for first study; $r = 0.28$ and $r = 0.66$ for second study
Beach *et al.* (1974, Study 2) [1]	Chance device (cards)	Confidence in estimation	Larger confidence for proportion from larger sample ($r = 0.22$)
Levin (1975, Experiment 3) [1]	Price information from grocery stores	Preference rating of stores	Larger sample (of price information) had more influence on preference. Effect size cannot be calculated.
Evans and Dusoir (1977, Experiment 1) [1]	Chance device (text problem about coin tossing)	Decision about which experiment provides better evidence that coin is biased	15 out of 48 subjects took sample size into account in more than 20 out of 24 problems, 22 subjects ignored sample size for more than 20 out of 24 problems, and 11 subjects were inconsistent
Evans and Pollard (1982, PAIRS task) [3]	Chance device (coin tosses simulated on computer screen)	Decision about which experiment provides better evidence that coin is biased	If proportion was constant (condition PCSV), sample size was taken into account (66%, 81% and 65% correct solutions)
Jones and Harris (1982)[b] [2]	Chance devices (Galton board and counters)	Open-ended answer and preference judgment	94% of subjects (in two experiments) stated that in the Galton board the larger the sample the closer the proportion to the expected value. Only 22% of subjects showed sensitivity to sample size in the counters problem (one experiment)

(*Table continues on next page*)

Exhibit 2. Continued.

Authors [numbers of studies]	Kind of task	Measure	General results (effect sizes)[a]
Jepson, Krantz and Nisbett (1983) [2]	Various text problems	Open-ended answers	In Study 1, only problems 4 and 5 met criteria (32% of subjects took sample size into account in these two problems, on average). In Study 2, sample size was taken into account in 82% of solutions for 'probabilistic' problems (four out of five problems met criteria) and in 57% of solutions for 'objective' problems (one out of five met criteria; none of the 'subjective' problems met criteria)
Nisbett *et al*. (1983, Study 1) [1]	Various text problems	Estimated proportion of a property in population based on samples (three different sizes)	Sample size had largest effects with 'heterogeneous' properties ($r = 0.23$ for 'shreeble color' and $r = 0.52$ for 'Barratos obesity') and least with 'homogeneous' properties (ceiling effect)
Study 4 [1]	Various text problems	Choice of five explanations (only one referred to sample size)	56% and 59% of experienced subjects chose 'sample size' explanation, compared to 35% and 29% of inexperienced subjects (expected by chance: 20%)
Evans and Pollard (1985, Experiment 1) [1]	Mean IQ (information displayed as blocks of numbers)	Judgment of odds that mean of sample IQ is above or below 100	Sample size influenced odds judgments ($r = 0.52$)
Experiment 2 [1]	Mean IQ (information displayed as bar graphs)	Judgment of odds that mean of sample IQ is above or below 100	Sample size did not influence odds judgments
Experiment 3b [1]	Mean IQ (information displayed as blocks of numbers and bar graphs)	Judgment of odds that mean of sample IQ is above or below 100	Sample sizes influenced odds judgments (overall $r = 0.5$)
Kunda and Nisbett (1986a, Study 1) [1]	Text problems (course evaluations)	Prediction of which of two courses will get better evaluation	Larger sample leads to higher confidence (probability estimate) in prediction (average $q = 0.95$)[c]
Study 2 [1]	Text problems (attributes of people)	Prediction of which of two people will get higher ranking on various attributes	Larger sample leads to slightly higher confidence (probability estimate) in prediction (average $q = 0.08$)

(*Table continues on next page*)

Exhibit 2. Continued.

