Probabilistically Valid Inference of Covariation From a Single $x,y$ Observation When Univariate Characteristics Are Known

Michael E. Doherty, a Richard B. Anderson, a Amanda M. Kelley, a,*
James H. Albertb

aDepartment of Psychology, Bowling Green State University
bDepartment of Mathematics, Bowling Green State University

Received 1 June 2007; received in revised form 24 June 2008; accepted 1 September 2008

Abstract

Participants were asked to draw inferences about correlation from single $x,y$ observations. In Experiment 1 statistically sophisticated participants were given the univariate characteristics of distributions of $x$ and $y$ and asked to infer whether a single $x,y$ observation came from a correlated or an uncorrelated population. In Experiment 2, students with a variety of statistical backgrounds assigned posterior probabilities to five possible populations based on single $x,y$ observations, again given knowledge of the univariate statistics. In Experiment 3, statistically naive participants were given a problem analogous to that given in Experiment 1, framed verbally. Experiment 4 replicated Experiment 3 but added an ‘‘impossible to determine’’ response option. Models that rely on computing sample correlations make no predictions about these investigations. From a Bayesian perspective, participants’ inferences in all four experiments tended to make probabilistically valid inferences as long as the single datum was directional. The results are discussed in light of the Brunswikian notion of vicarious functioning.

Keywords: Psychology; Decision making; Judgment; Inference; Reasoning; Statistics; Correlation; Human experimentation

1. Introduction

It is a commonplace that people ‘‘jump to conclusions,’’ even to the extent that they may generalize to an entire ethnic group based on a single observation. For example, a person

Correspondence should be sent to Richard B. Anderson, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403. E-mail: randers@bgsu.edu

*Amanda Kelley is now at the US Army Aeromedical Research Laboratory.
may infer that a correlation exists between ethnicity and some personality trait after observing a single instance of a member of that group behaving in a way supposedly expressing that trait. In this paper we argue that, under limited conditions, inference of correlation based on a single observation may be warranted and will result in correct generalizations at a greater-than-chance level. The claim that probabilistically valid inferences about covariation can be drawn from a single observation is made against a backdrop of very different claims, ones made sufficiently often that they can be considered the received view of covariation assessment—at least among those researchers steeped in traditional correlational statistics. People undoubtedly do draw such inferences, but the important idea in the explicit claim in the title is carried by the phrase ‘‘probabilistically valid.’’ Inferences based on single observations may of course be wrong, but the claim herein is that such inferences are likely to be correct at a greater-than-chance level.

The idea that probabilistically valid assessment of covariation requires the observation of multiple, paired observations of the variables in question is virtually ubiquitous and is typically made by unquestioningly referring to ‘‘normative properties of correlation coefficients’’ (Hutchinson & Alba, 1997), or ‘‘the normative model’’ (Shanks, 2004; italics added). Numerous investigators have assumed that inferences about correlation, in order to be defensible, must be based on the data necessary to the calculation of sample correlation coefficients. Alloy and Tabachnik’s (1984) widely cited paper asserts that studies of the judgment of covariation between continuous variables ‘‘generally compare estimates of covariations with statistical estimates based on the Pearson r’’ (p. 120). In another widely cited paper on judgment of covariation, Crocker (1981) referred to the Pearson $r$ as ‘‘the normative, or statistically correct, model for making covariation judgments…’’ (p. 272, italics added). And the substantial literature on illusory correlation initiated by investigations by Chapman and Chapman (1967) is predicated on the assumption that the correct statistical model for inference of covariation is the sample $\varphi$, which is the Pearson $r$ for dichotomous data. Reference to the illusory correlation literature makes it clear that we are not claiming that people draw inferences about correlation by performing some operations that rely on the same data as involved in the computation of the Pearson $r$. On the contrary, the illusory correlation literature shows that people infer correlations in distinctly non-Pearsonian ways, and that such deviations from the normative models are deemed errors. See Arkes and Harkness (1983), in particular, for a variety of descriptive models which deviate from the statistical model of how people infer covariation.

1.1. Experimental research in the detection of contingency

Research in single-cue probability learning (e.g., Edgell, 1978) and multiple-cue probability learning (e.g., Holzworth, 2001) demonstrates that people can be highly accurate in detecting and using correlational relationships to make predictions. Research in illusory correlation (e.g., Chapman & Chapman, 1969), on the other hand, shows that people may infer relationships when none are present. In those paradigms, as well as in the extensive research in causal inference (e.g., Cheng & Novick, 1990), participants are typically presented arrays of $x,y$ observations with variance in both the $x$ and $y$ variables, on which the inference of
correlation is to be based (Shanks, 2004). Thus, stimulus environments are designed according to formal models based on sample variances and covariances. Theoretical accounts of covariation assessment are shaped by the same formal models (Gigerenzer, 1991; Gigerenzer & Sturm, 2007).

There are exceptions to the generalization that experimenters present complete data arrays to participants in studies on the perception of correlation. A few investigators have incidentally asked participants to infer contingency in situations in which at least one of the purported variables had no variance. Berndsen, McGarty, van der Pligt, and Spears (2001, Experiment 1) presented their participants such incomplete data matrices, yet their participants made covariation judgments that varied systematically with the indeterminate data. Clement, Mercier, and Pasto (2002) presented participants with samples of paired variables randomly sampled without replacement from a finite population, which resulted in some indeterminate data sets in which at least one variable had zero variance. Nevertheless, their participants validly rated the strength of relationship shown in these indeterminate conditions, and their ratings corresponded rather well with the correlations in the populations from which the data had been sampled.

