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):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 29-04-2019
4. Explanations for the Use of Well-Known Statistical Techniques

Abstract
In their own working environments, researchers showed several undesirable ways of analysing and interpreting data, as described in Chapter 3. Data were seldom visualised, possible violations of assumptions were seldom checked for, and the interpretation of the results was often too unbalanced. These results are in line with the present state of inference found in articles. For the study presented in this Chapter, we investigated reasons for this unwanted behaviour. It was found that lack of knowledge and nonchalance seem to explain this behaviour, rather than rational reasons.

4.1 Introduction
In Chapter 3, we showed that researchers show a lot of undesirable task behaviour when analysing data in their own working environment. Little is known, however, about the reasons why this behaviour was shown. In fact, we expected that researchers might have rational reasons for their task behaviour, and that lack of knowledge is insufficient to explain this undesirable task behaviour. Various other explanations (as mentioned in the different sections) play a role. To study this, a questionnaire on reasons behind the observed behaviour was administered to the same participants as in Chapter 3. The results of this study are reported in the present chapter.

4.2 Method
Details on the selection of the participants used in the present study can be found in Chapter 3, section 3.2.1. Participants were asked to fill out a questionnaire, in which open questions were asked about the participants’ way of statistically analysing data sets and interpreting results in general. As
in Chapter 3, we focussed on visualisation of data, the checking for possible violations of assumptions, the selection of statistical outcomes, and acknowledgement of uncertainty in conclusions. The experimenter was present to answer any questions, and could stimulate the participants to answer more extensively in case that was necessary, or ask them to reformulate their answer when they seemed to have misread the question, given their initial answer. There was no time limit for filling out the questionnaire, but all participants took between 35 en 65 minutes to complete it.

4.3 Visualisation of the Data

We examined the following three explanations for possible lack of visualisation of data, as observed in the observation study: (1) The participant did not feel capable of visualising the data in the desired way, (2) There was lack of attention for the need of visualising data in education, or (3) It was not clear to the participant how to interpret the visualised data.

4.3.1 Operational Definitions

For studying possible lack of capability, we asked the participants whether in general they were always able to realise the visualisation of the data as they wanted. For studying possible lack of attention in education, we asked how much attention there had been in their education for the importance of visualisation, and for how data can be visualised. For the third explanation (not knowing how to interpret certain graphical representations of the data), we asked whether they always know how to interpret the graphical representations of the data they make. Prior to these questions, we asked them to indicate whether making some kind of visualisation is for them a standard procedure when analysing data, to be able to compare what they claim to do with what they showed to do in practice, as described in
Chapter 3, assuming that what they showed is representative for how they usually analyse their data.

4.3.2 Results

Twenty participants (67%, CI = [49,81]) said that for them, making some graphical representation is a standard procedure when analysing data, although 14 of them (47%, CI = [30,64]) did not do this for all six data sets. Fourteen (47%, CI = [30,64]) of the participants stated that it has never or almost never happened that they did not make some form of visualisation when analysing data, and in this case nine (30%, CI = [17,48]) of those participants did not make some form of visualisation for any of the six data sets.

Lack of capability to make the desired graphical representations does not seem to play an important role: only eight participants (27%, CI = [14,45]) said that it has at least once occurred that they were not able to make a desired graphical representation. All of them added that those cases should be considered exceptions.

Lack of attention to visualisation in education, however, can partly account for the absence of visualisation being a standard part of the analysis of data. Eighteen participants (60%, CI = [42,75]) said there was little or no attention to the need of visualising data during their education. Of course, this finding does not exclude the possibility that there was attention for the need of visualisation in their education, but that they forgot about this. Notably, 17 (57%, CI = [39,73]) of the participants said it had occurred that someone other than their data analysis instructors, mostly the supervisor (37%, CI = [22,55]), had encouraged them to make graphical representations for their data. Twenty-two participants (73%, CI = [55,86]) stated that they almost always or always know how to interpret visualisations they made; the remaining eight (27%, CI = [14,45]) asserted that there had been more than
one occasion in which they did not know how to interpret a specific visualisation.

