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.
3. The Use of Well-Known Statistical Techniques

Abstract

In the literature on problems with inference, little attention has been given to the way researchers use inference when analysing data in their working environment. In the present study, researchers’ behaviour during the analysis and interpretation of fictitious data sets was observed. It was found that the data were often not visualised, that researchers hardly ever checked for possible violations of assumption, that NHST was used as the standard inferential technique, and that the interpretational mistakes found in articles are also found in practice.

3.1 Introduction

When a researcher in the behavioural sciences analyses a data set, he or she makes several statistical choices, some probably more consciously than others. Amongst these choices are the following: 1) Is visualisation of the data needed before they can be analysed? 2) Is it necessary to examine possible violations of assumptions? 3) Which statistical outcomes are needed for the interpretation of the data? 4) How can the inferential outcomes and the following conclusions be presented? For all these choices, much has been written on which choices are preferable from a data analytic point of view in given situations. Little is known, however, on which choices are actually made by researchers for given data sets. The results of a study investigating these choices is described in the present Chapter. In Chapter 4, a study will be presented that investigated why these particular choices are made.
3.1.1 Visualisation of the Data

Two reasons to visualise data are to get a first impression of the results (APA, 2001), and to present the data in a manner providing insight into the data in an efficient way (e.g., Loftus, 1993). Although this may seem obvious, it has been argued that the importance of visualising data is still often underestimated in practice (e.g., Loftus, 1993). And although in the past decades an increase in the use of graphs and charts in journals in the social sciences can be seen, usually the graph follows rather than precedes the analysis (Maltz, 2006). Loftus stressed the importance of visualisation as follows: “A picture is worth a thousand $p$ values” (p. 250), stressing both the importance of visualising and the, according to him, often observed overestimation of the importance of $p$-values. Others (e.g., Morrison & Weaver, 1995) disagree with Loftus that graphical representations can replace NHST, and point out that making a “plot plus-error bar”, as Loftus suggested, is not always as straightforward as Loftus seems to suggest. They argue that especially in the case of mixed designs and factorial within-subjects designs, there is no single value that appropriately represents the standard error, which is essential to Loftus’ plot plus-error bars. Nevertheless, there seems general agreement on the idea that making graphical plots of data is valuable. The statements regarding the alleged lack of visualising data in research practice in the mentioned articles, however, are without any reference, and therefore seem to be based on experience of the authors rather than on study outcomes.

3.1.2 Examining Possible Violations of Assumptions

Most statistical techniques require that one or more assumptions be met, or, in the case that it has been proven that a technique is robust against a violation of an assumption, that the assumption is not violated too extremely. Many articles have been written on the robustness of certain techniques with
The Use of Well-Known Statistical Techniques

The Use of Well-Known Statistical Techniques

3.1.3 Selecting Outcomes

After data have been analysed, conclusions need to be drawn based on the statistical outcomes. In published articles, researchers have to decide which statistical outcomes they are going to use for the presentation of their data. From previous research (e.g., Finch et al., 2001), we know that in published articles some form NHST is included, and that it is usually the p-value that is interpreted. The same study by Finch also shows that confidence intervals (hereafter, referred to as CIs) are rarely reported, means
or effect sizes are most of the time reported but seldom connected to the other outcomes: The significance of an effect seems far more important than the observed mean or effect size.

In the sequel, we will study which outcomes are used when analysing data for the first time. The outcomes as are reported and interpreted in a paper are not necessarily those that are interpreted during the first analysis of the data set. For this reason, this study might give more direct insight in what outcomes researchers are interested in, since they are not restricted by what they think is required for a manuscript. To our knowledge, it has not been studied so far to what extent $p$-values, CIs, and effect sizes are actually examined before a report on the study is written.

3.1.4 Indicating Uncertainty when Presenting Conclusions

As stated in Section 3.1.3, most studies report NHST outcomes. As mentioned in Chapter 1, almost since the introduction of NHST as a technique, its usefulness has been questioned. Whether due to the technique or to the user, many interpretational errors are made when using NHST in published articles (see Chapter 2, i.e., Hoekstra et al., 2007; Finch, 2001), as well as under experimental conditions (Tversky & Kahneman, 1971; Oakes, 1986; Lecoutre, Poitevineau & Lecoutre, 2003).