Study 4 [2]	Text problems (abilities, traits)	Prediction of which of two people will get higher ranking on various abilities or traits	Larger sample leads to higher confidence (probability estimate) in prediction (laypeople: $q = 0.42$ and $q = 0.31$; psychologists: $q = 0.55$ and $q = 0.19$ for abilities and traits, respectively)
Study 6 [1]	Text problems (abilities, traits)	Prediction of which of two people will get higher ranking on various abilities or traits (within-subjects design)	Larger sample leads to higher confidence (probability estimate) in prediction ($q = 0.69$ and $q = 0.81$ for abilities and traits, respectively)
Kunda and Nisbett (1986b, Study 1) [1]	Text problems (abilities, traits)	Prediction of which of two people will get higher ranking on various abilities or traits (both between- and within-subjects conditions)	Presence of larger sample mostly leads to higher confidence (probability estimate) in prediction (within-subjects condition: $q = 0.59$ and $q = 0.26$; $q = 0.74$ and $q = 0.14$; $q = -0.07$ and $q = -0.1$; between-subjects condition: $q = 0.42$ and $q = 0.23$; $q = 0.61$ and $q = 0.12$; $q = -0.08$ and $q = -0.13$)[d]
Study 2 [1]	Text problems (abilities, traits)	Prediction of which of two people will get higher ranking on various abilities or traits (within-subjects design)	Presence of larger sample mostly leads to higher confidence (probability estimate) in prediction ($q = 0.46$ and $q = 0.68$; $q = 0.74$ and $q = 0.41$; $q = -0.24$ and $q = -0.02$)
Koslowski *et al.* (1989) [3]	Text problems (stories)	Rating of likelihood of causal relationship between an 'effect' and a 'target factor'	College students took sample size into account both when a 'target factor' covaried with an 'effect' and when it did not; 6th and 9th graders did so only in the latter case
Sanitioso and Kunda (1991) [2]	Text problems (sports)	Prediction of which of two people will get higher scores in athletic competition	Larger sample leads to higher confidence (probability estimate) in prediction ($r = 0.46$ in Study 1 and $r = 0.4$ in Study 2)

Note. The criteria for the selection of studies are explained in the text and footnote 3.
[a]We calculated the effect sizes r, η and q (Cohen, 1988; Rosenthal and Rosnow, 1991) when sufficient information was available.
[b]The choice tasks in this article have already been discussed.
[c]Kunda and Nisbett (1986a,b) transformed subjects' probability estimates into correlations. The q's reported here express the differences between the Fisher z transformed correlations (taken from Kunda and Nisbett's 1986a,b figures) of the large (total-to-total) and the small (item-to-item) samples.
[d]The 'item-to-item' condition was compared with the 'total-to-total', 'total-to-item', and 'item-to-total' conditions for abilities and traits in each case.

typical study relates to the difference in confidence estimates for small and large samples, whereas in Exhibit 1 it relates to the percentage of participants choosing one out of three possible answers. Thus, in Exhibit 1, 33% is the expected result by chance (e.g. if participants choose randomly among the three alternatives), which would correspond to an effect size of zero. As mentioned earlier, the median percentage in the studies using a sampling distribution task was exactly 33% (Exhibit 1). Therefore the medium- to large-sized effects in Exhibit 2 are consistent with the pattern for frequency distribution tasks (as opposed to the sampling distribution tasks) shown in Exhibit 1. People seem to apply the empirical law of large numbers not only to frequency distribution tasks that are directly comparable to sampling distribution tasks but also to frequency distribution tasks with a wide range of dependent variables.

Constructing distributions

We will now use the distinction between frequency and sampling distributions to suggest what participants do when they are asked to construct sampling distributions. The empirical law of large numbers by itself is not sufficient to explain how sample size affects the variance of sampling distribution. Therefore, intuitions about sample size as expressed by the empirical law of large numbers cannot help in constructing sampling distributions.

Two major results have been obtained in construction tasks to date (Fischhoff, Slovic, and Lichtenstein, 1979; Kahneman and Tversky, 1972; Olson, 1976; Teigen, 1974a): (1) sampling distributions did not vary with sample size, and (2) they were flatter than what would be expected for even the smallest sample size.[4] We propose a tentative explanation for these two results: participants construct *frequency distributions* when asked to construct sampling distributions. This proposal can account for both results.

Does the *variance* of a frequency distribution change systematically with sample size? The sample variance s^2 is an unbiased estimator of the population variance. That is, its expected value is equal to the value of the population variance, irrespective of sample size (e.g. Huntsberger and Billingsley, 1973, p. 138). With a sample of any size, the best estimate of the population variance σ^2 is s^2. Thus if participants construct frequency distributions, the distributions should not vary with sample size — the first result. A frequency distribution, in addition, can always be expected to be flatter than a corresponding sampling distribution for $n > 1$. Therefore, if participants construct frequency distributions, these distributions should be flatter than sampling distributions even for small sample sizes — the second result.

