The use and usability of inferential techniques
Hoekstra, Rink

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
7. Conclusions

The practice of inference in general and the use of NHST in particular has been criticized for decades. The criticisms focussed on the fact that NHST seems to encourage people to interpret the outcomes of their analyses in a dichotomous way (Rosnow & Rosenthal, 1989), on the fact that it does not tell what the researcher wants to know (Cohen 1994), and on the fact that a null hypothesis is always false (Meehl, 1967; Cohen, 1994). Despite the fact that many articles have pleaded for changes in the practice of research, relatively little is known about the practice of inference itself. In this thesis, the present state of the practice of inference in psychology in both published articles and in the working environment of researchers was studied.

7.1 Summary of Findings

Chapter 2 describes a study in which it was examined to what extent 286 articles, submitted before and after the publication of the fifth edition of the APA manual, complied with the recommendations on inference this manual. Contrary to previous APA guidelines, the fifth edition stated that CIs are “in general the best reporting strategy”, because “they combine information on location and precision and can often be indirectly used to infer significance levels” (APA, 2001, p. 22). The findings in the study presented in Chapter 2 suggested that NHST is still used almost universally in psychological research, and CIs are seldom used. Despite changed guidelines, this pattern did not seem drastically different for articles submitted after the publication of the fifth APA manual compared to those before. The important mistake of accepting H₀ was made in about half of the articles, whereas claims of accepting the alternative hypothesis were found in approximately one fifth of the articles. Effect size was usually reported,
but hardly ever interpreted. The results confirm the expectation that the reasoning in most articles is mainly of a binary nature.

Chapter 3 reports a study in which the use of basic statistical techniques (i.e., $t$-test, ANOVA and regression), of 30 researchers in their own working environment was examined. They were asked to analyse fictitious data sets as they would analyse their own data sets. Afterwards, they had to write in a few sentences what they thought they could conclude about the population, based on these data. It was found that violations of assumptions were seldom checked for, and that visualising data, as is often taught in statistics classes, was far from a standard procedure for most subjects. Furthermore, again, NHST seemed the standard means for inferential conclusions, whereas CIs were hardly ever used. The fact that this was found in published articles as well as in the practice of statistical analysis suggests that it is not only the risk of not getting an article published that withholds researchers from using CIs. Not surprisingly, the main reason why people seem to use NHST is habit: CIs are hardly ever considered as an alternative, and only irregularly as an addition to NHST. Furthermore, again the conclusions of the researchers contained frequent occurrences of accepting $H_0$, and, on a smaller scale, of interpreting significance as equivalent to the existence of a population effect, and effect size was seldom interpreted.

In Chapter 4, a study is presented which explores why researchers use some well-known statistical techniques in a specific way. To study explanations for the observed use of these techniques, the same 30 subjects as in Chapter 3 were questioned about their task-related behaviour. In advance, we expected that rational explanations would at least partly account for these findings. It was our expectation that researchers would be aware of the problems, but that they were doing inference in a pragmatic way. However, we found that lack of knowledge of the used techniques and
nonchalance seemed to be key factors for explaining the observed behaviour. Visualising data is often skipped when analyzing data, seemingly for pragmatic but, from a statistical viewpoint, unwarranted reasons. It was striking that most researchers thought visualisation to be a crucial part in the process of analysing data, and even stated it was part of their own standard procedure when analysing data, despite the fact that they showed just the opposite only moments before. Apparently, researchers’ task performance was more “quick-and-dirty” than they were aware of. The subjects proved to have little knowledge about which assumptions need to be met for well-known techniques. Even if these assumptions were given by the experimenter, they often did not know how to check for possible violations of these assumptions.

Lack of attention to CIs during researchers’ education, resulting in missing the information that CIs are in fact an inferential means, seems partly responsible for the fact that CIs are hardly ever and NHST is almost always used for answering inferential questions. On the other hand, knowledge about how to interpret both NHST outcomes and CIs is alarmingly limited, although people seem much more aware of their lack of knowledge of CIs. Furthermore, although some researchers had the expectation that the use of CIs in their articles lowers their chances of getting the article published, none of them could mention one example in which this was the case.

In Chapter 5, a study was described in which researchers were asked to estimate both the probability that there would be an effect in the population and the probability that an exact replication would result in a significant effect, given varying $p$-values and sample sizes. The estimates differed widely across subjects. Despite the widespread use of the $p$-value in psychological research and the relevance of these probabilities, apparently interpretations of the same data differ largely between researchers. This,
however, could also be explained by the alleged complexity of the task. Despite the fact that we think both probabilities should be important for researchers when interpreting data, many subjects complained about the difficulty of this task afterwards. In spite of the large variance, it was found that, in general, the subjects’ estimates for the two probabilities increased with fixed $n$ and decreasing $p$. It was also found, however, that estimates for the two probabilities increased with increasing $n$ and fixed $p$. Given certain assumptions, it can be proven that both probabilities are actually independent of $n$. This indicates that researchers seem to overlook the fact that sample size is already accounted for in statistical outcomes like $p$-values.