White (2000) and Kelley, Anderson, and Doherty (2006) set out to investigate contingency assessment with indeterminate matrices. White’s participants were given data sets, each of which was characterized by $n$ observations of three dichotomous variables. One was an effect variable with levels, “effect present” or “effect absent.” Another was correlated with the effect and had the levels “present” or “absent.” A third, referred to as a “common factor,” had no variance: it was always “present” and thus bore an indeterminate correlational relationship to the effect. The results indicated that people were prone to identify the common factor as a cause more often than the positive covariate, in spite of the fact that the common factor had zero variance. White maintained that participants formed a focal set (a concept also described by Cheng & Novick, 1990) in which they considered only cause-present data, even when both cause-present and cause-absent data were available. Kelley et al. (2006) tested White’s hypothesis concerning a focal set. Participants were asked to rate the strength of the causal effect of $x$ on $y$, in a number of samples. Each sample contained nine pairs of binary, $x,y$ data, but some samples were correlationally indeterminate in that only one level of $x$ (or one level of $y$) was represented. Participants systematically drew inferences from these indeterminate data.

The two studies most directly relevant to the present research are Griffiths and Tenenbaum (2005) and McKenzie and Mikkelson (2007), both of which utilized a Bayesian approach to understand the detection of covariation. Both found evidence that people appreciate the power of a single observation to inform causal judgment. In Griffith and Tenenbaum’s Experiment 2, participants were asked to rank a variety of data samples in terms of the degree to which the data supported the hypothesis of a causal relationship between two stimulus variables. A sample containing a single datum was ranked higher than a sample containing no data, which was taken as evidence for their Bayesian causal support model. In the second task of McKenzie and Mikkelson’s Experiment 1, participants were presented two cases, one with a single observation in cell A (present-present) and one with a single observation in cell D (absent-absent) of the contingency table. They were asked
which provided better evidence for the relation in question. Participants had no difficulty with such incomplete information, with their choices being dependent on the relative rarity of the observations, which had been established experimentally by a set of learning trials.

Participants’ behavior in the investigations of both Griffiths and Tenenbaum (2005) and McKenzie and Mikkelsen (2007) suggests that they understood the value of the single datum, in that they correctly ranked the evidential value of sample containing a single observation. Griffiths and Tenenbaum (2005) and McKenzie and Mikkelsen (2007) thereby established the theoretical point that a Bayesian view of covariation assessment is tenable. Note that in the experimental procedures employed by these investigators, the task was to choose a datum given a hypothesis. In contrast, the tasks in the present studies required participants to choose a hypothesis given a datum. Thus, the present studies expand the scope of the inference task to assess people’s ability to choose the correct correlational hypotheses given a datum, rather than choose the correct datum given a hypothesis. The research presented herein also differs from that previously reported in that the evidence variables are continuous rather than dichotomous. In contrast to dichotomous variables, the use of continuous variables allows presentation of a datum with 0, 0 coordinates, thus allowing the assessment of Bayesian inference when the datum does not provide any information about the directionality of correlation.

In sum, the empirical research on the detection of covariation indicates that people may correctly infer covariation from complete data arrays that reflect the data required by traditional normative models, that people may also infer correlations that do not exist according to such models, that people may infer covariation from data that are inadequate according to such models, and that people can appreciate the evidentiary value of a single datum. We emphasize that there is a nearly universal tendency for investigators to regard deviation from the normative model for covariation assessment as error, the two studies just described being clear exceptions.

1.2. Two approaches to the detection of contingency

In the above pages, we have distinguished between approaches to the assessment of covariation based on sample variances and covariances versus Bayes’ theorem. We have been careful not assert that one or the other approach is wrong. They are different. Griffiths and Tenenbaum (2005, p. 348) distinguish between two components of causal induction, structure learning and parameter estimation, the implications of which distinction are germane to the assessment of covariation in general. They posit that statistical hypothesis testing applies to structure learning – whether a relationship exists or not – whereas measures of effect size apply to parameter estimation. “As a consequence,” they go on to note, “there is no logically necessary relationship” between them. McKenzie and Mikkelsen (2007) drew the same distinction, but with different terminology. They used the term descriptive to label traditional statistical methods for the study of covariation assessment and contrasted it with their own Bayesian inferential approach. Hence, the terms structural, inferential, and Bayesian are contrasted with parameter estimation, sample variance-based, and descriptive, in somewhat different contexts. In the research to be presented, we ask
participants to draw an inference of covariation from a single datum, in the light of back-
ground knowledge, and to model their inference behavior with Bayes’ theorem.

1.3. Bayesian inference

Bayes’ theorem has been widely used as a model for successful adaptation to many task environments. The present paper is not proposing Bayes as the sole model of adaptation in environments calling for inference. Rather, we take as a fundamental feature of Bayesian inference that people are testing substantive hypotheses against one another, in contrast to classical statistical inference which posits that substantive hypotheses are tested indirectly by reference to a null hypothesis. This use of Bayes’ theorem is both as an analytical tool and as a theory of inference (Gigerenzer, 1991; Gigerenzer & Sturm, 2007). Note that a Bayesian approach was taken in developing the research to be reported, but the results are consistent with any frequentist estimator based on the likelihood function.