This tentative explanation is supported by three pieces of further evidence. First, participants can construct realistic frequency distributions. Teigen (1974b) reported that the distributions participants constructed for the heights of male and female students were close to the actual (population) distributions (after probabilities and frequencies were normalized to add up to 100%). Second, participants tend to recall sampling distribution tasks as frequency distribution tasks. Well, Pollatsek, and Boyce (1990, Experiment 4) had their participants recall the contents of a sampling distribution task and found that 11 of 21 participants who failed on the sampling distribution task recalled the task as a frequency distribution task and only 3 participants recalled it as a sampling distribution task. Third, participants tend to construct identical distributions when asked to construct frequency or sampling

[4] We could only find one study in which participants showed sensitivity to sample size in constructing subjective distributions. Peterson, DuCharme, and Edwards (1968) studied conservatism in probability revision. In Experiment 1, participants had to construct subjective binomial sampling distributions for various levels of p and sample size ($n = 3, 5,$ and 8). In contrast to other construction studies, which provide information in the form of texts, the authors used poker chips to represent the binomial probabilities.

distributions. The present account implies (as has been demonstrated in previous research) that participants' sampling distributions do not vary with sample size and, more interestingly, that their sampling distributions should be indistinguishable from their frequency distributions. Sedlmeier (1994, Study 2) extended Kahneman and Tversky's (1972) study on sampling distributions of the heights of Israeli soldiers. In Sedlmeier's study, participants were asked to construct both sampling and frequency distributions of different-size samples. One group of participants ($N = 55$) constructed sampling distributions of the height of Israeli soldiers for sample sizes of 20 and 200 (similar to previous research). A second group of participants ($N = 56$) constructed frequency distributions of these heights for sample sizes of 20 and 200. If participants construct frequency distributions when asked to construct sampling distributions, then the distributions they constructed should be the same in all four conditions. Comparison of the distributions constructed for sample sizes of 20 and 200 showed that for each height category, the median difference between the two sample-size conditions ($n = 20$ and $n = 200$) was zero, similar to what has been found in earlier studies. The new result was that this held both for sampling distributions (where it should not) and for frequency distributions (where it should). The median distributions in the four conditions were virtually identical. This result is consistent with our argument that participants construct frequency distributions when asked to construct sampling distributions: (1) participants' sampling distributions are indistinguishable from their frequency distributions and (2) their sampling distributions show the frequency-distribution characteristic of being independent of sample size.

DISCUSSION

Why did one group of studies report that people take sample size into account when they should, while another group reported that people ignored sample size? We proposed the hypothesis that human intuition conforms to the empirical law of large numbers and distinguished between two kinds of tasks — one for which this intuition is sufficient (frequency distributions) and one for which it is not (sampling distributions).[5] A review of the literature showed that this distinction can explain a substantial part of the apparently inconsistent results. Specifically, the evidence showed that (1) frequency distribution problems that are directly comparable to sampling distribution problems elicit substantially higher percentages of participants who take sample size into account, with almost no overlap between the distributions of percentages; and (2) frequency distribution problems not directly comparable to sampling distribution problems result in participants' generally taking sample size into account. We also proposed, tentatively, what participants do if they have to construct sampling distributions: they construct frequency distributions.

We do not mean to imply that there are no factors aside from the distinction between frequency and sampling distribution tasks that influence the use of sample size. For choice tasks, for instance, several such factors have been reported, including 'hands-on' experience with a simple Galton board (Jones and Harris, 1982), different ratios of sample size (Murray *et al.,* 1987), extreme 'cut-off' percentages (Bar-Hillel, 1979, 1982; Evans and Dusoir, 1977), and the part of the distribution to which the question refers (e.g. Kahneman and Tversky, 1972; Reagan, 1989). Comparatively high attention to sample size has been reported when the question posed to the participants referred to the center (as opposed to the tails) of the distribution (Well, Pollatsek, and Boyce, 1990), but this variant seems to have been studied

[5] One might think that people try to solve a sampling distribution task by splitting it up into many frequency distribution problems (such as 365 frequency distribution problems — one for each day of the year — in the post office problem), but this is not what people commonly do. After all, the intuition behind the empirical law of large numbers applies to a single mean or proportion, and not to a distribution of means or proportions.

only for sampling distribution tasks. There is some indication that population size can influence judgements about sample size (Evans and Bradshaw, 1986). For instance, Bar-Hillel (1979) proposed that it is not (absolute) sample size but relative sample size (relative to the population size) that people attend to, but her data only partially support this hypothesis (e.g. the results of her Problems 6 and 7 seem to be inconsistent).[6]

All in all, the distinction between frequency and sampling distributions — a distinction which has received little attention so far — seems to be a powerful one in differentiating between tasks in which people do or do not take account of sample size. For the studies reviewed in Exhibit 1, we do not know of any other factor that can produce a similar clear separation.

