Predictive Judgments in Situations of Statistical Analysis

MARIE-PAULE LECOUTRE AND HENRY ROUANET

Groupe Mathématiques et Psychologie, Université René Descartes

Probabilistic judgments made by researchers in psychology were investigated in statistical prediction situations. From these situations, it is possible to test the “representativeness hypothesis” (Tversky & Kahneman, 1971) and the “significance hypothesis” (Oakes, 1986). The predictive judgments concerned both an elementary descriptive statistic and a significance test statistic. In the first case, the predictive judgments were generally coherent and fit comparatively well to Bayesian standard predictive probabilities. In the second case, they were generally incoherent and fit poorly to Bayesian standard predictive probabilities. As for the two hypotheses tested, our findings are compatible with the significance hypothesis, but go against the representativeness hypothesis. © 1993 Academic Press, Inc.

INTRODUCTION

Through the statistical analysis of experimental data, a given conclusion is drawn about the effect being investigated. What is the probability that this conclusion would be reached again were the experiment replicated under the same conditions? This is a problem of statistical prediction which every experimentalist has surely met some time in the course of his or her research.

The idea that the replication of findings from one experiment to another is the essence of scientific enterprise is commonplace among scientists: “The essence of science is replication: a scientist should always be concerned about what will happen when he or another scientist repeats his experiment” (Guttman, 1977). In a previous work (Lecoutre, 1983), we reported that researchers often bring up this predictive idea. Moreover, statistical techniques are available to bring this idea into practice, namely, the procedures of Bayesian statistical prediction. The development of these techniques goes back to Laplace (see Laplace’s celebrated “seventh principle,” 1825) and has been pursued by contemporary Bayesian statisticians, beginning with Jeffreys (1961), Lindley (1965), and many others.

In statistical prediction, probability calculations pertain to observations

We thank Bruno Lecoutre for assistance in calculating Bayesian predictive probabilities, and for helpful comments. Address correspondence and reprint requests to Marie-Paule Lécoutre, Groupe Mathématiques et Psychologie, CNRS, URA 1201, Université René Descartes, Sorbonne, 12 rue Cujas, 75005 Paris, France.
to come, which are dependent upon available data. We call this probability a natural probability because it goes from the known (the data) to the unknown (observations to come), as is natural for probabilities. It should be pointed out that in the most commonly used statistical technique, namely, significance testing, the conventional argument is not natural since it is based on a probability that pertains to data and is dependent upon a hypothetical proposition (the null hypothesis). Such a probability, which goes from the unknown (the null hypothesis) to the known (the data) is surely not a natural one.

The difficulty of interpreting significance tests is notorious and is known to persist even for experienced researchers. This fact was confirmed by our findings (Lecoutre, 1983). It seems legitimate to trace these difficulties to the artificial nature of the underlying probabilistic reasoning.

In the work presented here, the intuitive probabilistic judgments of researchers in situations of statistical prediction are investigated. Previous experimental work related to this subject may be found in the literature on probabilistic judgments (Nisbett & Ross, 1981; Kahneman, Slovic, & Tversky, 1982; Oakes, 1986). Especially relevant are Tversky and Kahneman (1971) and Oakes (1986). Tversky and Kahneman asked researchers to estimate the probability, given a significant result, that that result will be significant a second time for an additional group of subjects. The probability values given as responses were found to be markedly higher than reference Bayesian probability values. Similar findings are reported by Oakes. Two hypotheses have been proposed to interpret these findings.

(1) Tversky and Kahneman invoke the "representativeness hypothesis," according to which the overestimation of the replicability of an experimental result is due to the underestimation of sample fluctuations, and to an unjustifiably high degree of confidence that any two samples resemble each other.

(2) Oakes invokes the "significance hypothesis," according to which the outcome of a significance test is interpreted in terms of a dichotomy: an effect either "exists" (when it is significant), or "does not exist" (when it is nonsignificant), leading in the former case to an overestimation of its replicability, and in the latter case to an underestimation of the chance of finding an effect. To test these two rival explanations, Oakes asked 54 psychologists three questions, one of which led to differential predictions. The findings were markedly in favor of the significance hypothesis.

The work presented hereafter focuses on testing the representativeness hypothesis. The following three types of results were defined in the first experiment.
• Situation 1: strong effect (significant).
• Situation 2: small effect (nonsignificant).
• Situation 3: no effect (nonsignificant).

