
SAMPLE SIZE AND THE DETECTION OF CORRELATION: A SIGNAL DETECTION ACCOUNT

Richard B. Anderson
Michael E. Doherty
Neil D. Berg
Jeff C. Friedrich

BOWLING GREEN STATE UNIVERSITY

Richard B. Anderson, Dept. of Psychology
Bowling Green State U., Bowling Green, OH 43403
Phone: (419) 372-9908, FAX: (419) 372-6013
E-mail: randers@bgnet.bgsu.edu
Abstract

Simulations examined the hypothesis that small samples provide better grounds for inferring the existence of a population correlation than do large samples. The simulations employed a signal detection methodology in which samples of 5, 7, 10, 15, or 30 data-pairs were drawn either from a population with a correlation of zero or from a population with a correlation greater than zero. When accuracy was computed independently for each level of the decision criterion, there was a criterion-specific small-sample advantage. For liberal criteria, decision accuracy was greater for large than for small samples, but for conservative criteria, the opposite result was obtained. This pattern occurred for intermediate values of the population correlation, as well as for symmetrical (Fisher’s $z$) and skewed sampling distributions. However, there was no evidence for a small-sample advantage when decision accuracy was measured as the area under an ROC curve, as the posterior probability of a hit, or as a function of the false alarm rate. The findings indicate that a small-sample advantage can occur, though only under very limited conditions.
Sample Size and the Detection of Correlation: A Signal Detection Account

Recent work by Kareev (1995, 2000; Kareev, Lieberman, & Lev, 1997) and by Juslin and Olsson (2005) has yielded conflicting views concerning the effect of resource limitation on judgments about relationships between variables: Kareev concludes that correlation detection is enhanced by limitations on sample size, whereas Juslin and Olsson find little support for such a conclusion. In the present paper, standard signal detection theory is used to argue that correlation detection can indeed be more accurate with small than with large samples, but only under extremely limited circumstances.

All judgments—including judgments of correlation—must be made with limited resources; these include limited opportunities for gathering information, as well as limited cognitive capacity for information processing. Intuition suggests that such limits ought to be detrimental to accurate decision-making. Indeed, researchers have long cited capacity-limitation as an explanation for sub-optimal performance (e.g., Miller, 1956; Broadbent, 1958). However, Kareev and his colleagues (Kareev, 1995, 2000; Kareev, Lieberman, & Lev, 1997) have suggested that resource limitations—specifically, limits on samples size—might actually be an asset in the detection of correlation. These conflicting conceptions of the effects of resource limitations reflect a longstanding split concerning fundamental assumptions of the field, first articulated as the “two camps on rationality” by Jungermann (1983), revisited by Doherty (2003), and highlighted in an exchange between Gigerenzer (1996) and Kahneman and Tversky (1996).

Citing work by David (1954) and Hays (1963), Kareev (e.g., Kareev et al., 1997) noted that the sampling distribution of the Pearson correlation coefficient is skewed, and that the amount of skew increases as $n$ decreases (where $n$ is the number of elements contained within each sample). The curves in the top graph of Figure 1 illustrate the phenomenon for a case in which the population correlation is .5, where $n$ is 5, 7, 10, 15, or 30. As $n$ becomes smaller, the distribution of the sample correlations ($r$) becomes progressively more skewed (other details of the figure will be described later). Thus, Kareev et al. (1997) argued that the skew amplifies correlations (when $n$ is small) that might otherwise fail to exceed the threshold for detection.
Kareev demonstrated this correlation-amplification effect, not just mathematically, but also experimentally. In one experiment (Kareev, et al., 1997, Experiment 1), subjects performed a pre-test consisting of a digit span task. Subjects scoring above the median were classified as having high working-memory capacity, while those scoring below the median were classified as having low working-memory capacity. Following the pre-test, subjects performed a prediction task. On each prediction trial, subjects viewed a red or a green envelope, made a prediction about whether the envelope contained the symbol "x" or the symbol "o," and then opened the envelope to reveal the symbol inside (which was either an "x" or an "o"). The researchers were interested in subjects’ estimates of the correlation between the envelope color and the symbol contained within the envelope. These estimates were measured indirectly by computing, for each subject, the correlation between the cues (the envelope colors) and the subject’s predictions. It was assumed that subjects based each of their predictions on their memory for the color/symbol pairings of previous trials, and that subjects with high working-memory capacities would remember a larger number of trials than would those with low working-memory capacity. Thus, it was assumed that the high-capacity subjects based their judgments on large samples whereas those with low-capacity based their predictions on small samples. The results indicated that the match between the derived correlation estimates and the objective correlations was more extreme for low-capacity than for high-capacity subjects, which was interpreted as support for the hypothesis that population correlations are amplified in small samples.