The empirical law of large numbers reflects an intuition that people seem to apply to a variety of situations. Where does the intuitive quality of the empirical law of large numbers come from? Research on animal foraging might provide some hints. Bumble-bees, for example, have been noted to be highly sensitive to frequency distributions. Their behavior covaries with changing means and variances in distributions of nectar (Real, 1991). Birds also show sensitivity to means and variances of variables relevant to their survival (Real and Caraco, 1986). When choosing between several foraging alternatives (e.g. flowers of a certain type), it is evolutionarily adaptive to have computational rules (or 'intuitions') about how to estimate the 'gain' (mean) and the 'risk' (variance) associated with a specific alternative. Although the literature on foraging documents that animals adapt to changes in frequency distributions, we know of no studies that focus on animals' use of sample size in sampling distributions.

Foraging is one important adaptive task in which intuitions about frequency distributions play a central role. But estimates of means and proportions are of more general importance in everyday life. For instance, many cultures value the knowledge of older men and women, presumably because they can draw on a larger sample of observations than younger people in making predictions about the behavior of nature and humans. So the intuition that estimates and predictions based on larger samples tend to be more accurate might just be the cumulative result of millennia of experience.

Why is the role of sample size in sampling distributions so hard to grasp? One consideration is that frequency distributions are involved in everyday problems of estimation and prediction, whereas the rule that the variability of a sampling distribution decreases with increasing sample size seems to have only few applications in ordinary life. In general, taking repeated samples and looking at the distribution of their means is rare in the everyday and only recent in scientific practice. For instance, the pioneers of systematic experimentation in the nineteenth century, such as the British agriculturist James F. W. Johnston and German physicist and mathematician Gustav Radicke, seemed to have no intuitions about the concept of a sampling distribution in making inferences about means (Gigerenzer *et al.,* 1989, pp. 72, 130–132). In the twentieth century, sampling distributions have played a key role

---

[6] One reviewer suggested that sampling distribution tasks are less often solved than frequency distribution tasks because they are more difficult to understand. One of us (Sedlmeier, unpublished data) tested this conjecture by making participants' task as clear as possible, using a visual demonstration of the task on a computer screen. Participants who had to solve frequency distribution tasks saw on a computer screen how samples of a particular size (e.g. the number of births on one day in a hospital) were drawn and how the proportion/mean for a sample was calculated (e.g. the birth rate for boys in hospital for one specific day). Participants who had to solve sampling distribution tasks watched the same demonstrations and, in addition, were shown how several proportions/means were placed as points in a corresponding sampling distribution. Only one sample size was used for the demonstrations. After the demonstrations, participants were prompted to ask when they found it difficult to understand the tasks; none indicated such a difficulty. One group of participants saw demonstrations for three sampling distribution tasks from Kahneman and Tversky (1972; 'maternity-ward', 'word-length', and 'height' tasks) and afterwards performed these tasks; and another group saw demonstrations for three corresponding frequency distribution tasks and afterwards performed those. The result was that 39% of the sampling distribution tasks and 73% of the frequency distribution tasks were solved ($n = 11$ in each group). This difference is consistent with those reported in this paper, despite an effort to eliminate potential misunderstandings of 'what the task is' in sampling distribution tasks. Thus, there was no evidence that mere lack of understanding can explain the difference between sampling and frequency problems.

in theories of hypothesis testing, but the role of sample size in sampling distributions still seems to be poorly understood by many contemporary researchers. For example, the power of significance tests (which depends on sample size) is widely ignored and (possibly, as a result) low in many experiments (Oakes, 1986; Sedlmeier and Gigerenzer, 1989; Tversky and Kahneman, 1971). Even in statistics proper, where the theoretical concept of repeated sampling from the same population (proposed by Jerzy Neyman and Egon S. Pearson, among others) is widely used, doubt has been expressed about its relevance to actual scientific practice. Ronald A. Fisher (1956), for instance, did not believe in the reality of repeated sampling in science and ridiculed this conception as having stemmed from 'the fantasy of circles [i.e., mathematicians] rather remote from scientific research' (p. 100).

