PROBLEMS WHEN INTERPRETING RESEARCH RESULTS USING ONLY *P*-VALUE AND SAMPLE SIZE

Rink Hoekstra, Henk A.L. Kiers, Addie Johnson and Marleen Groenier
University of Groningen, The Netherlands
r.hoekstra@rug.nl

The following three probabilities seem crucial when interpreting data, especially in the behavioral sciences:1) the probability that an effect is present in the population, 2) the probability that a replication is significant; and 3) the probability that the effect for a single individual in the population is in the expected direction. In our study, we asked 51 subjects (university students and lecturers in psychology) to estimate these probabilities after reading a short description of a hypothetical experiment with as outcomes only p-value and sample size given. Large variations in estimated probabilities were found. However estimates of the probabilities tended to increase as a positive function of sample size for a fixed p-value. Simulation studies show that, assuming a uniform prior distribution for the parameter, this turns out to be incorrect for all three probabilities.

INTRODUCTION

One of the main goals of statistics is to draw inferences from the data provided by a sample to the population one is interested in. One of the most frequently used methods for this purpose is null hypothesis significance testing (NHST). Although this approach has been heavily criticized for decades, for many researchers it prevails as the most common way to analyse data (e.g., Cohen, 1994; Masson and Loftus, 2003; Tryon, 2001).

NHST is designed to assess the strength of the evidence against a null hypothesis (H_0). Such a null hypothesis usually states that there is no effect in the population. In NHST the probability is calculated of finding a test statistic with a value as extreme as or more extreme than the actually observed value, assuming that there is no effect in the population. The smaller this *p*-value, the stronger the evidence against H_0 (Moore and McCabe, 2003).

From previous research it is known that errors are often made in interpreting the outcomes of a significance test. For example, Oakes (1986) investigated the knowledge and misconceptions about significance testing of 70 psychology researchers. A large majority (67) of those psychologists believed incorrect interpretations of the significance test to be correct. Lecoutre *et al*. (2003) found that even statisticians had difficulties interpreting the significance test. It appeared especially difficult to interpret the results of a nonsignificant effect.

When researchers analyse their data, they have some questions in mind that they want to get answered. Some of those questions cannot be answered with the available techniques, but nevertheless they are important to the researcher. When an effect is found in a sample, three important questions seem the following: 1) How certain is it that the effect also holds in the population? 2) What percentage of people in the population will show the effect found in the sample? 3) How certain is it that a similar study will show a similar effect? The first question we call the "certainty question". When the difference between two means is at interest, we can define the degree of certainty as $P(\mu>0 \mid n, p)$, with μ the unknown population mean, n the given sample size and p the given p-value. This seems to be an unallowable statement, because μ should be a fixed value (although we do not know the value of μ). We will come back to this problem later. Without further information, the certainty question cannot be answered. Although it is not possible to answer the certainty question using only the outcomes of NHST, there are indications that people interpret the p-value as a measure of certainty of the existence of the effect (e.g., Oakes, 1986).

The second question, regarding the percentage of people in the population who will show the effect, we call the "extrapolation question." We define the corresponding probability as $P(x_i>0 \mid n, p)$, with x_i being the score of a random person from the population. This question can, for example, be relevant for researchers who are interpreting a given effect when testing a new medicine. They might by interested in whether the medicine works only for a smaller subgroup,

or whether the medicine works marginally for the entire population. When one is only interested if there is an effect, this question seems less relevant.

The third question, the "replicability question," pertains to the extent that the effect is due to idiosyncrasies of the specific study, or whether a comparable study would produce comparable results. We define the associated probability as $P(p(M_{rep}) < .05 | p, n)$, with $p(M_{rep})$ being the *p*-value of an exact replication of the first study, using the same sample size.