In a second study (Kareev et al., 1997, Experiment 2), the subjects’ task was to view pairs of abstract stimulus attributes—line-lengths and angle-sizes—and to use one attribute to predict the value of the other (some subjects predicted angle-sizes from line-lengths; others predicted line-lengths from angle-sizes). Each subject performed only one prediction trial after having been pre-exposed to a sample of stimulus pairs; the sample size was either greater than, equal to, or smaller than the subject’s working-memory capacity, which had been pre-assessed by means of a digit span task. Some subjects were allowed to view the sample while making their prediction; others had to rely on their memory of the sample. The results showed that when
subjects were forced to rely on memory, their predictions were more accurate when they had viewed smaller rather than larger samples. Yet, this pattern was reversed when the sample remained in view during the prediction judgment: For subjects in this condition, predictions were less accurate with smaller than with larger samples. It is not clear why the data pattern reversed itself as a function of sample viewing condition. Thus, the experimental data provided only mixed support for the hypothesis of a small-sample advantage in correlation detection.

In response to Kareev’s theoretical work, Juslin and Olsson (2005) addressed the question of how false alarms should impact the small-sample advantage, an issue that Kareev (2000) had not explicitly addressed in his simulations. Having a small rather than a large sample might facilitate subjects inferring a population correlation when they should infer a correlation, but might also increase the frequency with which subjects infer a correlation when they should not infer a correlation—possibly nullifying what might otherwise be a small-sample advantage.

To address the question of false alarms, Juslin and Olsson (2005) conducted a set of simulations in which a simulated decision-maker drew a sample of $xy$ pairs from a population, compared the sample correlation to a critical correlation, and on that basis made an inference about the population correlation. In their simulations, Juslin and Olsson employed a non-standard adaptation of signal detection theory. The simulated decision-maker was presented with signal trials and noise trials, but the noise trials were not necessarily drawn from a population in which the correlation was zero. This unusual definition of noise was consistent with (but not identical to) Kareev’s (2000) conception of a usefulness criterion, developed in the context of a task in which a binary variable $x$ is used to predict the level of a binary variable $y$. Given unequal marginal distributions and given small sample sizes, it is possible for the $xy$ correlation to be so weak that prediction accuracy can be maximized by using the base rate instead—that is, by ignoring $x$ as a cue and simply predicting that $y$ will take on the value that it has most frequently taken on in the past. Juslin and Olsson did not use Kareev’s definition, per se, but they did adopt the general idea that a correlation can be non-useful even though it is nonzero. Thus, if $r$ fell below the critical $r$, then Juslin and Olsson (2005) counted the trial as a noise trial—the logic being that $r$, even when it is nonzero, should sometimes be considered too weak to be useful and should therefore be treated by the decision-maker as if it were zero. Alternatively, if $r$
exceeded the critical $r$, the trial was counted as a signal trial. The simulation then proceeded according to the prescriptions of signal detection theory: When the sample $r$ exceeded the critical $r$, it was called a hit if the trial was a signal trial, or a false alarm if it was a noise trial. When the sample $r$ did not exceed the critical $r$, it was called a miss or a correct rejection, for signal or noise trials, respectively. The simulations yielded a clear pattern of better decision-making accuracy (when measured as the proportion of correct decisions, as the hit rate minus the false alarm rate, or as posterior probabilities) for larger than for smaller samples, thus contradicting Kareev’s theoretical claim of a small-sample advantage.