For each situation, subjects (researchers) were asked two predictive questions about two statistics to be calculated in the replicated experiment.

(1) One question (question B) involved Student’s $t$ test statistic. More specifically, the question concerned the probability of finding, in the replication, a result that was either at least as significant (in situation 1) or at least as nonsignificant (in situations 2 and 3) as in the first experiment. For this question, the representativeness and the significance hypotheses both led to the same prediction, namely an overestimation of replicability.

(2) The other question (question A) was about a descriptive statistic, namely the observed difference in means.

At this point, it would be in order to introduce our conceptual distinction between natural statistics and sophisticated statistics. By natural statistics, we mean procedures for which there is an immediate link between data and outcome, and for which a “natural interpretation” (in the Feyerabend sense, 1975) is available. The mean is the prototype of natural statistics. We have often heard researchers make a comment like this when faced with a table of means: “Here, there is no need for statistics to draw conclusions!,” as if a set of means would not really belong to the “realm of statistics.” Unlike natural statistics, “sophisticated statistics” are based on a complex nonintuitive argument, where the link between data and outcome is easily lost. The prototype of sophisticated statistics is the significance test.

To summarize in regard to our experiment, one of our questions pertained to a sophisticated statistic, the other, to a natural statistic. For the natural statistic question, the two hypotheses lead to different predictions.

According to the representativeness hypothesis, an overestimation of the replicability of a result should be found for the natural statistic as well as for the significance test statistic. The significance hypothesis, however, is based on the test statistic and there is no reason to expect overestimation in predictive judgments. Our own hypothesis in this case is that for the natural statistic, the three predictive judgments (which are based on available natural interpretations) will be more coherent than for the test statistic and will be better fitted to reference Bayesian probability values.

The predictions regarding our data can be summarized as follows.

(1) For the significance test statistic, overestimation in predictive judgments is expected in all three situations.
(2) For the natural statistic, some response coherence across situations is expected, along with an acceptable fit to the Bayesian probability values.

(3) For the test statistic, a good number of subjects lacking a specific intuitive interpretation will attempt to use available ones, especially those related to natural statistics. This is why one can expect a nonnegligible proportion of identical responses for the test and natural statistics.

METHOD

Experimental Materials and Design

The three situations are presented in the same general context, as follows:

"May I ask you to put yourself in the following situation? In a study on problem solving you have performed an experiment whose aim was to compare solving times for two problems presented to the same subjects. You have thus presented the first problem, P1, to 20 adult subjects, and you have recorded the time (in seconds and tenths of seconds) taken by each subject to solve the problem. You have obtained the following results.

(The experimenter shows the subject a card indicating the results obtained for P1).

Then, to the same 20 subjects, you have presented the second problem, P2, for which you have again recorded each one's solving time. In what follows, three possible results for P2 will be considered, while the P1 results remain the same throughout.

The three situations differed by the results of problem P2, as indicated in Table 1. For all three situations, the observed mean difference, $d$, between the solving times of the two problems always had the same sign. For situation 1, $d$ was large and the outcome of the significance test was significant. For situation 2, $d$ was small, and the outcome of the significance test was moderately nonsignificant. For situation 3, $d$ was smaller still, and the outcome of the significance test was largely nonsignificant. The standard deviation of the difference was the same in all three situations.

<table>
<thead>
<tr>
<th>Situation</th>
<th>$d$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1.820</td>
<td>+2.093</td>
<td>.05</td>
</tr>
<tr>
<td>2</td>
<td>+0.920</td>
<td>+1.058</td>
<td>.30</td>
</tr>
<tr>
<td>3</td>
<td>+0.220</td>
<td>+0.253</td>
<td>.80</td>
</tr>
</tbody>
</table>

* Observed mean difference $d$, test statistic $t$, and description level $p$ (two-tailed).
The three situations were presented one after the other. After each presentation, the subject was told "Now suppose you plan to replicate this experiment, taking the same number of observations as in the first experiment. I am going to ask you some questions about this replication."

Two questions, A and B, were then asked in one of the three situations.

Question A was about the sign of the difference, $d'$, found in the replication, and was formulated as follows: "What, for you, is the probability that the observed difference, $d'$, will have the same sign in the second experiment as $d$, that is, that it will be positive ($d' > 0$)?"