Chapter 6 describes an experiment in which researchers’ interpretations of inferential results presented by means of CIs and NHST were compared. It was found that researchers made fewer mistakes when confronted with data presented by means of CIs instead of by means of NHST, and also made more references to effect size. Besides that, researchers seemed more certain of significant effects when presented by means of NHST than by means of CIs, whereas such a difference could not be found for non-significant effects. This suggests that CIs stimulate people to make more conservative estimates. Note that is not necessarily desirable behaviour, it only indicates that apparently both ways of presenting data lead to different interpretations. This outcome should not be regarded as an unconditional plea for the replacement of NHST by CIs in general. In both cases, awareness of the essentials of inference, which we showed is rather limited in general, seems more important than which inferential technique is used.

All in all, many psychological researchers, if not most of them, seem not fully aware of the importance of a balanced way of inference. Many may see it as a necessary step in their analyses and articles, but not as a tool to gain insight into their data. Often, inference seems only used as a way to
distinguish interesting results from less interesting ones. Even more so, it seems that the goal of many studies is to make exactly that, often unjustifiable, distinction, instead of interpreting the effect size that can be expected in the population.

7.2 Limitations of the Studies

We argue that the present state of the practice of inference is worrying, thus supporting with evidence the numerous articles pleading for a change in this practice. However, some limitations of the studies may limit the scope of the conclusions. First, the tasks subjects were given in the studies described in this thesis were different from the task of analyzing their own data. Therefore, it might be argued that the results found in this thesis are not representative for the practice of doing research. Researchers have more interests in studying their own data thoroughly, than analyzing fictitious data for someone else. This does not, however, necessarily mean that the frequencies of erroneous task behaviour we found here, are overestimates of what happens in practice. Subjects were well aware that they were being observed or that their results would be analysed afterwards, and therefore it might be expected that subjects were even more attentive to possible problems and pitfalls than they would normally be. For this reason, we think there is little reason to believe that the estimates of undesired task behaviour based on these studies could be an overestimate of what happens in reality.

A second important limitation of our studies was the relatively small sample size in the sample-based studies (i.e., all studies except for the one described in Chapter 2), resulting in relatively low power. In these studies, the population we sampled from was that of researchers in psychology working at Dutch universities. For pragmatic reasons, we chose mainly experienced Ph.D. students as subjects. Because their number is limited, and
because we did not want subjects to take part in more than one study (with the exception of the strongly related studies described in Chapters 3 and 4), it was relatively hard to find larger numbers of subjects. For a closer study of the practice of inference, it would be interesting to broaden the focus of the population of interest to researchers outside the Netherlands, as well. Nonetheless, there do not seem obvious reasons why Dutch researchers should behave very differently from researchers in other countries with relatively high university standards, so whereas lack of power due to a limited number of potential subjects might be a problem, representativeness is probably not.

A final limitation may lie in the fact that subjects frequently complained about the complexity and monotony of the tasks, which might have had the undesired consequence of subjects rushing to finish the tasks described in this thesis, thus resulting in task behaviour that is clearly different from task behaviour that they would show when not being observed. These effects caused by the artificial setting cannot be excluded in individual cases. On the other hand, the tasks were designed to resemble research practice, so both complexity and monotony might well play a similar role in actual research practice. Moreover, we expect that the desire to perform well in a task in which one is observed will, in most cases, prevent this effect from having too much influence. This is also supported by the fact that most subjects took a substantial amount of time for performing the tasks, indicating that they performed the task seriously.

7.3 Gravity of the Present State of Inference in Psychology

Our key recommendations for the present practice of inference in psychological research focus on the lack of attention for effect size, and on a binary and therefore too rigid interpretation of the results. As mentioned before, these recommendations are also given by the fifth edition of the APA
Manual. In our opinion, it would improve the scientific practice if people would follow them on a large scale, not because it is important in itself that APA regulations are followed, but because the recommendations are useful and usable. The APA Manual is regarded a standard for scientific publications in the behavioural sciences and thus it can be taken/assumed that these recommendation are known. To many researchers, however, the discussion on the alleged problematic practice of inference may not seem directly relevant for them. Probably, many of them are either not aware of this discussion and of changed guidelines in the first place, or do not consider this discussion important.