The present work used a simulation methodology to evaluate further the hypothesis of a small-sample advantage in correlation detection. Like Juslin and Olsson (2005), the simulations employed concepts from signal detection theory to examine the impact of false alarms on performance. However, the present simulations used a more traditional application of signal detection theory in which signal populations were defined as having a nonzero correlation, and noise populations were defined as having a correlation of zero. In addition, the present simulations examined correlations between continuous variables as well as between dichotomous variables (previous simulations dealt only with dichotomous variables). Also, to determine whether a small-sample advantage occurs for some settings of the criterion but not for others, the present work explicitly contrasted overall decision-making accuracy (the area under an ROC curve) with criterion-specific accuracy measured as the hit rate minus the false alarm rate.$^{1}$

Simulations with Continuous Variables

Each simulation consisted of multiple repetitions of the following procedure. A statistical software package was used to generate a set of 100,000 pairs $(x$ and $y$) of normally distributed random numbers ($\mu = 0$, $\sigma = 1$). This constituted a noise population in which the $xy$ correlation was zero. Next, a second set of 100,000 normally distributed random numbers was generated ($\mu = 0$, $\sigma = 1$); these were single, unpaired numbers. The numbers (in the second set) were then paired with duplicates, yielding a set of $xy$ pairs in which $x$ was equal to $y$, thus yielding a signal population in which the correlation was 1.0. Next, random-normal error was added to $x$, and to $y$, in both the signal and the noise populations so that the population correlation ($\rho$) for the signal was
reduced to .9, .8, .7, .6, .5, .4, .3, .2, or .1, depending on the variance of the added error. The addition of error to both the signal and noise populations insured that the variances of $x$ and $y$ in the signal population equaled the variances of $x$ and $y$ in the noise population. Of course, the addition of random error to the noise population (in which $\rho$ was always zero) had no systematic effect on the value of $\rho$ for that population.

A sample of 5, 7, 10, 15, or 30 $xy$ pairs was drawn randomly from either the signal or the noise population. Next, the sample correlation ($r$) was computed and then compared to one of a set of critical $r$ values. These criteria ($c$ and $-c$) were set to ±.1, .2, .3, .4, .5, .6, .7, .8, or .9, respectively. If $r$ exceeded $c$ or fell below $-c$, then it was registered as a hit if the sample had been drawn from the signal distribution, or as a false alarm if the sample had been drawn from a noise distribution. The sampling process was repeated so that there were 3,330 samples at each of the nine criterion-settings and for each of the five values of $n$, yielding 45 conditions. The entire process was replicated for each level of $\rho$.

Figure 1 illustrates the computation of the hit and false alarm rates for a case in which signal $\rho$ is .5, and in which $c$ and $-c$ equal .8 and -.8, respectively. The vertical dashed lines show the placement of the decision criteria, where $|c| = .8$. The symbols $H_1$, $H_2$, $F_1$, $F_2$, $L$, and $Q$ denote regions of the sampling distributions. Region $H_1$ defines the set of signal samples that exceed $c$, and $H_2$ defines the set of signal samples that fall below $-c$. Likewise, region $F_1$ defines the set of noise samples that exceed $c$, and $F_2$ defines the set of noise samples that fall below $-c$. Thus, the hit rate is the sum of $H_1$ and $H_2$, divided by the sum of $H_1$, $H_2$, and $L$; the false alarm rate is the sum of $F_1$ and $F_2$, divided by the sum of $F_1$, $F_2$, and $Q$.

There may be a conceptual difficulty with counting $H_2$ samples as hits: Such samples would be consistent with $\rho < 0$ despite the fact that the simulation only included populations in which $\rho = 0$ or $\rho > 0$. However the simulation algorithm's task can be conceptualized as a decision about whether $\rho = 0$ or $\rho \neq 0$. Under this conceptual framework, it would be appropriate to count $H_2$ samples as hits since such samples would indeed be consistent with $\rho \neq 0$. Alternatively, it is possible to conceptualize the task as a decision about whether $\rho = 0$ or $\rho > 0$. In this case a decision should be counted as a hit or as a false alarm if it is consistent with $\rho > 0$, but not if it is consistent with $\rho < 0$. Thus, under this alternative framework, the hit rates ($H$) should be computed as $H = H_1 / (H_1 + H_2 + L)$, and false alarm rates ($F$) should be computed as $F = F_1 / (F_1 + F_2 + Q)$. However, the alternative computations yielded results similar to (and lead to the same conclusions
as) the computations that counted both $H_1$ and $H_2$ as hits, and both $F_1$ and $F_2$ as false alarms (this was true for the dichotomous as well as the continuous simulations). Therefore the results for the alternative computations (in which hits and false alarms involve $H_1$ and $F_1$ only) will not be presented.