Here we provide a brief, formal argument that specifically addresses the question of inference. Suppose a single observation \((Z_x, Z_y)\) is taken from a bivariate normal distribution with zero means and unit standard deviations. The sampling density of this random observation is given by

\[
f(Z_x, Z_y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(Z_x^2 - 2\rho Z_x Z_y + Z_y^2)\right\},
\]

where \(\rho\) is the population correlation coefficient. Suppose one actually observes \((Z_x, Z_y)\). Then this observation is fixed and one can draw inferences about the correlation coefficient by its likelihood function,

\[
L(\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(Z_x^2 - 2\rho Z_x Z_y + Z_y^2)\right\},
\]

where \(-1 \leq \rho \leq 1\).

One method of estimating \(\rho\) is based on the idea of maximum likelihood; essentially one estimates the parameter by the value of \(\rho\) that maximizes the likelihood function over the space of allowable parameter values. For the sake of clarity, where predictions are framed in the remainder of the paper, we report posterior probabilities in lieu of likelihoods. The conversion into posterior probabilities was accomplished using the likelihood associated with \(\rho = .99\) instead of the unbounded likelihood associated with \(\rho = \pm 1\). Fig. 1 illustrates the function relating the posterior probability of \(\rho\) as a function of the value of \(\rho\), as \(\rho\) varies from \(-1\) to \(+1\), given \(Z_x, Z_y = 2, 2\).

In the present paper, we test directly what may be seen by many non-Bayesians as a radical proposition that people can infer the presence of correlation in a population based on a single data point. Specifically, four experiments test whether people will infer a probabilistically valid, correlational relationship between two variables based on a single, randomly drawn \(x, y\) observation and on knowledge of the characteristics of the univariate distributions of otherwise unspecified \(x\) and \(y\) variables.
The claim that people can draw probabilistically valid inferences from a single observation might seem radical to people knowledgeable in traditional statistics. It does not seem radical at all to a Bayesian. In a bookbag-and-poker-chip task (Edwards, 1968; Phillips & Edwards, 1966), a person is given the task of inferring from which of two bags a single chip has been chosen, given knowledge of the bag compositions, that is, the population proportions. Almost universally, people draw correct inferences about the source of the chip. Typically, in bookbag-and-poker-chip tasks, people’s posterior probabilities are less than the appropriate Bayesian posteriors, a finding that prompted the stream of research on “conservatism” (Edwards, 1968). By Bayesian reasoning, it is clear that one could as easily demonstrate successful inference in choosing between two populations of differing means and standard deviations based on a single element. And given that a correlation coefficient is a mean cross-product, we predict that people will successfully draw inferences about population correlations based on a single datum, again provided univariate means and standard deviations are known.

2. Experiment 1

Participants were sampled from three populations marked by high levels of statistical expertise. One sample was composed of faculty members from quantitatively oriented university departments. The other two samples were from societies in behavioral sciences that rely heavily on statistical reasoning and analysis.

2.1. Method

2.1.1. Group 1

2.1.1.1. Participants: All persons named in the university directory as having faculty rank in the Departments of Psychology, Sociology, Mathematics, Applied Statistics and Operations Research, Accounting, and Management Information Systems were sent an e-mail questionnaire (see Fig. 2). A total of 123 faculty members were solicited to serve as participants.
2.1.1.2. Materials: The one-item questionnaire is shown in Fig. 2. The order of the response options was counterbalanced.

2.1.1.3. Procedure: The faculty members in the named departments were individually e-mailed questionnaires. Three of the senders, two faculty members in Psychology and one in Mathematics, were identified by name only, but many of the recipients would have recognized one or more of the senders by name. Of the 123 sent, 24 were completed and returned.

Fig. 2. The electronic mail message sent to faculty in the departments named.
2.1.2. **Group 2**

2.1.2.1. **Participants:** The 116 members of the Brunswik Society were solicited via electronic mail addressed to the society e-mail list.

2.1.2.2. **Materials:** After a brief request for participation in a research project, respondents were referred to a Web site for the questionnaire. The one-item questionnaire shown in Fig. 2 was used, but with a different salutation. Technical limitations on the procedure for electronic administration did not allow counterbalancing of the order of the response options.

2.1.2.3. **Procedure:** The procedure was as above, except that for this group the message was sent only in the name of the first author, who is a longstanding member of the Brunswik Society. Of the 116 sent, 43 were completed and returned.

2.1.3. **Group 3**

2.1.3.1. **Participants:** The 948 members of the Society for Judgment and Decision Making were solicited via electronic mail addressed to the society e-mail list.

2.1.3.2. **Materials:** After a brief request for participation in a research project, respondents were referred to a Web site for the questionnaire. The one-item questionnaire shown in Fig. 2 was used, but with an appropriate salutation. The response options were not counterbalanced.

2.1.3.3. **Procedure:** The procedure was as above. For this group the message was also sent only in the name of the first author, who is a longstanding member of the Society. Of the 948 sent, 225 were completed and returned.

After all of the data had been analyzed, an explanation of the results in terms of Bayes’ theorem was sent to all participants in Group 1 and to all members of the Brunswik Society and the Society of Judgment and Decision Making.