Although many articles (e.g., Bakan, 1966; Batanero, 2000; Cohen, 1994; Falk & Greenbaum, 1995; Lecoutre, Poitevineau & Lecoutre, 2003; Meehl, 1978; Vacha-Haase, 2001.) discuss misinterpretation of NHST outcomes, to our knowledge, no systematic observational study into possible explanations for this behaviour has been published. The following questions remain unanswered: Do researchers not know how to interpret NHST outcomes, or might there be another explanation of why often inadequate interpretations of data sets are observed?
3.1.5 How to Answer Research Questions

In the process of statistically analysing data, various decisions are taken. It is known that in taking these decisions frequently interpretational errors are made, although in many cases this has not been directly observed. The present study examines how researchers analyse data in their own working environment. In Chapter 4, a study will be presented to examine why such interpretational errors are made. Are researchers just ignorant of misconceptions in their way of analysing data, or careless, or are they with reason pragmatic? And if the latter is the case: What are these reasons to explain their behaviour?

In the present study, researchers were observed while analysing fictitious data sets. Specifically, we observed whether certain frequently made mistakes, such as failing to check for possible violations of assumptions, accepting $H_0$, or ignoring effect sizes, were shown.

3.2 General Method

3.2.1 Participants

For the studies described in this Chapter and in Chapter 4, 30 Ph.D. students, 13 men and 17 women, working at psychology departments throughout the Netherlands were used as participants. All of them had at least two years experience conducting research at the university level, and none of them was or had been working in a methodology or statistics department. Ph.D. students were selected because we thought that, in contrast to advanced researchers such as professors, they are actively involved in the collection and analysis of data. Moreover, they were likely to have had their statistical education relatively recently, assuring a relatively up-to-date knowledge of statistics. They were required to have at least once applied a $t$-test, a linear regression and an ANOVA, although not necessarily in their own research project. The participants were selected from
universities in Tilburg, Groningen and Amsterdam, three cities in different regions of the Netherlands. From every city, ten participants were selected, as follows. To reduce possible university biases, first a list with psychology Ph.D. students per city was made. Then, the Ph.D. students were selected from that list until we found ten who were willing to participate. In order to keep the sample as heterogeneous as possible in order to be able to draw more general conclusions, we selected Ph.D. students from social, developmental, clinical, experimental, and industrial and organisational psychology groups.

3.2.2 Analysis

All results are presented as percentages of the total number of participants or of the total number of tasks, depending on the specific research question. CIs are given, but should be interpreted cautiously, because the sample cannot be regarded completely random.

The CIs for percentages were calculated by means of the plus four estimate (Agresti & Caffo, 2000). In cases that this method resulted in CIs with limits smaller than 0 or larger than 100, these were corrected to 0 and 100 respectively. All CIs are 95% CIs.

Given the relative recency of the participants’ statistical education, it was expected that their knowledge of statistics would be up-to-date, and for that reason it was expected that the frequency of misuse of statistics we found will probably be an underestimation of that for a random group of similar size of other researchers. Therefore, the lower bounds of the CIs can be considered a very conservative estimate of the frequency in the population of all Dutch researchers in psychology.
3.2.3 Tasks

All but three participants performed the two tasks at their own workplace. In the remaining three cases, an unoccupied room on the participants’ department was used, in order to not disturb one or more roommates. All participants needed between 30 and 75 minutes to complete the task.

3.3 Details of Method, and Results

Participants were asked to analyse six data sets. Because every participant indicated that he or she used SPSS as a standard statistical package to analyse data, the data were offered in SPSS format. For every data set a short description of a research question was given, without giving a hypothesis. The participants were asked to analyse the data sets and interpret the results as they would analyse and interpret their own data sets. They were allowed to consult any statistical books or the internet, although only two used this opportunity. It was mentioned that they were to write down their conclusions, taking into account what these results told them about the population from which the samples were drawn.

The short description of the research questions was written in such a way to suggest that, respectively, a $t$-test, linear regression and ANOVA were most suitable for analysing the data sets without explicitly naming the analysis technique. The results of a pilot of the task indicated that the descriptions were indeed sufficient to choose the desired analysis technique. The $t$-test, ANOVA and regression analysis were chosen because they are relatively simple, frequently used, and because it was expected that most participants would be familiar with those techniques. In Figure 3.1, an example of such an instruction is presented. In this case the participants were expected to use a $t$-test. This example has been translated from the Dutch.
All six research question descriptions, also in translated form, can be found in the Appendix to this Chapter.

