האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

ON THE ACCENTUATION OF CONTINGENCIES:
THE SENSITIVE RESEARCH DESIGNER
VERSUS THE INTUITIVE STATISTICIAN

by

YAAKOV KAREEV and KLAUS FIEDLER

Discussion Paper # 346 **December 2003**

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

On the Accentuation of Contingencies:

The Sensitive Research Designer versus the Intuitive Statistician

Yaakov Kareev

The Hebrew University of Jerusalem

and

Klaus Fiedler

University of Heidelberg

Running Head: THE ACCENTUATION OF CONTINGENCIES

Address:

Yaakov Kareev

School of Education and Center for Rationality

The Hebrew University

Jerusalem, Israel

email: kareev@vms.huji.ac.il

Abstract

The information used in reaching a decision between alternatives is often gleaned through samples drawn from the distributions of their outcomes. Since in most cases it is the direction of the difference in value, rather than its magnitude, that is of primary interest, the decision maker may benefit from sampling data in a way that will accentuate, rather than accurately estimate, the magnitude of that difference, as it helps to reach a decision swiftly and confidently. A reanalysis of performance in a study by Fiedler, Brinkmann, Betsch, and Wild (Journal of Experimental Psychology: General, 2000, 129, 399-418), in which participants had the freedom to sample data any way they wished, demonstrates that their apparently poor performance as estimators of conditional probability may actually reflect sophisticated sampling, which resulted in accentuating the sample value of the degree of contingency in the data. Thus, participants might be characterized as "sensitive research designers", intent on increasing the chances of detecting an effect (if one existed).

On the Accentuation of Contingencies:
The Sensitive Research Designer versus the Intuitive Statistician

Decisions often involve a choice based on the differences between alternatives on some criterion variable. The values to be compared may be means or probabilities, with the sign of the difference indicating which alternative is to be preferred. In such choice tasks, it is the direction of the difference (the ordinal information), rather than its magnitude (the interval information), that is of primary interest. When that is the case, an amplification, rather than an accurate estimate, of differences may be important, in that it helps the decision maker to reach a decision swiftly and confidently.

Indeed, in spite of the distortion it involves, accentuation of differences has often been observed. For example, people tend to amplify differences between categories and attenuate differences within them (Goldstone, 1994; Goldstone, Lippa, & Shiffrin, 2001; Harnad, 1987; Livingston, Andrews, & Harnad, 1998). This tendency is particularly pronounced in social settings, where accentuation – amplification of differences between social groups beyond what is objectively warranted – has long been investigated (Krueger 1991, Krueger & Rothbart 1990, Tajfel, 1957). Accentuation effects, though reflecting a perspective-related bias, facilitate inter-group discrimination and decisions based on differences between categories or social groups.

The detection of differences and contingencies is also of major interest to scientists, who are always eager to detect effects and relationships.[1] In fact, the sophisticated methods employed by scientists to quantify and infer the significance of effects have long been used as a yardstick against which the performance of lay people is evaluated. However, whereas

lay people's behavior in cases that call for the application of descriptive and inferential statistics has been studied extensively (see, for example, the reviews in Allan, 1993; Alloy & Tabachnik, 1984; Peterson & Beach, 1967; Pollard, 1984), their use of principles of research design – another essential facet of scientists' work – has not. This latter oversight apparently reflects the fact that, in most experimental setups, participants have very little control over aspects corresponding to research design. In real life, in contrast, people often have control over which groups or situations to observe and which samples to draw. Thus, they may have learned how to explore the environment in ways that increase the chances of detecting important differences and contingencies, and may conceivably use this procedural knowledge if presented with the opportunity to do so.

When scientists test for the existence of an effect, they conduct a statistical test that reflects the ratio between effect size (the difference between two means, $MS_{between}$) and error (the standard error of the difference between two means, MS_{within}). To increase the chances of detecting an effect, textbooks of research design (e.g., Kerlinger, 1973) explicitly advise would-be researchers to design their studies so as to increase the expected value of that ratio. While following that advice will not increase the number of false alarms, it may result in inflated estimates of effect size. Still, since scientists' primary concern is often to determine whether an effect exists at all, the amplification of effect size is viewed as a small price to pay for an increase in the probability of its detection.