2.2. **Results**

Table 1 shows the results for all three groups. Of the 24 respondents in Group 1, 19 chose the option corresponding to the population correlation $\rho = .50$, $\chi^2(1) = 8.17$, $p < .01$. Response frequencies to the two forms created by the counterbalancing did not differ. The majority of Group 2 participants’ frequencies of selection favored the correct selection (as defined by Bayes’ rule) as well, but not significantly so. Group 3 respondents favored the correct selection by a ratio a little greater than 2:1, and the large number of respondents ($n = 225$) was such that the resulting $\chi^2$ was 30.62, $\chi^2(1) = 30.62$, $p < .01$. The value of $\chi^2$ was, of course, still larger for the total number of respondents, $\chi^2(1) = 38.48$, $p < .01$. 
2.3. Discussion

The data show that a significant proportion of quantitatively sophisticated people can draw valid generalizations from a sample of \( n = 1 \), given distributional knowledge or reasonable assumptions thereabout. Some of our participants did understand the Bayesian underpinnings of the task; one e-mailed back the correct likelihood ratio on getting the original request, and two did so after the Bayesian explanation was provided. While some participants undoubtedly learned in their statistical training that such inferences can be made, it appears highly unlikely that formal statistical training is responsible for very many of the correct responses, in that standard statistics texts almost universally stress the need for variance-based conclusions. In the terminology of classical statistical inference, the inferences just described are based on 0 degrees of freedom, again stressing the caveat that such inferences can be made only given knowledge, either explicit or implicit, of the characteristics of the univariate distributions.

Clearly, one cannot generalize the claim that people can draw probabilistically valid conclusions from a single observation to all possible \( x, y \) values. Nor can one generalize broadly from the performance of such statistically sophisticated participants. The assessment of the generalizability of the assertion that people can infer covariation from an \( n = 1 \) requires providing a variety of respondents with a variety of different \( x, y \) pairs and response modes.

3. Experiment 2

This next experiment was designed to provide some generalizability across situations and populations. It posed a somewhat more demanding task on the participants, who were less advanced in their careers than those in Experiment 1, but were nevertheless sufficiently statistically sophisticated to understand problems framed in a way similar to that in Experiment 1. Specifically, graduate students and advanced undergraduates were asked to assign subjective posterior probabilities to each of five equally likely populations varying in \( \rho \), given a variety of single data points. The prediction was that participants would be able to use those single data points to assign valid probabilities to the several populations, each characterized by a different \( x, y \) correlation.
3.1. Method

3.1.1. Participants

Forty-three graduate students in the Department of Sociology were solicited to serve by placing packets with a call for participants in individual departmental mailboxes. In addition, 20 students in an undergraduate class of advanced statistics in the Psychology Department were solicited in class to serve in the study. Packets consisting of the experimental materials were distributed at that time.

3.1.2. Materials

The packets had 10 pages plus an informed consent page. The first page identified the authors and opened with the assertion: “It can be shown that, in principle, one can learn something about the correlation between two normally distributed variables from a single randomly sampled \(x, y\) observation, provided the means and standard deviations of both \(x\) and \(y\) are known.” The instructions went on to note that we already had data from university faculty and from two scientific societies that intuition rather than calculation was involved, and that the student would receive $10 on completion of the task. Instructions were also given as to how to return the consent form and packet so as to preserve anonymity.

The task instructions per se were conveyed via a sample problem on a second page. Participants were instructed to suppose that we were going to draw a datum at random from a population with an unknown \(x, y\) correlation, and that there were five equally likely population correlations, \(-1.0, -0.50, 0, +0.50\) and \(+1.0\). They were further informed that before the \(x, y\) coordinates of any datum were known, the probability that each of the above correlations had been the source of the datum was 20%. A Cartesian coordinate system labeled only \(Z_x\) and \(Z_y\) was shown with a data point that was patently high positive on both \(Z_x\) and \(Z_y\), with the assertion that “the probabilities for the various population correlations would not be equal.” Finally, participants were told that the problems were independent of one another, and that the percentages that they assigned to each should sum to 100%.

The packets had nine problems, one per page. The varying loci of the data points constituted the nine problems. There were four problems with \(Z_x, Z_y\) values of \(+1.8, +1.8, -1.8, +1.8, -1.8\), and \(+1.8, -1.8\), and four corresponding problems with \(0.9\) in lieu of \(1.8\), as well as one problem with \(Z_x, Z_y\) values of \(0.0\). The problem pages were assembled in a Latin Square design, but with self selection such a design was not reflected in the actual returns. The participants were told simply that “\(x\) and \(y\), in the population, have a mean of 0 and a standard deviation of 1.0.”

The participants’ judged probabilities for the \(Z_x, Z_y\) values of \(+1.8, +1.8, -1.8, +1.8, -1.8, -1.8\), and \(+1.8, -1.8\), for the corresponding problems with \(Z\) values of \(±.9\), and for \(0\) can be evaluated by comparison with the posterior probabilities, shown in Fig. 3. Note that the function for the data point \(0, 0\) is U-shaped. The posterior probabilities shown in Fig. 3 constitute the predicted responses for Experiment 2. At best, a modest difference in judged probabilities between the points with \(Z\) scores of \(±1.8\) and those with \(Z\) scores of \(±.9\), as well
as pronounced quadratic functions for the relation between judged probabilities and the values of $q$, are predicted, based on the figures.