Question B was about the significance test statistic, $t'$, in the replication. The formulation varied according to the outcome of the first experiment. For situation 1, the question was "What, for you, is the probability that in the second experiment, the observed difference, $d'$, will have the same sign as $d$, and that the result of Student's $t$ test will be at least as significant as in the first experiment, that is, $t' > +2.093$ or $d' > 0$ and $p' < .05$?"

For situation 2, it was formulated as follows: "What, for you, is the probability that in the second experiment, the observed difference, $d'$, will have the same sign as $d$, and that the result of Student's $t$ test will be at least as nonsignificant as in the first experiment, that is, $0 < t' < +1.058$ or $d' > 0$ and $p' > .30$?"

For situation 3, the same question was asked, replacing $0 < t' < +1.058$ by $0 < t' < +0.253$, and $p' > .30$ by $p' > .80$.

Subjects and Procedure

Fifty researchers from various laboratories in Paris, all with practical experience at processing experimental data, participated in the experiment. They were requested to respond in a spontaneous fashion, without making calculations, and it was stressed that the task was in no way a test of their knowledge of statistics. The responses were gathered individually by means of semidirective interviews recorded on tape. Interview times ranged from half an hour to over an hour, depending on the subject.

For 25 of the researchers, the three situations were presented in this order: situation 1 (first result significant), then situations 2 and 3 (first result nonsignificant). In all situations, question A was asked first. For the other 25 researchers, situations 2 and 3 were first presented and then situation 1. In this case, question B was asked first.

RESULTS

There was no appreciable effect of situation presentation order (the difference between the two groups of researchers was negligible in all situations: $t(46) = 0.63$ for situation 1, NS; $t(46) = -0.28$ for situation 2,
NS; and \( t(46) = -0.38 \) for situation 3, NS). Question presentation order had no effect either \( t(46) = 0.25, \) NS, for question A, and \( t(46) = -0.14, \) NS for question B).

These findings led us to pool the data, so that the results and comments given below concern all 50 researchers.

**Predictive Judgments Made by Researchers**

The results for the sign of the observed difference, \( d' \) (question A), are reported in Fig. 1.

In all three situations, the response spread is great. For situation 1, more than 75% of the response values (0.78) are above 0.50. For situations 2 and 3, they are on the whole lower and concentrated around the response stereotype, 0.50. For situation 3, there is an intriguingly high percentage (38%) of responses below 0.50, with comments like: "Since \( d \) is near zero, the probability of again finding something positive is very low." This of course is a fallacy. Since in the first experiment \( d \) is positive, though quite small, the probability that in the experiment to come the observed difference will be positive cannot decrease. The fallacy can be understood by the following argument, which is in line with Cohen's

\begin{figure}
\centering
\includegraphics[width=\textwidth]{situation1.png}
\caption{Situation 1}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{situation2.png}
\caption{Situation 2}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{situation3.png}
\caption{Situation 3}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{predictions.png}
\caption{Predictions about the sign of the observed difference (question A): distribution of the 50 responses by situation.}
\end{figure}
(1982) suggestion for rationalizing the well-known "gambler's fallacy." In the first experiment, the weakness of the observed difference leads one to "take a position" in favor of the inexistence of a systematic difference. Predicting a further positive difference would thus go against this taken position.

The analysis of the response patterns of the three responses given in succession to question A revealed that these patterns are generally (56%) coherent in the sense that the response values decrease from situation 1 to situation 3. Eighteen percent of the subjects gave the same response for all three situations, and only one subject gave response values that increase from situation 1 to situation 3.

The findings for the test statistic, $t'$ (question B), are reported in Fig. 2.\textsuperscript{1}

A substantial amount of response variability was found for the three situations. In situation 1 (first result significant), the probability values are generally high: most response values (80%) are higher or equal to 0.50. Taking the two situations in which the first result is nonsignificant, we can see that in situation 3, the probability values that are higher or equal to 0.70 are much more frequent (40%) than in situation 2 (24%). A typical

\textsuperscript{1} On question B, there were only 48 responses for situation 2, and 49 for situations 1 and 3. One researcher, "not having the slightest idea about the test," systematically gave no answer, and another researcher gave no answer for situation 2.
comment was: "The significance level is more clear-cut than just before; therefore the probability is necessarily even higher," etc. This may be merely a "surface" phenomenon linked to a literal interpretation of the significance level, a value near 1 meaning a clear-cut, reliable result that is perfectly likely to be found again. Or, it may be a "deeper" phenomenon resulting from a mistaken interpretation of the significance test in terms of decision making (Oakes, 1986), or more generally, from a cognitive tendency to "take a position," and more specifically here at the end of a study, both when the outcome of the test is nonsignificant and when it is significant. Once a result is classified as a "nondifference," an elevated probability of finding a similar result by replicating would be attributed.