Clearly, then, the employment of methods that magnify the size of an effect and thereby increase the chances of detecting it may be inherently incompatible with the attainment of another venerable goal – that of obtaining an estimate of the original effect size or, more generally, an accurate description of the environment. Thus, if people were shown to act in ways that amplify differences, if they were shown to act as "sensitive research

designers", that might explain some of their failures as intuitive descriptive and inferential statisticians (e.g., Hamilton & Guiford, 1976; Kahneman, Slovic, & Tversky, 1982; Kelly, 1967, 1971; Langer, 1975; Ross, 1977; Tversky & Kahneman, 1974). In what follows, it will be shown that, when given the opportunity to sample data freely – to "design their research" – people will often spontaneously act in ways corresponding to methods employed by scientists to increase the chances of detecting effects. This will be done by reanalyzing the data observed in a recent study by Fiedler, Brinkmann, Betsch, & Wild (2000), from the perspective of the "sensitive research designer".

In that study, participants were requested to estimate the value of a single conditional probability on the basis of a sample they could draw from a population of cases characterized by a combination of two binary variables. Unlike the case in most studies, participants actively sampled the data and hence enjoyed some of the freedom that research designers routinely have. Efficient sampling and unbiased estimation of the parameter requested could be obtained either by drawing a sample exclusively from one specific subset of the population or by sampling from the two subsets in proportion to their size in the population. However, sampling from only one subset of the population made it impossible to detect the contingency in the population, and proportional sampling did not accentuate that contingency (whereas some non-proportional sampling would have done so; see Fiedler, 2000, for an extensive discussion of the effects of sampling on estimates of contingencies). Thus, to the extent to which participants in the study deviated in their sampling from straightforward, normative sampling, which would have resulted in unbiased estimates of the requested parameter, that deviation may indicate the degree to which the goals of the "sensitive research designer" were prevailing over those of the "intuitive statistician", even when the attainment of the latter was explicitly called for.

Description of the Study

More specifically, participants were presented with four medical or diagnostic problems involving two binary variables, a predictor and a criterion. The judgment task called for an estimate of the conditional probability of a crucial criterion outcome (e.g., prenatal lung damage) given one specific predictor level (e.g., drug intake), based on a series of bi-variate stimulus observations describing the joint occurrence of criterion values (lung damage vs. no damage) and predictor values (drug intake vs. no drug). The other three tasks involved assessing the probability that dangerous complications would be suffered given hospitalization in one of two hospitals; of anorexia given unresolved sex role conflicts; and of breast cancer given positive results of a mammography test. Each participant estimated the conditional probability for all four data sets. Note that, although the stimulus information always made up a full 2x2 contingency, only two cells of the contingency table were relevant to the judgment of p(damage|drug). The two cells referring to the no-drug condition were fully irrelevant to the judgment task proper. Nevertheless, judges may act not like intuitive statisticians, concentrating only on the proportion of the crucial criterion event (lung damage) within the relevant condition (drug intake), but as intuitive research designers, who test for the presence of an effect by comparing an experimental and a control group (the damage rate when drug intake is present or absent). In other words, rather than judging a conditional probability, they may re-interpret the task as a full-blown contingency task.

Three of Fiedler et al.'s (2000) experiments, those that afforded the freedom either to choose a sample or to judge the adequacy of one, are relevant for the purpose of the present note. In two of the experiments (Exp. 2 and Exp. 3), participants were presented with a set of cases, each represented by a card that had a predictor value on one side and a criterion value on the other. When an item was drawn, its other side was exposed, revealing the item's value

on the other variable. In both Experiment 2 and 3, participants could draw any number of items (and in whatever order they wished) from either group. The two experiments differed in one important respect: In Experiment 3 participants first had to choose whether to use a file in which items were arranged by their values on the predictor or a file in which items were arranged by their values on the criterion; in Experiment 2, however, participants had no such choice, being provided instead either with a file arranged by predictor values or with one arranged by criterion values. The cards were arranged in two groups, either by the two predictor values or by the two criterion values. The number of cards in the two subfiles was clearly visible, thus revealing the relative frequency of the two values of the variable by which the file was arranged, even before any sample was drawn. In another experiment (Exp. 4), participants were presented with a description of the sampling employed in three fictitious studies, allegedly conducted by scientists in order to estimate the same conditional probabilities described above, and the results observed in them. Participants were asked to evaluate the adequacy of the research.