If there is no further information available, except for the *p*-value and the sample size, it is not possible to answer any of the three questions. Therefore there is no absolute "correct" or "incorrect" answer. However, when we vary either the *p*-value or the *n*, keeping the other fixed, it is possible to give the correct direction of every possible comparison of answers. When *n* is fixed and the *p*-value is decreased, the correct direction of answers to the three questions seems straightforward: when *p* is smaller one should be more certain, the proportion of the population with the effect should be higher, and the replicability should be higher. When we reverse the roles, (a fixed *p*-value and an increasing *n*) the situation is no longer so intuitively clear, because this relation between *n* and the *p*-value differs for the three questions.

To study the relation between *p*-values and sample size, we conducted three simulation studies using Bayesian statistics. When the outcomes of a sample study are known, one can still imagine the population mean having every possible value. Given an infinite number of possible μ's, we want to know the proportion of those μ's which are larger than 0. Thus, it is necessary to treat μ as a stochastic variable. If we assume an a priori distribution of μ, Bayesian statistics can be used to compare probabilities when changing *n* and keeping the *p*-value fixed. An a priori distribution in which every value has the same probability is called a uniform prior. Uniform priors can be used when no meaningful prediction about the a priori distribution can be made. Since that was the case here, we assumed a uniform prior distribution for μ.

For the certainty question, the simulation study assuming a uniform prior for μ showed, perhaps contrary to popular belief, that the associated probability only depends on the *p*-value and not on *n*. It can be proven that, given this distribution of μ, the certainty probability is equal to (1-*p*). For the extrapolation question, the simulation study showed that the extrapolation probability is dependent on the *p*-value as well as on *n*: With increasing *n* and fixed *p*-value the extrapolation probability decreases. This finding can be explained as follows: The larger the sample size, the smaller an effect needs to be to result in the same *p*-value. Thus, effects with the same *p*-value are smaller with larger *n*, and on average larger effects come from populations with more people scoring above the mean assumed under the H_0.For the replication question, the simulation study showed that, just like the certainty probability, the replication probability only depends on the *p*-value and not on *n*.

In summary, we showed that it is possible to deduct the direction of the probability for each question when *n* increases while keeping the *p*-value fixed, assuming a uniform prior for the population mean. In the behavioral study, we try to answer two questions. The first question is: Are researchers consistent when estimating the probabilities for the certainty, extrapolation and replication questions? The second research question is: Do researchers draw correct conclusions from the comparison of statements with the same *p*-value, but different *n*?

METHOD

Fifty-one University of Groningen students and staff, aged 19-60 (mean = 31.02, standard deviation = 11.50) took part in the research. The subjects were asked to make probability judgements based on briefly presented results of hypothetical research. The scenarios and questions were presented on the computer and responses were made via the computer keyboard. Filling in the questions took 20-30 minutes. Every trial started with a short scenario of a hypothetical experiment testing the efficacy of a blood pressure drug. The subject was required to read the scenario and make a judgment about the results. Sample size and *p*-value varied in the scenario, as was the form of the question asked. Means and measures of variance were not given because they are difficult to interpret on an unknown scale. Sample size was either 10, 50 or 100, the *p*-value was either .01, .03, .05, .08 or .10, and the question form was either the certainty, the

extrapolation, or the replication question. These three factors (sample size, *p*-value and question form) were factorially combined, resulting in a total of 45 trials per subject.

Following the presentation of each question, subjects were required to type in an estimate between 0 and 100 percent. As soon as the percentage was confirmed by pressing 'Enter', the next question appeared, without a possibility to return to earlier questions. This was done to lower the risk that the subject would be influenced by previous answers. The order of the trials was randomized for each subject. It was stressed that the exact probabilities could not be calculated and they should fill in what they thought was the most reasonable answer. It should be noted that almost all subjects complained afterwards about the difficulty of the task.

RESULTS

Figure 1 shows the effects of *p*-value and *n* on the certainty question. For a given *p*-value the average estimates increase with *n* for the question of how certain one was of finding the effect in the population. Note that the standard deviations are relatively high, varying from 24.9 to 34.2. The figures for the other two question types show similar findings.