Thus with the exception of statisticians and their kin, humans may have experienced little selective pressure to develop intuitions about the impact of sample size on the variance of sampling distributions. The empirical law of large numbers, in contrast, seems to be an intuition sufficient for understanding the role of sample size in everyday life. We may conclude that Jacob Bernoulli was correct in asserting that humans possess an 'instinct of nature' that attends to sample-size information. The psychological literature reviewed here suggests that this instinct is akin to the empirical law of large numbers.

APPENDIX

This appendix clarifies what the law of large numbers is, why it does not apply to the psychological research on sample size, and which mathematical results do apply.

What is the law of large numbers?

Simeon Denis Poisson (1837) was the first to introduce the term 'law of large numbers' for Bernoulli's theorem, which had been published posthumously in *Ars Conjectandi* (1713). In modern notation, Bernoulli's version of the theorem can be stated as follows (Stigler, 1986, p. 66). Suppose an experiment with two possible outcomes is to be repeated many times. If p is the probability of success in any single experiment, and if non-negative numbers ε and c are specified, then the number of trials n can be determined such that the number of *observed* successes m in n trials satisfies

$$P\left(\left|\frac{m}{n} - p\right| \leqslant \varepsilon\right) > cP\left(\left|\frac{m}{n} - p\right| > \varepsilon\right). \tag{A1}$$

The setup described above is known today as a 'Bernoulli process'. Bernoulli himself used an urn model with r 'fertile' and s 'sterile' equally likely cases so that $p = r/(r+s)$. He set ε equal to $1/(r+s)$ and proposed making c large enough to ensure 'moral certainty'. Bernoulli calculated the number of trials required for the case in which $r = 30$ and $s = 20$ and, because he had high standards of moral certainty, for $c = 1000$, 10,000, and 100,000 (Bernoulli, 1713, p. 238). For $c = 1000$ — where the probability P of m/n falling within the interval [29/50, 31/50] is at least 1000 times larger than the probability of m/n falling outside of that interval — he calculated that he would need at least $n = 25{,}550$ observations. This discouragingly large number might have been one reason for the abrupt conclusion of his *Ars Conjectandi* (Stigler, 1986, p. 77).

A first confusion abut the theorem stems from Bernoulli himself. The theorem assumes that p is known. However, Bernoulli also seems to have wanted to apply his theorem (illegitimately) to calculate the probability that the observed ratio m/n equalled an unknown p (Daston, 1988, p. 232; Pearson, 1925, p. 205).

The modern reformulation of Bernoulli's theorem (e.g. Maistrov, 1974, p. 201) considers only the limiting case. In a more general form (which applies to means as well as proportions) going beyond Bernoulli processes, the law of large numbers can be stated as follows: Assume that the $X_i (i = 1, 2, \ldots)$ are independently and identically distributed random variables, each having a finite mean $E[X_i] = \mu$. Then, as n becomes arbitrarily large, the probability that the deviation of the mean of the random variables X_i from their expected value μ exceeds ε approaches 0. In formal terms (for a proof see Scheaffer, 1990, p. 282),

$$\lim_{n \to \infty} P\left(|\frac{1}{n} \sum_{i=1}^{n} X_i - \mu| \geqslant \varepsilon \right) = 0. \tag{A2}$$

Equation (A2) is a version of what is known as the 'weak law of large numbers'. The most general form of the weak law of large numbers was proven by the Russian mathematician Khintchine (Feller, 1957, p. 229). The *strong* law of large numbers is often taken as the theoretical basis for deriving probabilities from relative frequencies (Feller, 1957, pp. 189–190). For Bernoulli trials, where x_i is a 0–1 indicator variable, it can be written as

$$P\left(\lim_{n \to \infty} 1/n \sum_{i}^{n} x_i = p \right) = 1 \tag{A3}$$

(Fine, 1973, p. 95). There are several versions of both the weak and strong law of large numbers (Révész, 1968). When we refer to the 'law of large numbers' hereafter, we refer to equation (A2).

Why does the law of large numbers not apply to psychological research on sample size?