The importance of these matters greatly depends on the kind of research questions researchers want to study. As stated before, many authors seem to assume that researchers are actually interested in the size of a population effect, and not only in whether there is a population effect or not. This, however, is not necessarily the case. It might be that many researchers are mainly interested in whether a certain variable plays a role in a certain psychological process, in whether two variables are related, or in whether one group has a higher mean than another group. In most psychological journals, it is usually such research questions that are described, more than research questions concerning the size of an effect. Indeed, most null hypotheses in NHST in psychological research reflect “no effect”, whereas a nonzero lower bound for a positive effect (or a nonzero upper bound for a negative effect) might be more interesting as null hypothesis.

As stated before, effect size is usually mentioned in articles, but is seldom interpreted. If a researcher only wants to verify whether there is an effect, the answer NHST provides might seem just as useful as the answer resulting from a CI, or a combination of NHST and effect size. The supposed advantages of CIs over NHST almost completely vanish when the size of an effect is not of interest to the researcher. Therefore, if the size of an effect is
generally not the main focus of what a researcher wants to know, it makes sense that the earlier mentioned criticisms on the usability of NHST do not have a large influence on the practice of inference in psychology. From this point of view, NHST is not as problematic as is often stated, and would even fit researchers’ wishes for answering their binary research questions, making NHST an appropriate technique in many instances.

However, the situation is not as simple as this. If the questions psychologists are interested in are focussed on whether there is an effect in the first place, there are still reasons why using NHST as a means to answer their research question is not the optimal strategy, despite the fact that it seems to answer their questions well.

First of all, as our results in Chapters 2, 4 and 6 show, many researchers make the mistake of accepting the $H_0$, so apparently they are at least sometimes interested in stating that $H_0$ is true or approximately true, that is, that the effect is negligible. By using NHST outcomes, however, it is relatively hard to make statements about an effect being small, negligible or even absent. CIs have as an important advantage over NHST outcomes that they allow for stronger conclusions about the viability of the null hypothesis than does NHST (Aberson, 2002).

Secondly, whereas the author might not be interested in testing other $H_0$s than the presented one, may be some readers are. A single CI makes it possible to test an infinite number of null hypotheses, whereas a NHST outcome gives information on only one $H_0$. By showing inferential results by means of CIs, this option is available to the reader.

A third and related reason why NHST outcomes are insufficient is that even if the author of an article is not interested in the size of an effect, the readers of the article might be. By presenting both information on whether there is an effect and on the size of an effect, readers who are interested in either of the two types of research questions can be satisfied
simultaneously. Note that just presenting a point estimate of the effect is not sufficient for this purpose: For an inferential interpretation of the data information on the degree of uncertainty in the estimate needs to be given as well. It seems needless to say that CIs seem more suited than NHST outcomes for providing the reader with just that.

A fourth reason why even researchers only interested in the size of an effect should not only rely on NHST outcomes, is an argument that relates to meta-analysis. A key goal in research is to combine information found in different studies in order to give a more general answer to the research question at hand, as is often done in meta-analysis. Usually, individual studies are relatively small and the samples are relatively homogeneous, compared to a combined sample used in a meta-analysis. Therefore, combining studies broadens the scope of the conclusions. For such a meta-analysis, measures of effect size are essential. This implies that even if a researcher is only interested in whether there is an effect, it is still crucial to present effect sizes for the case that the results will be used for a meta-analysis. Note that our studies showed that effect sizes are usually given (thus making meta-analysis possible), but that inferential information on the effect sizes is not given. The advantage of giving CIs is that effect sizes and the inferential information on the effect sizes are given automatically, and directly interpretable.

In the above, the assumption has been that researchers predominantly use NHST, because they are only interested in whether or not an effect exists, and not in its size. The situation might, however, also be the reverse: It could also be that researchers come up with dichotomous questions on whether or not an effect exists, because NHST is the default inferential approach. That is, in NHST a strict distinction is made between significant and non-significant results, and therefore researchers might erroneously think that those outcomes reflect the existence or the absence of
an effect respectively, and adjust their research questions to this binary outcome. Implicitly, the explanation of NHST’s popularity causing dichotomous questions means that researchers are actually interested in estimating sizes of effects as well, but adjust their research questions to what they think is required for the available or dominant inferential techniques. If this would be the case, it would be sufficient to focus more on the possibilities and advantages of alternative and less known techniques like CIs.

In summary, whatever the reasons for the dominance of NHST, focusing on the importance of the use of alternative inferential techniques seems necessary. For some researchers this will show them that there are possibilities, notably using CIs, to answer the questions they actually want to get answered, and for others this is necessary to make their work more available to readers and future researchers who want to use their data, and to give them the opportunity to make inferential statements about sizeable, but also about negligibly small effects.

7.4 Suggestions for Improving the Present State of Inference

The results in this thesis indicate that in the practice of inference in psychology incomplete, carelessly worded, or even incorrect inferential conclusions abound, despite the duration and intensity of the debate on the usefulness of NHST. Apparently, this mainly statistical debate did not have a decisive influence on researchers’ behaviour. Despite the fact that the importance of this problem is now explicitly recognized by an influential organisation as the APA, large changes in the practice of inference cannot be noticed so far.