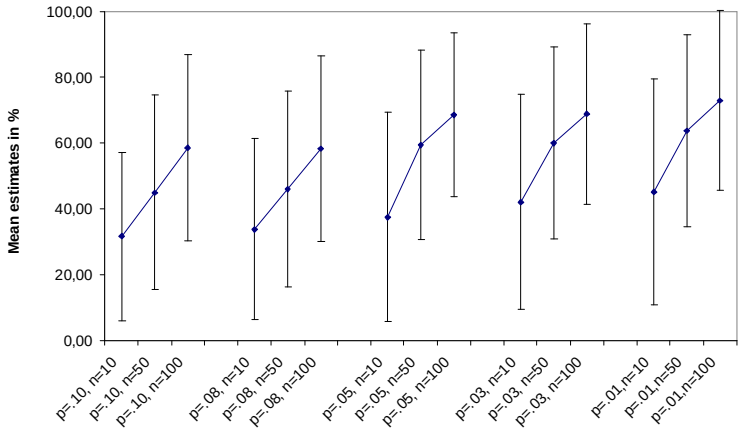

Figure 1: Probability estimates as a function of *n* and p-value for the certainty question The bars indicate standard deviations.

Two main conclusions can be drawn from the data as far as the certainty question is concerned: First, even if *n*, *p*-value and question type are taken into account, the variations between probability estimates are large when related to the range of possible probability estimates (0% until 100%). This indicates large variations between subjects. The second conclusion is that, despite these large variations, clear estimation differences can be found when *n* is varied and the *p*-value is fixed. In the next paragraph, we will focus on that relation.

The observed effects due to *p* and *n* indicate that the probability estimates were not completely random, which would be the case if subjects were not able to perform the estimation task. Increasing probability trends were found with fixed *p* and increasing *n*. This does, of course, not necessarily mean that this holds for every subject. We can get more insight into this matter when we look at all triads (n=10, n=50 and n=100) for every *p*-value and question type for every person. These triads can be increasing (just as the general trend showed), decreasing, equal, or inconsistent (meaning that the probability estimate for n=50 was not between the estimates for n=10 and n=100). Table 1 shows that a large proportion of those triads showed an increasing trend, whereas there are only small minorities for the decreasing trend, equal estimates and inconsistent estimates. Only 12% of all estimates were in agreement with the results from our simulation study, based on a uniform prior distribution of the population mean.

Table 1: The direction of the triads (n=10, 50 and 100), when p-value and question type are fixed. Bold percentages represent estimates in agreement with the simulation study outcomes.

Probability: \ *Trend:*	Monotonic increasing	Monotonic decreasing	Equal estimates	Inconsistent estimates
Certainty	64%	4%	*15%*	18%
Replication	50%	9%	***15%***	26%
Extrapolation	48%	***6%***	18%	27%

DISCUSSION

We asked psychologist to estimate probabilities on three questions highly relevant for interpreting research, while *n* and the *p*-value were varied. The data were characterized by large differences between subjects, making it unlikely that people have comparable interpretations when presented with the same data as far as the three questions (certainty of an effect in the population, extrapolation to individual cases and replication) are concerned.

These findings seem somewhat disturbing for scientific practice. We would argue that a researchers' interpretation of a study should at least partly depend on the interpretation of the presented statistical results. As our results show, these interpretations differ widely. It could be that researchers consider more variables than only the *n* and *p*-value, which were the only variables given in our experiment. We cannot exclude this possibility, but we regard it unlikely that adding other statistics, for example means, standard deviations, *t*- or *F*-values (frequently given statistical information in articles), would lead to radically different results.

It was found that, in general, the three probability estimates were higher with larger *n* and fixed *p*. A possible explanation is that subjects use *n* as a measure of reliability of the study, and use this information to interpret the *p*-value. This, however, ignores the fact that the *p*-value already includes the value of *n*.

REFERENCES

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

Lecoutre, M. P., Poitevineau, J., and Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis tests. *International Journal of Psychology*, 38, 37-45.

Masson, M. E. J. and Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57, 203.

Moore, D. S., and McCabe, G. P. (2003). *Introduction to the practice of statistics.* New York: W. H. Freeman and Company.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chicester: John Wiley and Sons.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6, 371-386.