A researcher is interested in the extent to which group A and group B differ in cognitive transcentivity. He has scores of 25 randomly selected participants from each of the two groups on a cognitive transcentivity test (with the range of possible scores from 0 to 25). Column 1 of the SPSS file gives the scores of the participants on the test, and column 2 the group membership (group A or B).

Figure 3.1: An example of one of the research question descriptions. In this example, participants were supposed to answer this question by means of a t-test.

Note that “transcentivity” is a fictional concept, and was used in order to prevent biased conclusions based on prior knowledge. Fictional concepts were only used for those tasks where a t-test was expected, and therefore this argument of preventing bias does not hold for conclusions for the other tasks.

In almost all instances the expected technique was chosen (in the remaining instances, ANOVA was used to analyse the data sets that were supposed to be analysed by means of a t-test. However, ANOVA in this case is completely equivalent to an independent-samples t-test).

The six constructed data sets differed with respect to the effect size, the significance of the outcomes, and to whether there was a strong violation of an assumption. Table 3.1, shows the different conditions, with four of the six data sets containing significant effects; for one of the two data sets for which a t-test was supposed to be used, the assumption of normality was clearly violated, for one of the two data sets for which ANOVA was supposed to be used, the assumption of heterogeneity was clearly violated,
and effect size was relatively large in three data sets and relatively small in the other data sets.

Table 3.1: An overview of the properties of the six scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Technique to be used</th>
<th>Effect size</th>
<th>p-value</th>
<th>Violations of assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>t-test</td>
<td>Medium</td>
<td>.04</td>
<td>Normality</td>
</tr>
<tr>
<td>2</td>
<td>t-test</td>
<td>Very small</td>
<td>.86</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>Regression analysis</td>
<td>Large</td>
<td>.00</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>Regression analysis</td>
<td>Medium</td>
<td>.01</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>ANOVA</td>
<td>Large</td>
<td>.05</td>
<td>Homogeneity</td>
</tr>
<tr>
<td>6</td>
<td>ANOVA</td>
<td>Close to 0</td>
<td>.58</td>
<td>None</td>
</tr>
</tbody>
</table>

To get more information on which choices were made by the participants during task performance, and why these choices were made, participants were asked to think aloud during task performance. This was recorded on cassette. During task performance, the selections made within the SPSS program were noted by the first author, making use of previously conducted task analyses. Furthermore, participants were asked to save the SPSS syntax files. For the analysis of the task performance, the information from the notes, the recordings and the syntax files were combined.

In order to know how techniques are used, it is important to know which necessary part of the task (that is, drawing inferential conclusions on the basis of a data set) was neglected or not executed correctly. As stated before, in this study we focused on (1) whether the data are represented visually, (2) whether possible violations of assumptions are examined, (3) which outcomes are used for the interpretation of the results, and (4) how these outcomes are interpreted.
3.3.1 Visualisation of the Data

3.3.1.1 Operational definitions. To study visualisation, we checked for each participant and for each task whether the participant made any kind of visual representation of the data. Any of the graphical representations available in the menu “Graphs” in SPSS were considered visualising data, thus including scatter plots and boxplots.

3.3.1.2 Results. In total, each of the 30 participants analysed six data sets, adding up to a total of 180 analyses, 60 for each type of analysis. Only six of the participants actually visualised their data at least once for each of the six data sets they analysed before filling in the questionnaire, and only nine participants did this at least four times. In total, in 76 of the 180 analyses (42%, CI= [35,50]), the participants visualised the data. For the t-test, this was 33% (CI= [23,46]), for ANOVA 37% (CI= [26,49]) and for regression the data were most frequently visualised: 57% (CI= 44,68]). If participants visualised data, the plots or graphs the participants made were usually histograms or boxplots for t-test and ANOVA, whereas for regression this was mostly a scatterplot. The results show that, at least for t-test and ANOVA data, visualising the data seems not to be part of researchers’ standard procedure for analysing their data.