It is also possible to compute $H$ as $H_1 / (H_1 + H_2 + L)$, while computing $F$ as $(F_1 + F_2) / (F_1 + F_2 + L)$. However, a problem with this coding is that the hit rate can be smaller than the false alarm rate (particularly when the signal $\rho$ is small)—which can imply negative sensitivity. Therefore, this coding was not adopted in the present analyses (except for Figure A1 in the appendix).

**ROC Curves**

Plots of the ROC curves (i.e., plots of the hit rates as a function of the false alarm rates) were constructed in order to assess the effect of $n$ per sample on the discriminability of signal samples from noise samples. Discriminability (often called sensitivity) was measured as the area under the ROC curve. Figure 2 shows that, contrary to the small-sample advantage hypothesis, discriminability was greater for large than for small samples, except when the signal $\rho$ was sufficiently small so that discriminability was near zero, irrespective of $n$, as shown in the upper left graph of Figure 2.

Criterion-Specific Accuracy

The preceding findings show that when discrimination accuracy is measured without regard to a particular critical value ($c$), performance is more accurate for large than for small samples. However, the findings do not establish whether there might be a criterion-specific small-sample advantage—that is, a small-sample advantage that occurs for particular values of the decision criterion. Figure 1 shows an example of hit and false alarm rates when the criterion, $c$, is set to $\pm .8$. It is possible to measure decision accuracy for $c = \pm .8$, or for any value of $c$, by calculating the hit rate ($H$) and false alarm rate ($F$) and then subtracting the $F$ from...
H. Thus, the simulation results were further analyzed by plotting accuracy \((H - F)\) as a function of \(c\), for various values of signal \(\rho\), as shown in Figure 3. The results indicate that decisions were generally more accurate for large than for small samples, but there was a marked tendency toward greater accuracy for small than for large samples for the combination of midrange values of \(\rho\) and conservative values of \(c\).

Following Juslin and Olsson (2005), decision accuracy was also computed as the posterior probability of a hit, and of a miss. The computations were done according to MacMillan and Creelman’s (1991) derivation from Bayes’ (1764) formula, as cited by Juslin and Olsson (2005). Because signal and noise trials were equally probable in the present simulations, the equations simplified to,

\[
H_p = 0.5 \times \left( \frac{H}{H + F} \right) \\
M_p = 0.5 \times \left( \frac{M}{M + V} \right)
\]

where, \(H_p\) is the posterior probability of a hit, \(M_p\) is the posterior probability of a miss, \(H\) is the hit rate, \(F\) is the false alarm rate, \(M\) is the miss rate (equal to \(1 - H\)), and \(V\) is the correct rejection rate (equal to \(1 - F\)). Note that the posterior probability of a hit is the probability that a sample is a signal sample given that the decision-maker has judged it to be a signal sample, and the posterior probability of a miss is the probability that a sample is a noise sample given that the decision-maker has judged it to be a noise sample.

Overall, the posterior probability plots (Figure 4) tended to show a large-sample rather than a small-sample advantage. However, for \(M_p\), there was a very small but systematic advantage at high levels of \(|c|\).
Simulations with Dichotomous Variables

In the dichotomous simulations, $x$ and $y$ took on values of either zero or one. In addition, the signal and noise populations were produced not by adding random error to the data but by converting a random proportion of the values to their alternative values—i.e., a proportion of the "ones" were changed to "zeros" and the same proportion of "zeros" were changed to "ones." Insofar as a given sample might contain only a single level of $x$ or a single level of $y$, the correlation between $x$ and $y$ was sometimes mathematically undefined. For consistency with Juslin and Olsson's (2005) method, the simulation algorithm classified all such samples as having come from populations with $p = 0$. This classification scheme was reasonable given that undefined correlations were, in all cases, as likely or more likely to have come from a noise distribution than from a signal distribution (see Table 1). All other aspects of the method were identical to those employed in the continuous simulations.

ROC Curves

The dichotomous-data simulations (Figure 5) showed a large-sample advantage rather than a small-sample advantage, as was the case with the continuous simulations.
Criterion-Specific Accuracy

Figure 6 shows that the results for dichotomous variables were similar in pattern to those obtained for simulations of continuous variables. When measured as the hit rate minus the false alarm rate, decision accuracy was generally higher for large than for small samples, but there was a tendency toward a small-sample advantage when \( \rho \) was moderate and \( c \) was conservative. When decision accuracy was measured as \( H_p \) or as \( M_p \) (see Figure 7), there tended to be a large-sample advantage rather than a small-sample advantage. However, for \( M_p \), there was a very modest small-sample advantage when the criterion levels were high.