The asymptotic feature (i.e. $n \to \infty$) of the law of large numbers makes it an inappropriate model for determining how participants should solve tasks in which sample sizes are finite. As far as we know, all empirical studies on the 'law of large numbers' have used finite sample sizes. However, as we can see from the previous section, the (mathematical) law of large numbers cannot justify these claims nor can Bernoulli's formulation of the theorem, which although it can be used for finite samples, is designed for a different purpose. Bernoulli's theorem allows for the determination of a finite n, given c, but the calculation of n rests on the knowledge of p. This is not the question addressed in research on the 'law of large numbers', where n is always given, p is sometimes given, and the question typically relates to c.

If not the law of large numbers, what else could serve as a normative basis for determining when and why to consider sample sizes in judgements?

What mathematical results justify the impact of sample size?

There are three different mathematical results that provide partial justifications. The simplest result pertains to the *variance of the sample mean*. For a sequence of independently and identically distributed random variables X_i with finite variance σ^2, the variance of the mean $\bar{X}$ (in a sample of size n) is σ^2/n. Thus the variance of the mean decreases with increasing n (hereafter, the term 'mean' is assumed to include 'proportion' as well). This consideration provides a first partial justification for the superiority of larger samples. However, it does not allow for specification of the distribution of the mean and, consequently, is mute as to how probable it is for $\bar{X}$ to lie within a specified interval. The second result, *Chebychev's inequality*, provides information about the upper bound of the probability that the difference between mean and expectation is greater than or equal to an arbitrarily small number ε.

Chebychev's inequality states that the probability of a deviation (ε or larger) of random variable X from its expectation μ is less than or equal to the variance σ^2 of X divided by the square of that deviation (e.g. Feller, 1957, p. 219). When Chebychev's inequality is applied to the mean $\bar{X}$, it yields

$$p(|\bar{X} - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{n\varepsilon^2}, \tag{A4}$$

because the variance of $\bar{X}$ is σ^2/n. Because the right-hand side of the inequality approaches 0 as n increases, the upper bound of the probability that the sample mean deviates from the population mean by at least ε decreases as n increases.

However, because Chebychev's inequality specifies only the upper bound, it is just a partial justification for the superiority of larger samples. Indeed, as Stigler (1980) discusses, under certain distributional assumptions, exceptions in which smaller samples give better estimates can be constructed.

Is there a stronger justification for why means of larger samples vary less around the true value? A third result that can be invoked is the *central limit theorem*, which states that for large n, the term $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution (e.g. Huntsberger & Billingsley, 1973, p. 131). Therefore, for large n, $\bar{X}$ has (approximately) the normal distribution with mean μ and variance σ^2/n. From the central limit theorem one can infer that the probability that $\bar{X}$ is very close to the population parameter increases monotonically with n. Thus we have a third partial justification for the superiority of larger samples. If n is very large or the population distribution is normal, then the central limit theorem provides the strongest justification because it deals with probability estimates rather than crude upper bounds on these probabilities.

However, if n is small and the population distribution deviates markedly from a normal distribution, then the estimate provided by the central limit theorem may be poor, and the two previous results can provide a better justification.

These three results provide (partial) mathematical justification for the impact of sample-size information, whereas the law of large numbers does *not*.

AUTHOR NOTE

This research was supported by a Feodor Lynen stipend from the Alexander von Humboldt Foundation and a Habilitationsstipendium from the Deutsche Forschungsgemeinschaft to the first author. We would like to thank Valerie Chase, Jonathan Evans, Dan Goldstein, Ralph Hertwig, Anita Todd, and two anonymous reviewers for their insightful remarks. Special thanks go to Berna Eden for her contribution and comments.

REFERENCES

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

Bar-Hillel, M. 'Studies of representativeness', in Kahneman, D., Slovic, P., and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and biases*. Cambridge MA: Cambridge University Press, 1982.

Beach, L. R., Beach, B. H., Carter, W. B., and Barclay, S. 'Five studies of subjective equivalence', *Organizational Behavior and Human Performance*, **12** (1974), 351–371.

Bernoulli, J. *Ars conjectandi*, Basilea: Thurnisius, 1713.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn) Hillsdale, NJ: Erlbaum, 1988.

Daston, L. *Classical Probability in the Enlightenment*, Princeton, NJ: Princeton University Press, 1988.