To appreciate and understand the participants' performance, it should be recalled that assessment of the *conditional probability* of a specific outcome event (the criterion), given a specific predictor condition, is straightforward: Only cases sharing the relevant predictor value have to be considered and the relative frequency of the relevant event in them noted. Cases with other predictor values are irrelevant for answering the question at hand and should therefore be disregarded; considering them is wasteful at best, and may lead to inaccurate estimates. Note that the normative method is also the simplest, as it requires only items bearing one predictor value. The method outlined above cannot be applied, of course, when the sample has to be drawn from a file organized by outcome (criterion) rather than predictor values. In that case, to assess conditional probability accurately, cases should be

sampled from each criterion value in proportion to their prevalence in the population. Such sampling will protect the sampler from the biased statistics that result from non-proportional sampling and will save him or her the need to employ the complicated correction procedure, calling for the use of the Bayes theorem required in that case.

Normative expectations for the participants' behavior in Fiedler et al. (2000) were therefore as follows: a) Participants given the choice between the two filing systems (Exp. 3) should choose the file arranged by predictor values. b) When sampling from a file arranged by predictor values, participants should sample only cases bearing the relevant predictor value. Such sampling should be performed by participants who chose the predictor filing system (Exp. 3) and by participants who were presented with files arranged by predictor values (half the participants in Exp. 2). c) Participants presented with files arranged by criterion values (the other half of the participants in Exp. 2) should sample items proportionally, to preserve the relative frequency of the two criterion values in the sample. d) Participants who evaluated the fictitious studies (Exp. 4) should rate the validity of studies highest when they involve predictor sampling, and only choose one predictor value and, when criterion sampling is used, should prefer studies that preserve the proportions of criterion values in their samples to studies that fail to preserve these proportions.

The results observed by Fiedler et al. fulfilled none of these expectations: Of the participants given a choice, many preferred the file arranged by criterion over that arranged by predictor values; a majority of the participants using files arranged by predictor values sampled cases of <u>both</u> predictor values; finally, virtually no-one used proportional sampling. Furthermore, the participants' estimates of the conditional probabilities reflected the biased proportions in the samples![2] These results are presented in greater detail in the next sections.

Empirical Data

a) Criterion Sampling

Our first analysis involves participants who had a choice by which variable – predictor or criterion – to sample. To the extent that many of them chose to engage in the laborious and error-prone (one is almost tempted to say "incorrect") criterion sampling rather than in the straightforward predictor sampling, this would be a strong indication that it was not conditional probability, per se, that participants were seeking.

Across all four scenarios, the proportion of participants who chose criterion sampling was .35 (these proportions were .37, .38, .42, and .23, for scenarios 1 through 4, respectively). Thus, a substantial proportion of the participants did not choose to use predictor sampling. A *t*-test, which assigned predictor sampling a score of 0 and criterion sampling a score of 1, revealed that in each of the four scenarios the average value (i.e., the proportion of criterion samplers) differed significantly from 0, the value expected if participants were to use the more efficient and appropriate predictor sampling (all *p*-values < .001).

b) Predictor Sampling

Our second analysis involved the proportion of participants who, having engaged in predictor sampling, did not restrict themselves to cases with the relevant predictor value, but instead sampled cases with both predictor values. In carrying out this analysis we distinguished between participants who were assigned to predictor-based sampling (in Experiment 2) and participants who freely chose to engage in such sampling (in Experiment 3). Of the participants assigned to predictor sampling, .71 sampled both predictor values (the values for problems 1 through 4 were .79, .70, .64, and .70, respectively). Of the participants choosing to sample by predictor, .65 sampled both predictor values in .65 (the values for problems 1 through 4 were .76, .59, .60, and .65, respectively). Scoring a choice to sample

both predictor values as 1, and a choice to sample only the relevant predictor value as 0, single sample *t*-tests showed that all eight values differed significantly from the value 0 expected if participants were to conduct the more efficient sampling (all *p*-values < .001). In other words, in both experiments, and for all scenarios, the proportion of participants (out of those engaged in predictor sampling) who restricted themselves to the relevant predictor value was small.