When the response patterns of the three responses given successively to question B are analyzed, we can see that on the whole (62%) they are noncoherent in the sense that there is no systematic relationship between the probability values given as responses and the results of the significance test in the first experiment. Only 20% of the response patterns are coherent since the response values decrease from the first situation to the last.

Relationship between the Two Types of Predictive Judgments

A comparison of responses to questions A and B is presented in Table 2 for each situation.

A striking finding is the number of researchers who gave exactly the same answer to questions A and B: 33% for situation 1, 21% for situation 2, and 13% for situation 3. Typical comments by these researchers were: "In fact, it's quite the same thing," "I think that the first one entails the other one," or "The value does indeed depend on the standard deviation, but weakly; it's the mean that mainly matters, so we will get the same thing," etc. When two different response values were given, they are often close to each other. The differences between the responses given to questions A and B are smaller than or equal to 0.10 in 50% of the cases in situation 1, 52% in situation 2 (here the differences are nonsignificant \( t(46) = 1.47 \)), and 28% in situation 3.

<table>
<thead>
<tr>
<th>Situation</th>
<th>( P_d &gt; P_t )</th>
<th>( P_d = P_t )</th>
<th>( P_d &lt; P_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation 1</td>
<td>0.54</td>
<td>0.33</td>
<td>0.13</td>
</tr>
<tr>
<td>Situation 2</td>
<td>0.56</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td>Situation 3</td>
<td>0.33</td>
<td>0.13</td>
<td>0.54</td>
</tr>
</tbody>
</table>

\( a \) \( P_d \) prediction for the difference (question A), \( P_t \) prediction for the test (question B).
Standard Predictive Probabilities

In situations like the present one, where there is no available information other than the data, the Bayesian methods based on diffuse initial distributions yield probabilities that can be taken as standard predictive probabilities (SPP) (see Appendix).

The standard predictive probability values calculated for each of the three situations are reported in Table 3 for questions A and B.

Concerning the standard predictive probabilities, for question A all SPP are above .50, and decrease from situation 1 to situation 3. For question B, they decrease from a value near .50 (situation 1) to a very small value (situation 3).

Judgments Made and the SPP: Test of the Two Hypotheses

Comparison of the responses given by researchers (Figs. 1 and 2 and Table 3) with the SPP leads to the following comments.

Question A. For the first two situations, most response values are lower than the SPP. The means, which are equal to 0.70 and 0.56, respectively, are significantly lower than the SPP (p < .001 in both situations). This can be interpreted as reflecting a cautious attitude, that is, a reluctance to hastily generalize properties found on a sample.

Notice that such a cautious attitude prevails in situation 1 despite the significant outcome of the test, and in situation 2—perhaps—because of the nonsignificant outcome of the test. For situation 3, most response values are close to the SPP value (.57). But in order to interpret this finding, the proximity of the response stereotype, ½, must be kept in mind. Some underestimation again exists: the mean, which is equal to 0.44 here, is also significantly lower than the SPP (p < .001). Thus, systematic underestimation with respect to the SPP was observed. These findings go against the representativeness hypothesis which, on the contrary, would lead to the overestimation of probability values.

Question B. For situation 1, the bulk of the responses are either close to the SPP (although this coincides with the response stereotype, ½) or markedly higher. The mean, which is equal to 0.57, is significantly higher than the SPP (p < .05). For the other two situations, most responses are

<table>
<thead>
<tr>
<th>Situation 1</th>
<th>Question A</th>
<th>Question B</th>
</tr>
</thead>
<tbody>
<tr>
<td>.92 (0.70)</td>
<td>.50 (0.57)</td>
<td></td>
</tr>
<tr>
<td>.77 (0.56)</td>
<td>.27 (0.51)</td>
<td></td>
</tr>
<tr>
<td>.57 (0.44)</td>
<td>.07 (0.56)</td>
<td></td>
</tr>
</tbody>
</table>
higher than the SPP. The means, which are equal to 0.51 for situation 2 and 0.56 for situation 3, are significantly higher than the SPP ($\rho < .001$) in each situation. Overestimation is considerable for situation 3. Thus, there is overestimation in all three situations, especially in situations 2 and 3 (first result nonsignificant). These findings are consistent with the representativeness hypothesis, but also with the significance hypothesis, since in such a case, both lead to the same prediction.