4.4 Examining Possible Violations of Assumptions

We studied five explanations for the finding that assumptions were often not checked: (1) Unfamiliarity with the assumptions, (2) Unfamiliarity with how to check the assumptions, (3) Violation of the assumption not being regarded problematic, (4) Unfamiliarity with a remedy against an assumption violation, and (5) Influence of environment to ignore violations of specific assumptions. In the next section, these five explanations will be operationalised.

4.4.1 Operational Definitions

Unfamiliarity with the assumptions. In order to study familiarity with assumptions, the participants were asked for all three techniques to write down the assumptions they thought were necessary to check. We counted the assumptions that were mentioned, and compared these with our selection of assumptions. We selected the assumptions of normality and equality of variances because these are the most commonly described assumptions in statistical textbooks, and because these assumptions should be satisfied for any of the three selected techniques. The assumption of independent observations was not selected, because that is an assumption about the process of data collection, and not on the analysis of the data.

Unfamiliarity with how to check the assumptions. For examining unfamiliarity with how to check for possible violations of both assumptions, we simply asked whether the participants knew a way to investigate whether there was a violation of each of the two assumptions (normality and homogeneity of variance) for the t-test, ANOVA and regression analysis.
Specifying how to visualise the data in such a way that a possible violation was visible was categorized as a correct way of checking for assumption violations, even when no further information was given. The same holds for rules of thumb or tests for the assumption at hand that are incorporated in SPSS, such as Levene’s test for testing equality of variances.

Violation of the assumption not being regarded problematic. For techniques for which it has been shown that they are robust against certain assumption violations, it can be argued that it makes sense not to check for these assumptions, because the outcome of this checking process would not influence the interpretation of the data anyway. To study this explanation, we asked per assumption whether violation of this assumption was found important.

Unfamiliarity with a remedy against an assumption violation. In order to study the hypothesis of unfamiliarity with a remedy against a possible violation of an assumption, the participants were asked to mention a remedy against a possible violation, if they knew any. One could imagine that a possible violation of assumptions is not checked because in case a violation is found, the participant does not know how to solve this problem. Of course, this type of reasoning is not necessary in the case that the researcher is aware of the robustness of the technique at hand against a violation of the assumption. Again, our definitions for correct remedies were transforming the data (it was not required to specify which transformation), and increasing the sample size.

Influence of environment to ignore violations of specific assumptions. The fifth and last explanation is the possible influence of the environment to ignore violations of specific assumptions. It could be that,
although initially the participants were willing to check assumptions, their supervisor or a colleague told them not to do so, for example for pragmatic reasons. In order to study this hypothesis, the participants were asked whether it had ever occurred that someone told them it was not necessary to check for violations of any assumption.

4.4.2 Results

In Table 4.1, the results are presented for the first four explanations for lack of checking assumptions. Specifically, percentages of participants to whom each of the statements per combination of assumption and technique apply are given. For the fifth, this information was not available per assumption and therefore not given in Table 4.1. The results show that a majority of the participants were unfamiliar with the assumptions.

For every assumption, a minority of participants mentioned one of the correct ways to check violation of the assumption. Note that all mentioned ways were considered correct, implying that in all other instances no way to check violation of the assumption was mentioned.

The results show that for a majority of the participants the alleged robustness of an technique against assumption violations is not a reason not to check these assumptions in the first place. Frequently, the participants added that they had no idea whatsoever whether violation of an assumption was important or not.