c) Proportion of Each Value that Was Sampled

For the large majority of people who – because they would not or could not – did not restrict their sample to the positive predictor value only, proportional sampling (drawing samples that preserve the relative frequency of each value in the population) could help to obtain accurate estimates of the values of interest: For participants sampling by criterion values, a proportional sample would provide the data necessary to derive an unbiased estimate of the conditional probability they had been asked to assess. A proportional sample would also provide all those participants, irrespective of sampling method, with the data necessary for direct derivation of an unbiased estimate of the contingency in the population.[3] Therefore, an analysis of the biases resulting from participants' deviations from proportional sampling may provide a clue to the reasons underlying their sampling behavior.

In all scenarios employed by Fiedler et al., the two values of interest – the base rates of positive predictor value (e.g., sex-role conflict in youth) and the focal criterion value (e.g., anorexia) – were less frequent than their complementary values. The relative frequencies of the positive predictor value were .29, .30, .26, and .19, for scenarios 1 through 4, respectively; the relative frequencies of the focal criterion value were .05, .13, .09, and .05, for the four scenarios, respectively.

An analysis of the samples drawn revealed that, of the people engaged in criterion sampling, the proportion of cases with the focal criterion value were, in Experiment 2, .29, .44, .40, and .29, for problems 1 through 4, respectively; in Experiment 3, the corresponding values were .30, .49, .45, and .28. In all eight cases the proportion of cases with the focal criterion value was significantly higher that the actual proportion (all 8 p-values<.001). In all, the proportion of the focal criterion value, whose average frequency across all four scenarios was .08, rose in the samples more than four fold, to .37. Fully .97 of the participants engaged in criterion sampling sampled cases such that the proportions of the two criterion values in their samples were closer to .5:.5 than they were in the population.

Of the people engaged in predictor sampling,[4] the proportion of cases sampled having the positive predictor value were, in Experiment 2, .42, .54, .42, and .39, for problems 1 through 4, respectively; in Experiment 3 the corresponding values were .50, .53, .52, and .48. In all eight cases the proportion of predictor-positive cases was significantly higher than that in the population (all p-values≤.006, 5 of the 8 p-values<.001). In all, the proportion of the 'positive' predictor value, whose average frequency across all four scenarios was .26, was almost doubled in the samples, to .48. Fully .83 of the participants engaged in predictor sampling sampled cases, such that the proportions of the two predictor values in their samples were closer to .5:.5 than they were in the population. Thus, for all problems and in all conditions, sampling behavior was characterized by large and highly significant deviations from proportional sampling, with the proportion of cases sampled closer to .5:.5 than that in the population.

The results of Experiment 4 further indicated that equal sampling was considered adequate: When presented with the design of (fictitious) studies, in which the use of data gleaned through more proportional sampling would result in a less biased estimate of the

quantity in question, participants judged research employing .5:.5 criterion sampling to be superior to research employing a more extreme, but proportional, criterion sampling scheme, one that led to more accurate estimates.

Summary and Discussion of Empirical Data

The behavior of Fiedler et al.'s participants could be summarized as follows:

1. When given a choice, many participants preferred to employ the patently inefficient criterion sampling rather than predictor sampling.

2. Of the participants engaged in predictor sampling, a majority were not satisfied merely to sample cases with the relevant predictor value; in addition they also sampled cases with a predictor value that was completely irrelevant for the task at hand.

3. Irrespective of sampling mode, participants did not engage in (the more efficient) proportional sampling; instead they drew samples in which the relative frequency of the two values was more evenly distributed than it was in the population. In a similar vein, participants evaluating the validity of the fictitious research regarded equal, rather than uneven, sampling of criterion values to constitute superior research, even though uneven sampling resulted in a more accurate estimate of the conditional probability whose estimate was the objective of that fictitious study.