Decision Accuracy as a Function of Type I Error Probability

The present criterion-specific simulations assume that when decision-makers set their decision criteria, they do not attempt to keep the hit or false alarm rate constant across different levels of \( n \). However, there are situations in which a decision-maker would want to maintain a constant false alarm rate. Such is the case in formal statistical decision-making where the decision-maker sets \( \alpha \), the Type I error rate, to some conventional value (e.g., \( \alpha = .05 \)). The \( \alpha \) parameter, in the language of null hypothesis testing, is equivalent to the false alarm rate in the language of signal detection theory insofar as both variables represent the actual or estimated proportion of false positives.

Figure 1 illustrates the result of keeping \( c \) constant across \( n \), while allowing the hit rate and the false alarm rate (i.e., \( \alpha \)) to vary. However, given that decision-makers sometimes seek to hold \( \alpha \) constant, it is reasonable to ask whether a small-sample advantage can occur if \( \alpha \) is held constant across \( n \), thus (contrary to Figure 1) allowing \( c \) to vary with \( n \). The data in Figures 2 and 5 provide an answer to this question: If the hit
rate is taken as a measure of decision accuracy, then for any given level of \( \alpha \) (that is, for any given false alarm rate), accuracy is always higher for larger than for smaller samples. The point is further clarified by reanalyzing some of the data. Figure 8 shows the values from the center graph of Figure 3 recalculated to show the hit rate minus false alarm rate as a function of the false alarm rate rather than as a function of \(|e|\), thus allowing the comparison of criterion-specific accuracy (the center graph in Figure 3) to an \( \alpha \)-specific accuracy (Figure 8). The data pattern shows that whereas conservative decision-making (in which \(|e|\) is set very high) can produce a small-sample advantage, no small-sample advantage occurs when the decision-maker holds \( \alpha \) constant across \( n \).

Insert Figure 8 about here

Transformation-Independence of Criterion-Specific Accuracy

The present simulations were theoretically motivated by the skewness of the sampling distribution of the Pearson \( r \) (as illustrated in Figure 1). Thus, it is reasonable to suppose that any small-sample advantage should depend on the Pearson \( r \) being an appropriate measure of statistical association. Yet, it is not known whether real decision-makers represent associations as Pearson \( r \) or as some other measure such as Fisher's \( z \). It turns out, however, that the present simulation results must hold for Fisher's \( z \), for signed \( r^2 \) (in which \( r \) is squared, then multiplied by \(-1\) if \( r \) had initially been negative), or for any monotonic transformation of \( r \). The explanation for this is twofold. First, decision accuracy is always measured as a hit rate, a false alarm rate, or some combination (or complement) of the two. Second, a monotonic transformation of the critical \( r \) and the sample \( r \) can have absolutely no effect on hit rates or false alarm rates: If the untransformed sample \( r \) exceeds the untransformed critical \( r \), then the transformed sample \( z \) must exceed the transformed critical \( z \), or alternatively, the transformed sample signed \( r^2 \) must exceed the transformed critical signed \( r^2 \). A monotonic transformation cannot change the ordinal relationship between the values, which includes the relationship between the critical \( r \) and the sample \( r \), and therefore can have no effect on hit rates or false alarm rates.
Roles of Variance and Central-Tendency Displacement

The small-sample advantage (for extreme $|c|$ combined with moderate $\rho$) occurred even when the distributions were made symmetrical by means of a Fisher's $z$ transformation (illustrated in Figure 9). Therefore, the effect of $n$ on the hit rate minus false alarm rate cannot be explained solely by differential skewness, per se. However, an effect related to differential skewness does play a role. Figure 7 shows that as $n$ decreases from 30 to 5, there is a systematic displacement of the central tendency of the sampling distribution to a point more extreme than that of the population. Specifically, the mean of the Fisher’s $z$ sampling distribution (given $\rho = .50$) shifts by .082, from .557 to .639. However, at the same time, the sampling standard deviation increases by almost a factor of four, going from .192 to .707; and the relation between differential displacement and sampling variability in this range is almost perfect, with the correlation between them being .996. Consequently, the small-sample advantage results not from skewness, per se, but from the combined influences of central-tendency displacement and sampling variability.