Only in a minority of instances one of the earlier described remedies for a violation of an assumption was mentioned. This implies that there seems little knowledge among the participants on how to overcome a violation of one of these assumptions. This interpretation is supported by the finding that most participants admitted to never having looked for a remedy against a violation of an assumption.
### Table 4.1: Percentages of participants to whom each of the explanations apply per combination of assumption and technique. Between brackets are 95% CIs for the proportions in percentages.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Technique</th>
<th>( t )-test</th>
<th>ANOVA</th>
<th>Regression</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>Unfamiliarity with assumption</td>
<td>87%</td>
<td>93%</td>
<td>87%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td>[70, 95]</td>
<td>[77, 99]</td>
<td>[70, 95]</td>
<td>[80, 94]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unfamiliarity how to check assumption</td>
<td>53%</td>
<td>63%</td>
<td>60%</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>[36, 70]</td>
<td>[45, 78]</td>
<td>[42, 75]</td>
<td>[49, 68]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Violation not regarded problematic</td>
<td>27%</td>
<td>30%</td>
<td>27%</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>[14, 45]</td>
<td>[17, 48]</td>
<td>[14, 45]</td>
<td>[20, 38]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unfamiliarity with remedy</td>
<td>60%</td>
<td>67%</td>
<td>67%</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>[42, 75]</td>
<td>[49, 81]</td>
<td>[49, 81]</td>
<td>[54, 74]</td>
<td></td>
</tr>
<tr>
<td>Homogeneity of variance</td>
<td>Unfamiliarity with assumption</td>
<td>53%</td>
<td>57%</td>
<td>67%</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>[36, 70]</td>
<td>[.39, 73]</td>
<td>[49, 81]</td>
<td>[49, 68]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unfamiliarity how to check assumption</td>
<td>73%</td>
<td>60%</td>
<td>87%</td>
<td>73%</td>
</tr>
<tr>
<td></td>
<td>[55, 86]</td>
<td>[42,75]</td>
<td>[70, 95]</td>
<td>[63, 81]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Violation not regarded problematic</td>
<td>23%</td>
<td>30%</td>
<td>37%</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>[12, 41]</td>
<td>[17, 48]</td>
<td>[22, 55]</td>
<td>[22,40]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unfamiliarity with remedy</td>
<td>70%</td>
<td>63%</td>
<td>57%</td>
<td>63%</td>
</tr>
<tr>
<td></td>
<td>[52, 83]</td>
<td>[45, 78]</td>
<td>[.39, 73]</td>
<td>[53, 73]</td>
<td></td>
</tr>
</tbody>
</table>
Finally, nine participants (30%, 95%CI = [17,48]) said that it had happened at least once that someone told them not to check possible violations of an assumption. This proportion is substantial, but too low to indicate that suggestions from others had a large influence on the frequent absence of the checking of these violations.

4.5 Selecting Outcomes

We looked for reasons why effect size measures are often not used, and why it was often found that NHST is almost always used when analysing data, as opposed to CIs. For checking why conclusions are often not accompanied by effect size measures, we distinguished two possible explanations: (1) lack of interest in effect size, and (2) influence of others not to report effect size. To study the reasons why most researchers use NHST instead of CIs to present outcomes, we distinguished three possible explanations: (1) NHST is used by default because an alternative is not known, (2) There has been more attention in education for NHST than for CIs, and (3) The belief that the use of CIs lowers the chances of getting an article published.

4.5.1 Operational Definitions

To study the possible lack of interest in effect size, we asked whether the participants were interested at all in effect size, and whether they usually presented some effect size outcomes in their own work. To study the possible influence of others, we asked whether others had ever discouraged them from reporting effect size measures. We also asked whether others had explicitly asked them to add some measure of effect size.

For studying the first explanation for the preferred use of NHST, we asked how often the participants used NHST and CIs, and we asked whether the participants automatically use NHST when interpreting data analysis
results. For the second explanation, we asked the participants to estimate the ratio of the attention for NHST and CIs during their education. For the third explanation, we asked whether the participants thought using CIs influences the chance of getting an article published, and whether they had encountered such an influence in practice.

4.5.2 Results

Although Chapter 3 showed that effect size was hardly ever reported, 21 participants (70%, CI = [52, 83]) stated that they are always interested or very interested in the size of the effect. Apparently, however, for 14 of them (67%, CI = [45,83]) it was not important enough to report in their conclusions in the study reported in Chapter 3. Twenty participants (67%, CI = [49,81]) said they usually present some effect size measure in their own work.

Two participants (7%, CI = [1,23]) said they had been discouraged by their supervisors from reporting effect size measures in an article. Eight participants (27%, CI = [14,45]) reported that they were at least once explicitly asked to add effect size measures to the presentation of outcomes. It can safely be concluded that giving effect size or means is not standard practice when drawing conclusions on the basis of data analyses. As far as there is influence of others on the reporting of effect size, it seems to be an encouraging rather than a discouraging influence.