We might add that since the difference was underestimated and the test statistic was overestimated, it would not be possible to devise a Bayesian distribution better than the standard probability distribution for both kinds of statistics considered simultaneously.

Overall examination of the results suggests that when a natural (i.e., predictive) probability is requested about a natural (i.e., descriptive) statistic, the fit of the Bayesian model is acceptable, at least as a first approximation. But when "nonnaturalness" is introduced into the situation, either for the underlying probabilistic argument or for the type of statistic, intuitive judgments are largely biased.

DISCUSSION

The main finding here is that researchers do not exhibit the same tendency in their predictive judgments about the difference of means and the significance test statistic.

(1) Predictive judgments about the difference of means are most often coherent and fit comparatively well to standard predictive probabilities. Moreover, systematic underestimation with respect to SPP was observed, which presumably reflects a cautious attitude. These findings go against the representativeness hypothesis.

(2) Predictive judgements about the test statistic are most often incoherent, and SPP fit is poor. The systematic overestimation (whether the first result is significant or nonsignificant) is compatible with both the significance hypothesis and the representativeness hypothesis.

(3) A number of researchers hardly differentiate (and some of them not at all) their predictions about the observed difference from those about the significance test. This result strongly suggests, as expected, that a substantial proportion of researchers lack a specific intuitive representation for the significance test. This might suffice to interpret the observed overestimation, alleviating the need to invoke the representativeness or significance hypothesis.

(4) Finally, for the significance test, our findings suggest that subjects either have no specific interpretation and use an available one that is appropriate to another situation (hence the same prediction for significance test and difference of means) or develop biased interpretations,
such as those previously mentioned in connection with the significance hypothesis.

In conclusion, such a bias does not appear to be linked to the judgmental process itself, but to the object of the judgment, that is, the significance test. To reduce this bias, users should be provided with alternative interpretations of significance tests that are more natural than the conventional one. The Bayesian interpretation is one such alternative, of course. Another one might well be the set-theoretic interpretation, recently explored by Rouanet, Bernard, and Lecoutre (1986). Related situations might be investigated from an experimental viewpoint. For instance, it might be worth comparing the situation studied in this paper with one in which the null hypothesis deserves special prior credence (say, a well-established theory) in order to find out whether some specific intuition for significance testing would show up.

APPENDIX

Standard predictive probabilities were derived within the Bayesian framework, taking as a prior distribution the classical locally uniform distribution for parameter \( \delta \) (see e.g., Lindley, 1965, or Box & Tiao, 1973).

Since here we have \( n' = n \), the sampling distributions of the \( d \) and \( d' \) statistics have the same variance \( e^2 \). If \( e^2 \) were known, the predictive distribution pertaining to \( d' \) would be normal, centered around \( d \) (the observed mean in the first experiment), with a variance equal to \( 2e^2 \), i.e., \( d' \sim N(d, 2e^2) \).

Since \( e^2 \) is not known, it is estimated by \( e^2 \) (the denominator of the \( t \) statistic), and the normal distribution is replaced by a generalized Student distribution centered around \( d \) with a scale parameter of \( 2e^2 \), hence the standard predictive distribution \( d' \sim t_{19}(d, 2e^2) \).

For question A, this distribution provides the requested probabilities \( P(d' > 0) \) for the three situations. The values of these probabilities all lie between \( \frac{1}{2} \) and \( 1 - p/2 \), where \( p \) denotes the observed two-tailed level in the first experiment (\( 1 - p/2 = 0.975, 0.85, \) and 0.60, respectively).

As for question B, if \( e^2 \) were known, the requested probabilities could be simply derived from the one pertaining to \( d' \), namely \( P(d' > d) = .50 \) for situation 1, and \( P(0 < d' < d) = P(d' > 0) - \frac{1}{2} \) for situations 2 and 3. Since \( e^2 \) is not known, this procedure is only approximate. For the exact solution, see, for example, Lecoutre (1984); but the approximate procedure here yields two correct decimal places.

REFERENCES


**RECEIVED:** December 21, 1989