Conclusions

Overall, the results indicate that larger samples generally lead to better inferences about correlations than do smaller samples. But the results also demonstrate specific conditions under which a small-sample advantage can occur. Specifically, such an advantage is likely to occur when accuracy is defined in a criterion-specific rather than in a criterion-free manner, when the value of the population correlation is moderate rather than extreme, and when the decision-maker adopts a conservative rather than a liberal decision criterion.

Another limitation involves the nature of the investigator’s standard of rationality. When criterion-specific performance was measured as the posterior probability of a hit (or of a false alarm), there was no small-sample advantage. Yet, for conservative levels of the criterion, there was a very modest small-sample
advantage when decision accuracy was measured as the posterior probability of a miss (or of a correct rejection), and an even greater small-sample advantage when accuracy was measured as the hit rate minus the false alarm rate. The question of which standard of rationality is the appropriate standard, in this domain, cannot be resolved absent specific epistemic criteria and explicit payoff matrices.

The present work shows that the counterintuitive small-sample advantage hypothesized by Kareev and his colleagues (e.g., Kareev, 2000) can indeed occur. However, under most conditions the advantage is for large rather than small samples. In addition, given that the advantage occurs for symmetrically distributed Fisher's z-transformed coefficients, the small-sample advantage is not dependent on the skew, per se, of the sampling distribution. Thus, to the extent that small-sample advantages do occur, they result from a combination of differential central-tendency displacement and differential sampling variability. Future studies will investigate whether actual human decision-making exhibits the kind of small-sample advantage found in the present simulations.

The present findings are consistent with the counterintuitive proposition that resource limitations can be adaptive (J. Anderson, 1990; R. Anderson, 1998; R. Anderson, Tweney, Rivardo, & Duncan, 1997; Gigerenzer, 2000; Kareev, 2000). These findings also show that the conditions under which resource limitations can facilitate inferences about correlation are very limited. Finally, the present work, along with that of Kareev (2000) and Juslin and Olsson (2005), illustrates the need to analyze the ecology as a means of understanding the ways in which adaptive cognitive processes can exploit the structure of the ecology.
References


Appendix

Criterion-Specific Decision Accuracy for an Alternate Coding of Hits and False Alarms

It is possible to adopt an alternate coding of the hit rates (H) and false alarm rates such that, using the symbols described in Figure 1,  \( H = \frac{H_1}{(H_1 + H_2 + L)} \), and \( F = \frac{(F_1 + F_2)}{(F_1 + F_2 + L)} \). Such a coding is problematic with regard to signal detection theory because the coding allows \( F \) to be greater than \( H \), particularly when the signal population has a low correlation (\( \rho \)). Figure A1 illustrates that even when signal \( \rho \) is moderate, \( F \) can still exceed \( H \), yielding a negative value for \( H - F \). Nevertheless, Figure A1 illustrates the small-sample advantage, for moderate \( \rho \) combined with a high criterion, despite the alternate coding.

--------------------
Insert Figure A1 about here
--------------------
Author Note

Richard B. Anderson, Michael E. Doherty, Neil D. Berg, and Jeff C. Friedrich, Department of Psychology, Bowling Green State University.

The authors thank Peter Juslin, Yaakov Kareev, and two anonymous reviewers for careful and helpful comments on an earlier draft of this manuscript. We also thank Ryan Tweney for introducing us to literature on the small-sample advantage, James Sullivan for statistical consulting, and Amanda Kelley for assistance in running the simulations.

Correspondence concerning this article should be addressed to Richard B. Anderson, Department of Psychology, Bowling Green State University, Bowling Green, Ohio, 43403. Email: randers@bgnet.bgsu.edu
Footnotes

1 Some readers may be more familiar with an alternative measure—the mean of the hit rate and the correct rejection rate—which is equal to the percent correct when the number of signal trials equals the number of noise trials. It should be noted that this measure must necessarily yield a data pattern equivalent to that yielded by the hit-minus-false-alarm measure, since the correct rejection rate is the mathematical complement of the false alarm rate.

2 For consistency with Juslin and Olsson (2005), the posterior probabilities were multiplied by 0.5, as shown in Equations 1 and 2.
Figure Captions

Figure 1. Empirical sampling distributions of the Pearson correlation coefficient, where the population correlations are .5 (top graph) and zero (bottom graph).