DuCharme, W. M. and Peterson, C. R. 'Proportion estimation as a function of proportion and sample size', *Journal of Experimental Psychology*, **81** (1969), 536–541.

Evans, J. St B. T. and Bradshaw, H. 'Estimating sample-size requirements in research design: A study of intuitive statistical judgment', *Current Psychological Research & Reviews*, **5** (1986), 10–19.
Evans, J. St B. T. and Dusoir, A. E. 'Proportionality and sample size as factors in intuitive statistical judgement', *Acta Psychologica*, **41** (1977), 129–137.
Evans, J. St B. T. and Pollard, P. 'Statistical judgement: A further test of the representatives construct', *Acta Psychologica,* **51** (1982), 91–103.
Evans, J. St B. T. and Pollard, P. 'Intuitive statistical inferences about normally distributed data', *Acta Psychologica*, **60** (1985), 57–71.
Feller, W. *An Introduction to Probability Theory and its Applications. Vol 1* (2nd edn), New York: Wiley, 1957
Fine, T. L. *Theories of Probability: An examination of foundations*, New York: Academic Press, 1973.
Fischhoff, B., Slovic, P. and Lichtenstein, S. 'Subjective sensitivity analysis', *Organizational Behavior and Human Performance*, **23** (1979), 339–359.
Fisher, R. A. *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd, 1956.
Freedman, D., Pisani, R., Purves, R. and Adhikari, A. *Statistics* (2nd edn), New York: Norton, 1991.
Freudenthal, H. 'The ''empirical law of large numbers'' or ''the stability of frequencies'' ', *Educational Studies in Mathematics*, **4** (1972), 484–490.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L. *The Empire of Chance: How probability changed science and everyday life*, Cambridge: Cambridge University Press, 1989.
Huntsberger, D. V. and Billingsley, P. *Elements of Statistical Inference* (3rd edn), Boston: Allyn and Bacon, 1973.
Irwin, F. W., Smith, W. A. S. and Mayfield, J. F. 'Tests of two theories of decision in an ''expanded judgments'' situation', *Journal of Experimental Psychology* **51** (1956), 261–268.
Jepson, C., Krantz, D. H. and Nisbett, R. E. 'Inductive reasoning: Competence or skill?' *The Behavioral and Brain Sciences*, **3** (1983), 494–501.
Jones, C. J. and Harris, P. L. 'Insight into the law of large numbers: A comparison of Piagetian and judgment theory', *Quarterly Journal of Experimental Psychology*, **34**A (1982), 479–488.
Kahneman, D. and Tversky, A. 'Subjective probability: a judgment of representativeness', *Cognitive Psychology*, **3** (1972), 430–454.
Kahneman, D. and Tversky, A. 'On the study of statistical intuitions', *Cognition*, **11** (1982), 123–141.
Koslowski, B., Okagaki, L., Lorenz, C and Umbach, D. 'When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning', *Child Development*, **60** (1989), 1316–1327.
Kunda, Z. and Nisbett, R. E. 'The psychometrics of everyday life', *Cognitive Psychology*, **18** (1986a), 195–224.
Kunda, Z. and Nisbett, R. E. 'Prediction and the partial understanding of the law of large numbers', *Journal of Experimental Social Psychology*, **22** (1986b), 339–354.
Levin, I. P. 'Averaging processes and intuitive statistical judgments', *Organizational Behaviour and Human Performance*, **12** (1974a), 83–91.
Levin, I. P. 'Averaging processes in ratings and choices based on numerical information', *Memory & Cognition*, **2** (1974b), 786–790.
Levin. I. P. 'Information integration in numerical judgments and decision processes', *Journal of Experimental Psychology: General*, **104** (1975), 39–53.
Maistrov, L. E. *Probability Theory: a historical sketch*. New York: Academic Press, 1974.
Murray, J., Iding, M., Farris, H. and Revlin, R. 'Sample-size salience and statistical inference', *Bulletin of the Psychonomic Society*, **25** (1987), 367–369.
Nisbett, R. E. (ed.) *Rules for Reasoning*, Hillsdale, NJ: Erlbaum, 1993.
Nisbett, R. E., Krantz, D. H., Jepson, C. and Kunda Z. 'The use of statistical heuristics in everyday inductive reasoning', *Psychological Review*, **90** (1983), 339–363.
Oakes, M. *Statistical Inference: A commentary for the social and behavioral sciences*, New York: Wiley, 1986.
Olson, C. L. 'Some apparent violations of the representativeness heuristic in human judgment', *Journal of Experimental Psychology: Human Perception and Performance*, **2** (1976), 599–608.
Pearson, K. 'James Bernoulli's theorem', *Biometrika*, **17** (1925), 14–210.
Peterson, C. R. and Beach, L. R. 'Man as an intuitive statistician', *Psychological Bulletin*, **68** (1967), 29–46.
Peterson, C. R., DuCharme, W. M. and Edwards, W. 'Sampling distributions and probability revisions', *Journal of Experimental Psychology*, **76** (1968), 236–243.
Piaget, J. and Inhelder, B. *The Origin of the Idea of Chance in Children* (L. Leake, Jr, P. Burrel and H. D. Fishbein, trans.), New York: Norton, 1975 (original work published 1951).
Poisson, S. D. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédé des règles générales du calcul des probabilités, Paris:* Bachelier, 1837.