These results could not have been obtained if participants were only interested in estimating the conditional probability they had been asked to assess. Since a large majority of the participants sampled items so as to view both values of the predictor and the criterion, it is clear that they were intent on assessing the overall contingency, rather than the single conditional probability they had been asked about. With contingencies (i.e., differences between different levels of an independent variable) playing a vital role in the efficient functioning of organisms, their detection and assessment is obviously a major concern of the

cognitive system. The format in which the data in the Fiedler et al. study were presented, with both predictor and criterion values explicitly mentioned, surely rendered the 2x2 contingency highly prominent. Thus it is not surprising that most participants sampled cases that enabled them to sensitively detect the contingency. However, the sampling behavior exhibited by the participants was incompatible with an attempt to obtain an accurate estimate of either the contingency or the conditional probability in question. To obtain such an unbiased estimate, they should have engaged in proportional sampling, sampling each value of the predictor (or the criterion) in proportion to its incidence in the population. Instead, the less common value of the sampled variable was greatly over-represented.

To explain this "poor" performance, we suggest not only that the participants, unawares, were making an attempt to find out if a contingency existed, but also that their choice of sampling was designed to increase the chances of detecting that contingency, by sampling in a way that would amplify its sample value. In other words, we suspect that the results reflect a case of the intuitive research designer gaining the upper hand over the intuitive statistician.

As we shall demonstrate in the theoretical analysis that follows, under almost all conditions, non-proportional sampling such as that employed by Fiedler et al.'s participants amplifies the statistic used to infer an existing relationship (or a difference). Analysis of the data actually observed by the participants then shows that this was indeed the outcome for the participants of the study: The data they sampled conveyed a relationship much stronger than that actually existing in the population from which the sample had been drawn.

The Theoretical Analyses

The measure of contingency in a 2x2 table is Δ_p, defined as the difference between the two conditional probabilities – the probability of one criterion value given one of the

predictor values and the probability of observing the same criterion value given the other predictor value. With reference to Table 1, Δ_p is defined as

$$\Delta_p = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

$$= \frac{ad - bc}{[(a+b)(c+d)]} \quad .^{5}$$

To determine the significance of Δ_p requires using the Z test for the difference between two unknown proportions.[6] Under the null hypothesis that the proportion of the criterion value of interest is the same in both groups of predictor values, the formula for testing the difference is:

$$Z = \frac{\frac{a}{(a+b)} - \frac{c}{(c+d)}}{\sqrt{\frac{(a+c)(b+d)}{(a+b+c+d)^2(a+b)} + \frac{(a+c)(b+d)}{(a+b+c+d)^2(c+d)}}}$$

or, after some algebra,

$$Z = \varphi\sqrt{N} \quad .$$

The last formula indicates that, for a given sample size, any manipulation that renders the sample value of φ – the geometric mean of Δ_p and Δ_c – more extreme, will increase the chance of concluding that an effect exists.

Implications of Participants' Sampling

As described above, the participants sampled items in such a way that the distribution of the values of the variable by which they sampled would be less extreme (i.e., closer to .5:.5) than that in the population. What effect would such non-proportional sampling have on the expected value of the Z-test, the normative statistic calculated to infer the existence of a relationship between the two variables? Since the value of that statistic equals φ*√N, and it is assumed that N (i.e., the effort put into sampling) remains unchanged, the answer lies in the

effect of non-proportional sampling on the expected value of φ. Table 2 presents the entries of the 2x2 table following non-proportional sampling by criterion values (assuming that the less frequent value of the variable sampled is over-represented, but that its frequency in the new sample does not exceed .5). Table 3 presents the entries of the 2x2 table following non-proportional sampling by predictor values.