3.1.3. Procedure

The Sociology participants returned the consent form and the response pages to different envelopes in the Sociology Department office. The Psychology undergraduates returned the consent form to one author’s mailbox, and the response form to a second author’s mailbox. Of the 43 Sociology students solicited, eight returned the form. Of these, one failed to follow instructions, and gave 20% as the probability for all five population correlations for all loci. Of the 20 Psychology students solicited, 13 returned the form, 2 of whom gave 20% as the probability for all five populations. Hence, there are 18 participants in Experiment 2, each one having solved nine problems.

Fig. 3. Posterior probabilities of $\rho$, given the single data points in each of the four quadrants and for the 0, 0 data point, in Experiment 2.
3.2. Results

Table 2 reports the linear and quadratic correlations between the five possible values of $\rho$ and the mean percentages assigned to those values. A few participants appeared to reverse the scale, which resulted in a function relating average percentage to possible population correlation that was curvilinear. For example, for the $-1.8,-1.8$ point locus, the percentages assigned by four participants were inconsistent with the sign of the slope of the function relating posterior probability to the population $\rho$. However, of the 144 individual slopes calculated for the nonzero data loci, only 15 were of the wrong sign, and some had a 0 slope. Fig. 4 shows the regressions of the mean percentages on the possible $\rho$ values separately for each data point locus. The curvilinearity created largely by the scale reversal for a small number of participants is evident in the graphs.

3.3. Discussion

Inspection of the individual participants’ data reveals that only one participant gave a complete distribution of percentages that was qualitatively similar to that called for by Bayes’ theorem, assigning 50% to population correlations of $-1$ and 1 at the 0,0 locus. However, the mean percentages for the data points other than the one at the origin are highly consistent with the Bayesian predictions. The single observations in each of the four quadrants were regarded by most participants as evidentiary. The slopes of the mean probability estimates in Fig. 4 for the data loci in the four quadrants are not sufficiently extreme, but they are qualitatively consistent with the Bayesian posteriors. The function for the 0, 0 point is qualitatively inconsistent with the predictions. It takes considerably more thought to recognize that, counterintuitively, the point 0, 0 is less likely given $\rho = 0$ than given any $\rho \neq 0$. Participants evidently considered the 0, 0 datum to be no more informative than no data at all, given that they tended to assign 20% to each of the five possibilities.

Table 2
Linear and quadratic correlations ($r$ and $R$) between the five possible values of $\rho$ and the mean percentages assigned to those values in Experiment 2

<table>
<thead>
<tr>
<th>Point Locus</th>
<th>$r$</th>
<th>$p$</th>
<th>$R$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1.8, +1.8</td>
<td>.82</td>
<td>&lt;.10</td>
<td>.98</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>+.9, +.9</td>
<td>.88</td>
<td>&lt;.05</td>
<td>.99</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>−1.8, +1.8</td>
<td>−.82</td>
<td>&lt;.10</td>
<td>.99</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>−.9, +.9</td>
<td>−.92</td>
<td>&lt;.05</td>
<td>.99</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>−1.8, −1.8</td>
<td>.63</td>
<td>ns</td>
<td>.98</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>−.9, −.9</td>
<td>.71</td>
<td>ns</td>
<td>.99</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>+1.8, −1.8</td>
<td>−.69</td>
<td>ns</td>
<td>.99</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>+.9, −.9</td>
<td>−.88</td>
<td>&lt;.10</td>
<td>.99</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>0, 0</td>
<td>.08</td>
<td>ns</td>
<td>.44</td>
<td>ns</td>
</tr>
</tbody>
</table>
Fig. 4 shows that the participants appear to treat the evidentiary impact of $Z$ values of 1.8 similar to the impact of $Z$ values of .9 with regard to the presence of covariation, but as Fig. 3 indicates, the differences in diagnosticity between them are not great until extreme values of $\rho$ are reached.

Experiment 2 involved inferences concerning a distribution of possible population correlations from a single data point, given explicit knowledge of the univariate distributions and of the prior probabilities of the possible populations. Participants’ inferences in eight of the
nine problems clearly tended to be qualitatively correct. Experiment 2 was similar in conceptualization to Experiment 1 but varied in ways that permit some generalization across tasks and populations and that pinpoint an exception thereto. The differences include level of experience, the variety in loci of the data points, the nature of the dependent variables, and the mode of data presentation collection (paper and pencil vs. Internet). The results of Experiment 2 are, for the most part, consistent with and extend those of Experiment 1. Further discussion of the lack of correspondence with the Bayesian prediction found in Experiment 2 will be pursued in the general discussion.

4. Experiment 3

The experiment was designed to generalize our findings to the behavior of statistically naive participants. This entailed framing the problems in a way that was primarily verbal and concrete rather than statistical and graphical. Each participant was given a single problem in which the task was to decide whether a single $x$, $y$ data point came from a correlated or an uncorrelated population. Thus, the decision involved choosing between two concrete alternatives, that is, two populations, rather than between two possible different numerical characterizations (correlation coefficients) of one population. The following is an example of one version, tall-high, of the problem. The boldfaced text indicates the location of alternative phrases that the researchers used to manipulate the location of the data point (see Table 3).