As far as possible explanations for the preference of NHST over CIs are concerned, the following was found. All but one participant said they had (almost) always used the significance test when analysing data. As might be expected, we found a large difference between the frequency of use of NHST and CIs in the present study as well. Eight (27%, CI = [14,45]) participants had at least once constructed a CI for their own research, while five participants had considered constructing one, but eventually abandoned
this idea. Five (17%, CI = [7,34]) participants said they always construct a CI when analysing data, despite the fact that none of them constructed a CI in the study presented in Chapter 3. Fourteen participants (47%, CI = [30,64]), however, said they knew that CIs can be used as an alternative for NHST. The other 16 (53%, CI = [36,70]) said that they were not aware of this, despite the fact that the use of CIs is strongly encouraged in the fifth edition of the APA manual (APA, 2001).

Our results suggest an overall lack of knowledge of CIs, although 21 participants (70%, CI = [52,83]) said the use of CIs was taught during their education (implying that for 30% this was not taught, or was forgotten). They estimated, however, that on average almost six times more time was dedicated to the explanation of NHST compared to the time spent on CIs.

Eight participants (27%, CI = [14,45]) mentioned that the supposed difficulty of getting an article published when using CIs is a reason to continue not using CIs. None of the participants, however, had ever actually heard of a situation in which a reviewer requested CIs to be removed from the manuscript. Moreover, three participants had been explicitly asked by the reviewers to add CIs to their manuscript.

4.6 Indicating Uncertainty when Presenting Conclusions

We studied two possible explanations for the fact that researchers frequently give oversimplified conclusions after analysing the data sets: (1) the unfamiliarity with a more careful way of interpreting results, and (2) the influence of the scientific environment.

4.6.1 Operational Definitions

For the first, we examined participants’ knowledge on the correct interpretation of NHST as well as CIs. In order to study whether or not this knowledge is present, participants were asked to define how they interpret a
significant effect, a non-significant effect, and a $p$-value. Definitions of a significant effect as it being probable that there is an effect in the population were considered correct, whereas definitions of a significant effect as proof that an effect exists were considered incorrect. Finding a non-significant effect, without further information, implies that one cannot give a statement about the existence of the effect, and therefore all definitions implying a non-significant effect being proof for the absence of an effect were considered incorrect. Note that these are in fact lenient judgements as to the correctness of the definitions of (non-)significance. The adequate interpretation of a $p$-value is that it is the probability that the found effect or a more extreme one would have been found if the $H_0$ were true. Defining the $p$-value as the probability that the effect would have been found by chance was thus considered incorrect.

For CIs, we asked the participants to define how, according to them, a 95% CI should be interpreted, and we asked them to interpret a 95% CI with specified margins. Both questions were considered to be answered correctly whenever the answers explicitly referred to the inferential character of the CI, for example by explicitly mentioning that it is an estimate for the population value. It was not required to report 95% in the answers. As far as the influence of scientific environment is concerned, we asked the participants whether it had ever occurred to them that someone else asked them to rephrase a carefully considered interpretation in such a way that a less carefully considered interpretation was made.

4.6.2 Results

Only 10% (CI = [3,27]) of the participants gave an adequate interpretation of a significant effect. Also 10% (CI = [3,27]) gave an adequate interpretation of a non-significant effect, and 3% (CI = [0,18]) gave an adequate interpretation of the $p$-value, while 7% (CI = [1,23]) gave an
almost correct definition, but without mentioning the necessary “or more extreme”-part. For CIs a comparable pattern can be seen: only 13% (CI = [5,30]) of the participants gave an adequate interpretation of a CI in general. Seventeen percent (CI = [7,34]) of the participants gave an adequate interpretation for a specific CI given fictitious study results. None of the participants gave an adequate definition for all outcomes mentioned above. These dramatic results suggest that the capability of researchers to interpret the results of the two most important inferential techniques in an adequate way is limited.

As far as the influence of the scientific environment is concerned, none of the participants reported ever having been asked to rephrase a carefully considered interpretation into a less carefully considered interpretation. Again, however, this finding can be due to the fact that most participants were not aware of a careful interpretation (with a measure of uncertainty included) in the first place.