Compared to some decades ago, some clear initiatives for change can be seen. The APA has changed its guidelines, the number of articles criticizing the state of the practice of inference has kept growing, meta-
analysis seems to have achieved a more prominent place in psychological research, and in more recent textbooks the gravity of problems with NHST seems more often stressed than before. These changes, however, seem still too non-committal, and do not seem part of a coherent plan to change this practice radically. Our studies showed that erroneous inferential conclusions still abound in scientific literature as well as in researchers’ work environment when analyzing data. For a faster change, both explaining and maintaining guidelines and improved education seem necessary. A top-down approach could be organised by the APA, but on a much larger scale than by only changing the prescriptions for manuscripts. In the guidelines given in the fifth edition of the APA Publication Manual, a clear explanation of why this change is needed and, more importantly, how a researcher could implement these changes when presenting his or her data, is missing. Whereas, for instance, the proper lay-out of a reference list of a manuscript is extensively described, including many practical examples, there is no such thing for the practice of inference, which has, arguably, more relevance to the readers of the articles.

On the other hand, as asserted in the previous paragraph, it should be made clear why research questions on the size of an effect, rather than on the mere existence of this effect, would improve researchers’ understanding of their topics. It would also be recommended that it be made clear for researchers who are interested to change their behaviour where they could go for advice. An idea would be to launch some kind of electronic consultation helpdesk for questions related to inference. This might sound time and money consuming at first sight, but if indeed, as for example Schmidt (1996) argues, the progress of sciences depends on this process, the costs are negligible compared to the total sum of research money and time in psychology.
Only changing guidelines themselves, and providing those with examples how to improve the inferential practice and giving people easy access to consultation on this matter is probably not enough. Editors of APA journals should be asked to cooperate in this process as well. Ideally, they should not only ask authors to submit manuscripts that satisfy the guidelines of the APA as far as inference is concerned, but they should also support them with suggestions on how to change their manuscripts in a proper way. If editors would do this on a large scale, one could expect this should at least have a little influence on the practice of inference. The case of Memory & Cognition (Finch et al., 2004), where Gregory Loftus tried to make such a change for the journal he was the editor of, showed that although some changes can be reached, this is a long and painstaking process. Moreover, editors and reviewers are usually researchers themselves, which means that it would probably also require them, and also the reviewers, to change their own inferential practice.

The importance of a balanced way of interpreting results should also be stressed in classes and in textbooks, although it must be noted that this is partly happening already. Consultation in universities should be aimed at the importance of a balanced way of interpreting results as well. Given stricter demands for manuscripts, researchers will probably eventually adjust their behaviour.

7.5 Predictions

The present state of inference in psychology is problematic, and given the amount of erroneous behaviour it is important to change this state. This is, however, unrelated to the question whether research questions should focus on the existence of an effect, or on the size of an effect as well. Research questions only studying whether there is an effect can sometimes be interesting, but giving readers who are interested in the size of the effect
the opportunity to get their inferential questions answered seems a good reason to encourage researchers to include information on both in every manuscript they submit. This can be done more easily be means of CIs than by means of NHST outcomes. Whatever technique is used, however, a proper understanding of it is absolutely necessary, and our studies show that this is seldom the case. The supposed advantages of CIs only apply if they are used properly. If adding CIs to graphs, or even replacing NHST outcomes by CIs would only lead researchers to interpret CIs as if they were NHST outcomes, the net effect would be zero. It is, therefore, not only important that researchers realise the importance of adding CIs, but more that of interpreting the confidence limits with respect to the scale at hand. If not, there is large risk that in some decades many more CIs will be seen in articles, without any change in the interpretation of the results. It can be argued that techniques based on Bayesian computations of a posterior distributions are an even more useful alternative, since they offer considerably more flexibility than traditional techniques, and, the techniques force people to think rather than just apply ready-made procedures (Lecoutre, Lecoutre & Poitevineau, 2001). The latter could well be one of the main causes of the frequently observed undesired task behaviour as described in this thesis. Despite the alleged usefulness of Bayesian techniques, the efforts that are required for a researcher to capture the basic principles of Bayesian statistics are probably obstructing the techniques to be used on a much larger scale.

A revolutionary change in the practice of inference in psychological research is highly desirable, but, given earlier described limitations, probably not attainable in the near future. Until then, statisticians, methodologists and researchers in general who are convinced that binary thinking in psychological research is undesirable and effect size estimates are needed should probably try to attain changes on a much
smaller scale in their own environment, and hope that the forces of habit are not strong enough to counteract these efforts.