Imagine that there are two fraternities on campus. Each fraternity has 100 members, and the two fraternities are equal with respect to the members’ heights and GPAs (grade point averages). However, the relationship between height and GPA is different for the two fraternities. In Gamma Gamma Phi, shorter members tend to have lower GPAs, whereas taller members tend to have higher GPAs. In Kappa Kappa Mu, the GPAs of shorter members are no higher or lower than the GPAs of taller members.

Suppose you meet a fraternity member on campus. He is 6 feet 2 inches tall, which is much taller than most members of Gamma Gamma Phi, and also much taller than most members of Kappa Kappa Mu. He has a 3.7 GPA, which is much higher than

<table>
<thead>
<tr>
<th>Condition</th>
<th>Height</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tall-high</td>
<td>6 feet 2 inches</td>
<td>3.7</td>
</tr>
<tr>
<td>Short-high</td>
<td>5 feet 4 inches</td>
<td>3.7</td>
</tr>
<tr>
<td>Short-low</td>
<td>5 feet 4 inches</td>
<td>2.3</td>
</tr>
<tr>
<td>Tall-low</td>
<td>6 feet 2 inches</td>
<td>2.3</td>
</tr>
<tr>
<td>Average-average</td>
<td>5 feet 9 inches</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 3
Description of the experimental conditions of Experiments 3 and 4

Note: Details of the scenario description are provided in the main text.
GPA, grade point average.
most members of Gamma Gamma Phi, and also much higher than most members of Kappa Kappa Mu.

Given that this student is relatively tall and has a relatively high GPA, which fraternity is he more likely to be a member of? Place a check mark beside your selection, below.

___ Gamma Gamma Phi

___ Kappa Kappa Mu

The preceding example problem is a concrete analog of the abstract problem employed in Experiment 1. Thus, the single data point, in the present problem, describes the height and GPA of a hypothetical individual who belongs either to a fraternity in which height and weight are correlated, or to one in which height and weight are uncorrelated. The description of the data point includes information about the relation of x and y to the distributions of x and y in the populations (i.e., ‘He has a 3.7 GPA, which is much higher than most members of Gamma Gamma Phi.’). Thus, the description accomplishes informally what was done formally in Experiments 1 and 2.

4.1. Method

4.1.1. Participants

One hundred eighty students participated in partial fulfillment of the requirements for an introductory psychology course.

4.1.2. Design and materials

The location of the data point was a between-participants independent variable and was varied across five levels: tall-high (i.e., tall, with a high GPA), short-high, short-low, tall-low, and average-average. For half of the participants Gamma Gamma Phi was described as the correlated population; for the other half it was described as the uncorrelated population. Thus, the position of the normatively correct response option was counterbalanced (Gamma Gamma Phi was always the first option; Kappa Kappa Phi was always the second).

4.1.3. Procedure

Two-page packets containing an informed consent form and a single problem were distributed to students near the end of the class period. Each student had the option to participate or not. Those who chose to participate returned a completed packet to the instructor, prior to the end of the class period.

4.2. Results

Table 4 shows the numbers of participants at each level who chose each of the two response options. The table also indicates the numbers of normatively correct responses per
level, and the results of binomial tests to determine whether the proportions of normatively correct responses were significantly different from .5. Of all 180 participants, 130, or 72% (p < .001, for the binomial test), made a choice consistent with the fraternity indicated by Bayes’ theorem. If participants in the average-average scenarios are excluded, the percentage correct rises to 80%. In all conditions except the average-average condition (see Table 3), participants exhibited a significant tendency to choose the normatively correct answer.

4.3. Discussion

Table 2 and Fig. 4 clearly show that the judgments tended to reflect Bayesian predictions but that participants did not have the requisite intuition to make correct inferences given the average-average datum. Inspection of Table 3 reveals that the single observation available to a participant was either at the origin, or in one of the four quadrants of the implied Cartesian system. The data points in the first and third quadrant favor the positively correlated population intuitively as well as mathematically. The data points in the second and fourth quadrant intuitively favor a negative correlation, but that is not an option available to the participant. These data points are far more likely given a population with $\rho = 0$ than one that has a positive correlation. Participants were not given a specific value of the non-zero correlation, but for the sake of illustration, let us assume that the alternative under test is $\rho = .50$. If we assume that a height of 6 feet 2 inches corresponds to a $Z$-score of +1.5 and a GPA of 2.30 corresponds to a $Z$-score of −1.5, then the likelihood ratio in favor of $\rho = 0$,.
assuming that the alternative $\rho = .50$, is about 8.2:1. Under these assumptions, the likelihood ratios for the data points in the first and third quadrants and at the origin, in favor of $\rho = .50$ over $\rho = 0$, are 2.4:1 and 1.2:1, respectively.

The findings of Experiment 3 are consistent with those of Experiments 1 and 2 in that participants’ correlational inferences tended to reflect Bayesian probabilities. As in Experiment 2, there is an exception, specifically when the scenarios described the target person as average in both respects. Thus, Experiment 3 demonstrates that the ability to draw probabilistically correct inferences is not limited to experts making judgments about abstract, mathematical data. Rather, even statistically naïve participants making qualitative judgments about concrete situations tend to draw probabilistically valid correlational inferences from a single datum, as long as the datum itself provides some information about directionality.