Figure 2. Hit rate (H) as a function of false alarm rate (F), for simulations of continuous variables. H = (H₁ + H₂) / (H₁ + H₂ + L), and F = (F₁ + F₂) / (F₁ + F₂ + Q), as described in Figure 1. Each graph shows the results for a particular combination of signal ρ and noise ρ. Noise ρ is zero in all cases.

Figure 3. Hit rate (H) minus false alarm rate (F) as a function of the decision criteria and of n, for simulations of continuous variables.

Figure 4. The posterior probability of a hit (Hp) and of a miss (Mp) as a function of the decision criteria and of n, for simulations of continuous variables (signal ρ = .5). Probabilities are computed according to Equations 1 and 2. When n = 30 and |c| = .9, Hp is undefined (because of division by zero) and is therefore not plotted. The posterior probabilities of a false alarm and of a correct rejection are not presented because they are the inverse of Hp and Mp, respectively.

Figure 5. Hit rate (H) as a function of false alarm rate (F), for simulations of dichotomous variables.

Figure 6. Hit rate (H) minus false alarm rate (F) as a function of the decision criteria and of n, for simulations of dichotomous variables.

Figure 7. The posterior probability of a hit (Hp) and of a miss (Mp) as a function of the decision criteria and of n, for simulations of dichotomous variables where signal ρ = .5.
Figure 8. Hit rate (H) minus false alarm rate (F) as a function of n and of the false alarm rate, which is analogous to the Type I error rate, $\alpha$. The data are for simulations of continuous variables where signal $\rho = .5$.

Figure 9. Empirical sampling distributions of Fisher’s $z$. The dashed vertical lines show the placement of the decision criteria where $|c| = 1.47$ (which corresponds to a Pearson $r$ of .9). The symbol definitions are as in Figure 1. For the signal distributions, the mean (marked by vertical, solid lines for $n = 5$ and $n = 30$) shifts as a function of $n$.

Figure A1. Criterion-specific decision accuracy using an alternate computation of the hit rate minus false alarm rate. * $H_1$ samples are counted as hits, but $H_2$ samples are not. Both $F_1$ and $F_2$ samples are counted as false alarms. The results are for correlations between continuous variables.
Table 1

Proportion of samples for which there was no defined correlation, as a function of the population correlation and \( n \).

<table>
<thead>
<tr>
<th>Population Correlation (( \rho ))</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n ) per sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.10</td>
<td>0.096</td>
<td>0.096</td>
<td>0.086</td>
<td>0.075</td>
</tr>
<tr>
<td>7</td>
<td>0.029</td>
<td>0.029</td>
<td>0.020</td>
<td>0.027</td>
<td>0.029</td>
<td>0.028</td>
<td>0.027</td>
<td>0.026</td>
<td>0.026</td>
<td>0.020</td>
</tr>
<tr>
<td>10</td>
<td>0.0037</td>
<td>0.0042</td>
<td>0.0043</td>
<td>0.0036</td>
<td>0.0035</td>
<td>0.0043</td>
<td>0.0032</td>
<td>0.0031</td>
<td>0.0031</td>
<td>0.0027</td>
</tr>
<tr>
<td>15</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0</td>
</tr>
<tr>
<td>30</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Figure 2

Hit Rate vs. False Alarm Rate for different signal strengths and number of samples per signal.

- Signal $\rho = 0.1$
- Signal $\rho = 0.2$
- Signal $\rho = 0.3$
- Signal $\rho = 0.4$
- Signal $\rho = 0.5$
- Signal $\rho = 0.6$
- Signal $\rho = 0.7$
- Signal $\rho = 0.8$
- Signal $\rho = 0.9$

Number of samples per signal: 5, 7, 10, 15, 30.
Criterion (+ and −)

Hit Rate Minus False Alarm Rate

Figure 3
Figure 4
Figure 5
Hit Rate Minus False Alarm Rate

Criterion (+ and −)
Figure 7
False Alarm Rate (Critical $\alpha$)

$H - F$

$n$ per sample

Figure 8
Figure 9

Number of Samples vs. Sample Fisher's $z$ for different $n$ per sample:

- $H_2$, $L$, $H_1$
- $F_2$, $Q$, $F_1$

- $n$ per sample:
  - 5
  - 7
  - 10
  - 15
  - 30

$SIGNAL$ and $NOISE$ distributions are shown for each $n$.
signal $\rho = .5$

![Graph showing $H - F^*$ vs. Criterion (+ and -) for different $n$ per sample (5, 7, 10, 15, 30).](image)

Figure A1