When based on a sample in which the less common value of the variable used for sampling is over-represented (as is the case in Tables 2 and 3), the new value of φ tends to be more extreme than that in the population. This assertion is based on the results observed in a computational analysis that systematically covered all combinations of marginal distributions of predictor and criterion and all possible φ values. Distributions of predictor and criterion values were systematically varied in the range of .5:.5 to .05:.95. Then for each combination of marginal values, each possible φ value (from maximally negative to maximally positive in steps of .05) was generated, by gradually changing the cell entries. Each of these combinations defined a "parent" population with its characteristic value of φ. To explore the effects of non-proportional sampling, we now systematically modified the marginal distribution of the predictor values (and then that of the criterion values) for each of the parent populations, so that it became less extreme than that in the parent population. That modification was carried out in steps of .01, up to the point at which the marginal distribution of the variable the frequency of whose values was modified reached .50:.50. For each such new 2x2 table, we calculated the value of φ and compared it to the value observed in the parent population. That comparison revealed that with the less extreme margins, the new value of φ was more extreme than that in the original population in fully .933 of cases!

A regression analysis was then conducted, to determine the relationship between the pertinent variables and the degree to which the value of φ, following non-proportional

sampling, becomes more extreme than the original population value of φ. As independent variables we considered the original distribution of the variable by which values are sampled, the original distribution of the other variable, the degree of change in the marginal distribution of the variable used for sampling, and the population value of φ. Table 4 presents the correlation matrix for those variables. The regression analysis revealed a total multiple correlation of R=0.702. The variable most strongly related to the change in the strength of φ (r=.604) and the first to enter the multiple regression is the initial imbalance in the distribution of the variable by which values are sampled: The more skewed the distribution, the greater was the increase in the strength of φ when the values in the sample were less extreme than those in the original population. The second strongest predictor of the degree of change (r=.561) and the second to enter the multiple regression is the degree of change in the marginal distribution: The larger the change in the distribution of values, the larger was the increase in the value of φ. The third variable to enter the multiple regression equation was the imbalance in the population between values of the variable that was not used for sampling: The change in φ values was smaller, the more extreme the distribution. Its correlation with the degree of change was -.102. The last variable to enter the regression was the original correlation – φ. The more positive it was, the smaller was the change in its value. Its correlation with the degree of change was -.132.[7]

The computational analysis just described demonstrates that, in the vast majority of cases, the drawing of non-proportional samples with marginal distributions less extreme than those in the original population yields an accentuation of the value of φ. With that established, we inspected the actual outcomes of the non-proportional sampling conducted by the participants of Fiedler et al. (2000), to determine how it affected the value of the Z test in the data they observed. To that end, we calculated the value of Z test expected for each

participant if he or she were to engage in proportional sampling (by whatever variable they sampled), and compared it to the value of the Z test calculated on the basis of the sample actually drawn. As might be expected, the actual value of the Z test was much stronger than that expected under proportional sampling: An analysis of the difference between the observed and expected results of the Z tests revealed that for participants engaged in predictor sampling the mean difference was .26 ($t(170)=2.52$, $p=.013$). For participants engaged in criterion sampling, the mean difference was 1.99 ($t(156)=7.25$, $p<.001$). For all participants combined, the mean difference was 1.09 ($t(327)=7.28$, $p<.001$). In other words, the sampling engaged in by the participants in the study resulted in greatly exaggerated values of φ, and hence in the values of the Z test used to detect a relationship between variables.

Conclusion

Our reanalysis of the data observed by Fiedler et al. (2000) suggests that, given the freedom to design their own "research", people do that in a way that will help them to confidently answer the question of "is there a contingency?" (or more generally, "is there an effect?"), rather than to accurately assess the strength of the contingency (or the size of the effect). Even when explicitly asked to assess a single value, they sample data in a way that will not only make it possible to detect the original difference, but will also amplify it. In other words, not only do people care about differences between distributions; it is the information about the existence of a difference and its direction, rather than its size that they care most about.