5. Experiment 4

In Experiment 3, participants were asked to make a forced choice between the two hypothetical fraternities. A question remains as to whether participants who intuitively make the correct inference would, when asked directly, consider such an inference possible. A further question is whether a participant’s judgment concerning the sufficiency of the datum to support inference is related to subsequent performance on the inference question. Experiment 4 was a constructive replication of Experiment 3. It presented precisely the same scenarios, except for a prior question concerning the sufficiency of the information.

5.1. Participants and procedure

One hundred fifty-nine students from an introductory psychology course participated, with the questionnaires being distributed and completed toward the end of a class period.

5.2. Design and materials

These were as in Experiment 3, except that the response format was as follows:

In your opinion, has enough information been provided so that it is possible to determine which fraternity this student is more likely to be a member of. (Place a check mark beside ‘‘yes’’ or ‘‘no,’’ below.)

_____ yes

_____ no

Regardless of how you answered the previous question, please use your best judgment to decide which fraternity the student is more likely to be a member of. (Place a check mark beside your selection, below).
5.3. Results

A breakdown of the respondents by scenario type and response to the first question is provided in Table 4. Note that the significance tests in this section, and in Table 4, are binomial tests that assess whether a proportion is significantly different from .5. Of the 159 participants, only 38 indicated that they thought that enough information had been provided to answer the question concerning fraternity membership. Of all 159 participants, 117, or 74%, made a choice consistent with the fraternity indicated by Bayes’ theorem, \( p < .001 \). Of the 121 participants who responded that there was insufficient evidence, the percentage responding in a fashion consistent with Bayes’ theorem is 67%, \( p < .001 \). Of the 38 respondents who checked that there was sufficient information, fully 95% responded in a fashion consistent with Bayes’ theorem, \( p < .001 \). The 38 participants who answered “yes” to the sufficiency question performed better on the fraternity question than did the 121 participants who responded that there was insufficient information, 95% correct to 67% correct, \( p < .01 \). As in Experiment 3, participants in the average-average condition did not systematically differ about which fraternity was more likely, with an exactly even split between the answer dictated by Bayes and the alternative answer.

5.4. Discussion

The findings of Experiment 4 are consistent with those of Experiments 1, 2, and 3 in that participants’ inferences tended to reflect Bayesian probabilities, except when the scenarios described the target person as average on both dimensions. The results of Experiment 4 replicate Experiment 3 in detail: only the differences in frequencies in the short-low condition failed to achieve statistical significance. As in Experiment 2, participants’ responses to the average-average datum were inconsistent with the Bayesian predictions.

As noted, participants who responded “yes” to the question about whether the information provided was sufficient to draw an inference were almost unanimously correct. Such a difference was unexpected. This investigation was not designed to explore the reason for this difference, but one speculation is that the sufficiency question engaged metacognitive processing at a relatively analytical level, whereas the fraternity question may have engaged more intuitive processes. For some participants, no formal or analogical means of answering the fraternity question may have come easily to mind, and the response to the sufficiency question tended to be “no.” Of these who answered “no,” some may have subsequently just guessed on the fraternity membership question, while others came to terms with the issue and relied on a more intuitive process, specifically representativeness (Tversky & Kahneman, 1974). In this and some other domains, representativeness can be conceptualized
as instantiating Bayesian inference. By representativeness, we mean that the participants may have noticed a similarity relation between the target person and typical members of a particular fraternity and—except for the average-average condition—made an appropriate response consonant with Bayes’ theorem.

6. General discussion

6.1. Bayesian rationality

In general, the results of the four studies described herein are qualitatively consistent with the Bayesian views of covariation assessment put forth by Griffiths and Tenenbaum (2005) and by McKenzie and his colleagues (e.g., McKenzie, 2005; McKenzie & Mikkelson, 2007), as well as with the older Bayesian research associated with Edwards (1968) and others. Experiments 1 and 2 had sophisticated participants draw an inference concerning a datum drawn from a single bivariate population. In Experiments 3 and 4, unsophisticated participants inferred whether a datum was drawn from one population or another.

The Bayesian approach bearing most directly on our results is that of Griffiths and Tenenbaum (2005), described in the introduction. The present studies were not designed as a test of their causal support model, which was framed in terms of causal inference, but they complement their empirical results and generalize their conclusion in three ways. First, they generalize the conclusion to the assessment of association framed in noncausal terms. Second, they generalize the conclusion to the case in which the participants’ conclusion is about the hypothesis given a datum (rather than about a datum given a hypothesis). Finally, the present studies generalize Griffith’s and Tenenbaum’s conclusion to a wide range of cases involving the correlation between continuous variables—specifically, cases involving a directional (i.e., non 0, 0) datum.

Griffiths and Tenenbaum’s finding that participants ranked the samples containing one data point as providing greater causal support than those containing no data, taken together with the results of McKenzie and Mikkelson (2007) and the four experiments presented herein, provide converging operations (Garner, Hake, & Eriksen, 1956) that people can draw probabilistically valid inferences from a single observation. Our results may be considered, then, an extension of Griffith and Tenenbaum’s Bayesian causal support model to the more general domain of covariation assessment. This application of a Bayesian approach is clearly as a rational, functional model rather than a mechanistic, representational one (Brighton & Gigerenzer, 2008; Tenenbaum & Griffiths, 2001).