Reagan, R. T. 'Variations on a seminal demonstration of people's insensitivity to sample size', *Organizational Behavior and Human Decision Processes*, **43** (1989), 52–57.
Real, L. A. 'Animal choice behavior and the evolution of cognitive architecture', *Science*, **253** (1991), 980–986.
Real, L. and Caraco, T. 'Risk and foraging in stochastic environments', *Annual Review of Ecology and Systematics*, **17** (1986), 371–390.
Révész, P. *The Laws of Large Numbers*, New York: Academic Press, 1968.
Rosenthal, R. and Rosnow, R. L. *Essentials of Behavioral Research: Methods and data analysis* (2nd edn), New York: McGraw-Hill, 1991.
Sanitioso, R. and Kunda, Z. 'Ducking the collection of costly evidence: Motivated use of statistical heuristics', *Journal of Behavioral Decision Making*, **4** (1991), 161–178.
Scheaffer, R. L. *Introduction to Probability and its Applications*, Belmont, CA: Duxbury Press, 1990.
Sedlmeier, P. 'People's appreciation of sample size in frequency distributions and sampling distributions', unpublished manuscript. University of Chicago, 1994.
Sedlmeier, P. and Gigerenzer, G. 'Do studies of statistical power have an effect on the power of studies?' *Psychological Bulletin*, **105** (1989), 309–316.
Stigler, S. M. 'An Edgeworth curiosum', *The Annals of Statistics*, **8** (1980), 931–934.
Stigler, S. M. *The History of Statistics: The measurements of uncertainty before 1900*, Cambridge, MA: Belknap/Harvard University Press, 1986.
Swieringa, R., Gibbins, M., Larsson, L. and Sweeney, J. L. 'Experiments in the heuristics of human information processing', *Journal of Accounting Research*, **4** (Suppl.) (1976), 159–187.
Teigen, K. H. 'Subjective sampling distributions and the additivity of estimates', *Scandinavian Journal of Psychology*, **15** (1974a), 50–55.
Teigen, K. H. 'Overestimation of subjective probabilities', *Scandinavian Journal of Psychology*, **15** (1974b), 56–62.
Tversky, A. and Kahneman, D. 'Belief in the law of small numbers', *Psychological Bulletin*, **73** (1971), 105–110.
Well, A. D., Pollatsek, A. and Boyce, S. J. 'Understanding the effects of sample size on the variability of the mean', *Organizational Behavior and Human Decision Processes*, **47** (1990), 289–312.

Authors' biographies:
Peter Sedlmeier is currently an associate at the University of Paderborn with a Habilitationsstipendium of the Deutsche Forschungsgemeinschaft. He works on tutorial systems for judgement under uncertainty, and on models of frequency counting.

Gerd Gigerenzer is Director at the Max Planck Institute for Psychological Research, Munich. He works on ecological rationality, on fast and frugal 'satisficing' algorithms that make sound inferences, and on the domain-specificity of judgement and decision making.

Authors' addresses:
Peter Sedlmeier, Fachbereich 2 — Psychology, University of Paderborn, 33095 Paderborn, Germany.
Gerd Gigerenzer, Max Planck Institute for Psychological Research, Postfach 44 01 09, 80750 Munich, Germany.