Scientists, in designing their studies, often manipulate the situation, sampling data in a way that will expose a relationship, and will let an effect, if it exists, be detected. The present results show that lay people are also quite adept at using similar procedures when

deciding what samples to draw. Fiedler (2000) has demonstrated that sampling effects may underlie many of the biases evident in people's judgments, including such phenomena as illusory correlations (Hamilton & Guiford, 1976) and base-rate neglect (Kahneman & Tversky, 1972). He noted that, while people's judgments quite accurately reflect the value of the statistics in the samples they observe, they suffer from a deficit in metacognitive monitoring and control, as evidenced by their apparent unawareness of the distortions induced by the sampling schemes they employ. Our current analysis suggests one reason why that happens.

To our mind, the fact that the values observed in the non-representative samples that people draw are used, incorrectly, to estimate population parameters confirms that the detection of differences and relationships is more important than the accurate assessment of their values. Sampling that would help to reach correct decisions with confidence is apparently preferred to sampling that would lead to more accurate estimates, but that could render the difference more obscure and difficult to rely upon in deciding how to act. The sampling of data so as to accentuate differences should thus be viewed as an indication of the skills of the sensitive research designer, rather than of the shortcomings of the intuitive statistician.

References

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, **114**, 435-448.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, **91**, 112-149.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review,* **107**, 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**, 399-418.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General.* **123**, 178-200.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition,* **78**, 27-43.

Hamilton, D. L., & Guiford, R. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, **12**, 392-407.

Harnad, S. (1987). *Categorical perception.* Cambridge: Cambridge University Press.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-453.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlation. *Psychological Review*, **107**, 397-402.

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

Kelly, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.) *Nebraska Symposium on Motivation,* (pp. 192-238). Lincoln: Univeresity of Nebraska Press.

Kelly, H. H. (1971). Attribution in social interaction. In E. E. Jones, D. E. Kenause, H. H. Kelly, R. E. Nisbett, S. Valins, & B. Weiner (Eds.) *Attribution: Perceiving the causes of behavior* (pp. 1-26). Morristown, NJ: General Learning Press.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd. ed.). New York: Holt, Rinehart, & Winston.

Krueger, J. (1991). Accentuation effects and illusory change in exemplar-based category learning. *European Journal of Social Psychology,* **21**, 37-48.

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology,* **59***,* 651-653.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology,* **32**, 311-328.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **24**, 732-753.

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

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychological Research and Review*, **3**, 5-18.

Ross, L. (1977). The intuitive psychologist and his shortcomings. In L. Berkowitz (Ed.) *Advances in experimental social psychology,* **10**, 173-220. San Diego: Academic Press.

Tajfel, H. (1957). Value and the perceptual judgment of magnitude. *Psychological Review,* **64**, 192-204.

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

Footnotes

[1]Whenever two or more groups differ in their value on some criterion, it can be said that the values of the criterion are contingent on the values of the variable defining the groups. Thus, contingencies are synonymous with differences, with the variable defining the groups – the independent variable – regarded as the predictor. The correspondence of differences between groups with contingencies is most evident when both predictor and criterion are binary variables (see Kareev, 2000).

[2]Had a correction of sample values, resulting in accurate estimates, been observed, the participants' sampling schemes could have been dismissed as being merely less efficient.

[3]Recall that an estimate of the contingency was not required, but there could be hardly any other explanation why people who had a choice decided to sample by criterion or to sample cases from the 'negative' predictor value.

[4]We have excluded participants who engaged in predictor sampling and drew a sample consisting exclusively of cases with the positive predictor value (i.e., we excluded that minority of the participants who performed the task in the normatively prescribed, most efficient manner). Not only were these participants engaged in an undertaking other than that of the majority of their colleagues, but by having 100% of their cases drawn from the less common predictor value, their inclusion in the present analysis would result in even higher estimates of the degree of non-proportional sampling.