Typically, theories of how people infer causation or covariation, such as $\Delta P$ and its variants for dichotomous data, or the Pearson $r$ and its variants, are predicated on arrays of bivariate observations. Griffiths and Tenenbaum (2005) note that “$\Delta P$ and causal power are both undefined when there are no trials on which cause was absent. This is a problem for a general theory of the assessment of covariation, as people readily make causal judgments under such circumstances” (p. 364). An analogous argument can be made for the continuous case. The Pearson $r$ is undefined when only a single observation is available, yet people
readily make covariation judgments from single cases. A primary implication of these results for our understanding of everyday adaptive cognition is that when someone “jumps to a conclusion” from a single observation, the conclusion may be wrong, but the inference process may well be mathematically warranted.

6.2. Departure from Bayesian predictions

Our use of continuous variables provided some insight into a boundary condition of the Bayesian approach, in that the Bayesian predictions were not verified in the cases in which the datum provided no directional information about the putative relationship, namely the 0, 0 datum in Experiment 2 and the average-average datum in Experiments 3 and 4. These conditions are unique to the present studies (Griffiths & Tenenbaum, 2005; and McKenzie & Mikkelson, 2007 used dichotomous rather than continuous stimulus data) and cannot be explained in a strictly Bayesian fashion. This departure from Bayesian predictions in Experiments 3 and 4 can be reconciled with the predictions in an ad hoc fashion if one assumes that the participants understood the $\rho > 0$ response option to indicate a positive but not necessarily strong population correlation. This seems reasonable given that the experimental materials (in Experiments 3 and 4) described the population correlations qualitatively rather than quantitatively. Thus, given that the function for the posterior probability of $r$ given 0, 0, in Fig. 3 is relatively flat until values of $\rho$ in excess of ±.9 are reached, it is no surprise that the very small differences in posterior probabilities given 0, 0 were not reflected in behavior.

No similar ad hoc account can plausibly explain the behavior in the 0, 0 condition of Experiment 2. In that experiment, the various categories of population correlations, including the $-1$, and 1 categories were quantitatively explicit, leaving no room for participants to interpret extreme population correlations as anything other than extreme. Hence, it appears to us that the failure of participants to behave at least qualitatively according to the predictions is a real exception to the general proposition advanced herein and by Griffiths and Tenenbaum (2005) and McKenzie and Mikkelson (2007).

Why is behavior in the 0, 0 condition an exception? A plausible speculation is that only when observations deviate from the norm does the request for any inference about possible relationships make sense. People typically try to make sense of things that are out of the ordinary, but the 0, 0 condition conveys that the observation is average on both dimensions, and therefore entails no deviation from the norm. Hence, the participants may have no intuitive model for inference when faced with an experimental request for such an inference, and, as noted above, the appropriate analytical model is extremely counterintuitive. Consequently, our participants in Experiment 2 (and even those in Experiments 3 and 4) may have adopted a strategy of responding randomly when faced with a data point at the origin. This exception leaves us with the general conclusion that when people draw inferences from even very limited data, people’s judgments conform qualitatively with those dictated by the rational model proposed, except in the situation in which the datum provides no information about the direction of the inference.

A reasonably direct way to address such a conjecture would be to constructively replicate Experiment 2 with a parametric manipulation of the Zx, Zy variable. If the judgments vary
qualitatively with Bayesian posteriors for values very close but not exactly equal to 0, 0, but are still decidedly non-Bayesian when $Z_x, Z_y = 0$, such a result would be further evidence for a Bayesian approach, with a caveat about the need for directionality. A related issue is whether the 0, 0 exception would hold for the case of causal inference. A straightforward test of whether it does or does not would be a replication of Experiment 2, but framed concretely as a causal scenario instead.

In summary, though the 0, 0 condition of Experiment 2 is similar to the average-average conditions of Experiments 3 and 4, the average-average conditions occur in the context of qualitatively defined, rather than quantitatively defined inference alternatives. Thus, performance both in the 0, 0 condition of Experiment 2 and the average-average conditions of Experiments 3 and 4 may be attributable to participants using an inappropriate strategy—one of random responding when there is no intuitive model for producing a judgment. However, performance in the average-average conditions of Experiments 3 and 4 may also be explainable as a consequence of participants’ interpretations of imprecise, qualitatively defined response alternatives.

6.3. Broader issues

The concept of representativeness (Tversky & Kahneman, 1974) has been sharply criticized as being too vague to count as an explanation (Gigerenzer, 1996). We do not take issue in general with Gigerenzer’s critique, but we believe that in the situation in which our participants found themselves (related to the sufficiency question of the Experiment 4 procedure), and in situations in everyday life in which unusual observations are made, representativeness can be formalized in terms of Bayes’ theorem. Finally, Brunswik’s (1956) conception of vicarious functioning posited that a person can arrive at the same goal in multiple ways, depending on the ecological constraints. These results are very much in line with that conception. In light of the concept of vicarious functioning, we take the findings presented to mean not that psychological theories of covariation inference based on sample variances and covariances are wrong, only that they are incomplete to the extent that the nature of the ecology is not explored and varied. Indeed, any theory of the inference of covariation that does not take into account the mathematical ecology in which the inference is drawn must be incomplete.

Acknowledgment

This research was supported by National Science Foundation grant 0423825.

References