4.7 General Discussion

In this study, we sought for explanations for why three well-known techniques were used in the way as observed in the observation study described in Chapter 3. More specifically, we questioned the participants on four parts of the task of analysing data, namely visualising the data, checking for possible assumption violations, the selection of outcomes for drawing conclusions, and the carefulness of interpreting the results.

The finding that most participants asserted that they were aware of the importance of visualising data, and even stated that they usually visualised data, contradicts the frequent absence of graphical representations of the data in the observation study in Chapter 3. Apparently, the researchers overestimated their own use of visualisation during task analysis, probably at least partly due to socially desirable answering behaviour. We suspect that
the researchers, in general, were not sufficiently aware of the importance of visualising data. This view is supported by the fact that a majority complained about the lack of attention for this matter in their education, and by the fact that also a majority had at least once been advised to add graphical representations of the data to their results, indicating that they had not done this before. This contradicts most researchers’ claim that visualising data is part of their standard way of analysing data. Because researchers were obviously aware that they were being observed during the observation study, we expect them to have performed better rather than worse than usual. For this reason, we think that their observed behaviour is probably a better indication of their use of visualisation in practice than their answers from the questionnaire on this issue.

An important explanation for the observed lack of examining possible assumption violations seems to be the lack of knowledge about the assumptions, especially about the assumption of normality. Even when the assumptions were given, only a minority of the participants could mention a way to check the assumptions. Contrary to our expectations, the participants mentioned little influence of their environment not to check for assumption violations. Furthermore, we did not find evidence that lack of knowledge about a remedy against an assumption violation is an important explanation for the observed absence of checking for assumptions: Usually the researchers do not know the assumptions in the first place and therefore cannot be expected to know of a remedy. The same reasoning holds for the explanation of lack of knowledge about the robustness of a technique against assumption violations. Given the frequently observed lack of knowledge regarding the assumptions, it thus seems that if we consider checking for assumption violations essential for analysing data in order to prevent unjustified conclusions, the importance of this should be stressed more in education and in editorials.
The fact that researchers stated that they are usually interested in effect size, and usually report some form of effect size, contradicts the results in the observation study in Chapter 3, in which we found that effect size was seldom interpreted. It was also found that NHST seems to be used mainly because of its dominance in contemporary literature. The opposite seems to apply for the use of CIs: Most participants encountered CIs relatively rarely in their education as well as in most articles they read. Although not based on findings in studies described in this dissertation, the same holds probably for statistical packages. For example, in order to get CIs in SPSS, the statistical package that is very popular in the social sciences, for most techniques the user has to change the default settings in order to get these CIs, whereas \( p \)-values are always clearly given. In SPSS a CI is only automatically given in case of the \( t \)-test. In case of regression and ANOVA CIs can be obtained only with additional effort, although CIs are given when the “explore” option in SPSS is used. A second, and probably even more important reason, is the lack of knowledge of the participants of CIs, with 87% giving a definition that did not acknowledge the inferential nature of a CI. Our data did not support the explanation that NHST is reported because editors demand that manuscripts consist of results interpreted by means of NHST. This can, however, be explained by the fact that there was little reason for reviewers to do this, because in almost all submitted manuscripts NHST results are given anyway. The question why data analysis results are often interpreted in a too simplistic way when using an inferential technique, seems to be at least partly explained by researchers’ overwhelming lack of knowing how to interpret NHST outcomes properly, and even more how to interpret CIs.

In summary, contrary to our expectations, the reasons behind the inadequate task behaviour as shown in the observation study often do not seem the consequence of well-considered reflections. It seems that lack of
knowledge (when checking assumptions and interpreting inferential results) and carelessness (when visualising data, checking assumptions, and carefully interpreting inferential results) explain this behaviour for the most part. This does not imply that we can blame all errors on the researcher: Apparently, the scientific community tolerates this behaviour. Researchers, working under high pressure and having little time seem to use statistics as pragmatically as is tolerated by their scientific environment. The many suggestions that statisticians have given in consultations and in articles, and even in the APA Manual, seem too non-committal, and the tolerance for an opportunistic use of statistics is too high.