[5]A 2x2 table entails, in fact, two contingencies, the other, reflecting the contingency of predictor values on criterion values, which we shall call Δ_c, to distinguish it from Δ_p, is defined as

$$\Delta_c = \frac{a}{(a+c)} - \frac{b}{(b+d)}$$

$$= \frac{ad - bc}{[(a + c)(b + d)]} \ .$$

φ, the symmetrical measure of correlation in a 2x2 table, is the geometric mean of Δ_p and Δ_c. The value of φ is

$$\varphi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

[6]It is highly unlikely, of course, that lay people calculate the Z test for a difference between proportions. However, it should be recalled that the ANOVA model (Kelly, 1967) was found to be quite a good description of people's behavior in situations involving inferences of causal attribution. Similarly, Kareev, Arnon, & Horwitz-Zeliger (2002) found that, in a task that called for determining which of two distributions had the higher mean value, the result of the Z test for a difference between two means, and its significance, were the best predictors of lay people's choices and confidence in them. Therefore, inspecting if people "design their research" in a way that increases the chances of observing a more extreme value of the inferential statistic appropriate for the task at hand is not as far-fetched as might seem at first.

[7]All correlations are based on 3068 cases. All values of r reported are significant at the .001 level.

Authors Note

Yaakov Kareev, Center for the Study of Rationality and School of Education, The Hebrew University of Jerusalem; Klaus Fiedler, Institute of Psychology, University of Heidelberg. Work on this article was supported by Israel Science Foundation grant 712/98 to YK, and a grant from the Deutsche Forschungsgemeinschaft (within SFB 504) to KF. Correspondence concerning this article should be addressed to Yaakov Kareev, Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel. Email: kareev@vms.huji.ac.il.

Table 1: Standard 2x2 Table of Frequencies.

		Criterion		
		C_1	C_2	Total
Predictor	P_1	a	b	$(a+b)$
	P_2	c	d	$(c+d)$
	Total	$(a+c)$	$(b+d)$	$(a+b+c+d)=N$

Table 2: 2x2 Table of Frequencies Resulting from Non-Proportional Sampling by the Criterion, So that the Resulting Distribution of Criterion Values is Closer to .5:.5 Than the Original One.

		Criterion		
		C_1	C_2	Total
Predictor	P_1	$\left(\frac{a}{a+c}\right)(a+c-\varepsilon)$	$\left(\frac{b}{b+d}\right)(b+d+\varepsilon)$	$(a+b)-\frac{a\varepsilon}{a+c}+\frac{b\varepsilon}{b+d}$
	P_2	$\left(\frac{c}{a+c}\right)(a+c-\varepsilon)$	$\left(\frac{d}{b+d}\right)(b+d+\varepsilon)$	$(c+d)-\frac{c\varepsilon}{a+c}+\frac{d\varepsilon}{b+d}$
	Total[a]	$(a+c-\varepsilon)$	$(b+d+\varepsilon)$	$(a+b+c+d)=N$

[a] *$(a+c)>(b+d)$; $\varepsilon>0$; $(a+c-\varepsilon)\geq N/2$*

Table 3: 2x2 Table of Frequencies Resulting from Non-Proportional Sampling by the Predictor, So that the Resulting Distribution of Predictor Values is Closer to .5:.5 Than the Original One.

		Criterion		
		C_1	C_2	Total[a]
Predictor	P_1	$\left(\frac{a}{a+b}\right)(a+b-\varepsilon)$	$\left(\frac{b}{a+b}\right)(a+b-\varepsilon)$	$(a+b-\varepsilon)$
	P_2	$\left(\frac{c}{c+d}\right)(c+d+\varepsilon)$	$\left(\frac{d}{c+d}\right)(c+d+\varepsilon)$	$(c+d+\varepsilon)$
	Total	$(a+c-\frac{a\varepsilon}{a+b}+\frac{c\varepsilon}{c+d})$	$(b+d-\frac{b\varepsilon}{a+b}+\frac{d\varepsilon}{c+d})$	$(a+b+c+d)=N$

[a] *(a+b)>(c+d); ε >0; (a+b-ε)≥N/2*

Table 4: Correlation Matrix, for Variables Entering the Multiple Regression for Increase in φ Following a Change in Marginal Distribution.

	Increase in φ	Imbalance, P	Change (ε)	Imbalance, Q	Original φ
Increase in φ	--	.604	.561	-.102	-.132
Imbalance, P	.604	--	.549	.159	.040
Change (ε)	.561	.549	--	.087	.025
Imbalance, Q	-.102	.159	.087	--	.232
Original φ	-.132	.040	.025	.232	--