3.3.2 Examining Possible Violations of Assumptions

For each of the three techniques, several assumptions can be distinguished, but for practical reasons we limited ourselves to the important assumptions of normality and of homogeneity of variance (sometimes referred to as the assumption of equal variances, hereafter referred as the assumption of homogeneity of variance) for all three techniques. The assumption of normality requires that the scores in the population in case of a t-test or ANOVA, and the population residuals in case of a regression
analysis be normally distributed. The assumption of homogeneity of variance requires equal population variances per group in case of a t-test or ANOVA, and equal population variances for every value of the independent variable for regression. The assumption of independent observations was not taken into account because our focus is on data analysis and not on data collection. Also, we chose to ignore the assumption of linearity for regression analysis, because this assumption does not need to be met for the other two techniques. This, however, should by no means be interpreted as an implicit statement about the importance of this assumption.

3.3.2.1 Operational definitions. We counted occurrences of cases in which the participants checked for violations of the assumptions of normality and for assumptions of homogeneity of variance. We also counted occurrences of checking for irrelevant assumptions, such as equal group sizes for the t-test, or normality of all scores for one variable (instead of checking for normality per group) for all three techniques.

A check for the assumption of normality was recorded if, for the t-test and ANOVA a graphical representation of the different groups was requested, except in case this was only used to detect outliers. Merely observing the data, without making a visual representation of them, was also considered insufficient. For regression analysis, a check of the assumption of normality was recorded if a scatter plot or a residual plot was made, and was explicitly checked for a normal pattern or an absence of a normal pattern in such a plot. Deciding whether this was done explicitly was based on whether the participants made any reference to the normality when thinking aloud. A second option was to make a QQ- or PP-plot of the residuals.

As far as the assumption of homogeneity of variance is concerned, for the t-test and ANOVA four ways of checking for this assumption were considered adequate ways of checking: The first was to make a graphical
representation of the data in such a way that difference in variance between the groups was visible. This includes boxplots and scatter plots, provided that they are given per group. A second way was to make an explicit reference to the variance of the groups, and a third way was to conduct a test for equality of variance (e.g., Levene’s test). A final possibility was to compare standard deviations of the groups in the output, without or with making use of a rule of thumb to discriminate between violations and non-violations. For regression analysis, a scatter plot or a residual plot was considered necessary to check the assumption of homogeneity of variance. Although the assumption of homogeneity of variance assumes equality of the population variations, the participants were not required to make an explicit reference to the population.

3.3.2.2 Results. Violation of, or conformance with the assumptions of normality and homogeneity of variance was correctly checked 22 times and 57 times respectively (out of 180 cases for both assumptions). Table 3.3 shows for each of the three techniques how frequently possible violation of the assumptions of normality and homogeneity of variance occurred, and whether the checking was done correctly given the earlier definitions. Note that the assumption of normality for regression was never checked correctly. In the few occasions normality was checked, it was, erroneously, the normality of the scores instead of the residuals that was checked.

The assumption of homogeneity of variance for ANOVA was checked relatively often. An explanation for that is the fact that a clear violation of this assumption can often be directly deduced from the standard deviations, whereas measures indicating normality are less common. Another explanation for the fact that the assumption of homogeneity of variance was more frequently checked for than the assumption of normality, is the observation that most participants seem familiar with a rule of thumb
Table 3.3: The percentages of the 60 possible cases in which violations of the assumption of normality, and the assumption of homogeneity of variance were checked at all, and were checked correctly for the t-test, ANOVA and regression. Between brackets are 95% CIs for the percentages.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Technique</th>
<th>t-test</th>
<th>ANOVA</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>Checked</td>
<td>23% [14, 36]</td>
<td>30% [20, 43]</td>
<td>12% [6, 23]</td>
</tr>
<tr>
<td></td>
<td>Correctly checked</td>
<td>17% [9, 28]</td>
<td>20% [12, 32]</td>
<td>0% [0, 7]</td>
</tr>
<tr>
<td>Homogeneity of variance</td>
<td>Checked</td>
<td>33% [22, 46]</td>
<td>55% [42, 67]</td>
<td>7% [2, 17]</td>
</tr>
<tr>
<td></td>
<td>Correctly checked</td>
<td>33% [22, 46]</td>
<td>55% [42, 67]</td>
<td>7% [2, 17]</td>
</tr>
</tbody>
</table>

to check whether the assumption of homogeneity of variance for ANOVA is violated (e.g., largest standard deviation is larger than twice the smallest standard deviation), whereas such rules of thumb for checking possible violations of the assumption of normality were not known.

Apart from the assumptions of normality and homogeneity of variance, only the checking of the assumption of linearity for regression was observed in a few instances; no other known assumption seems to have been checked.

3.3.3 Selecting Outcomes

3.3.3.1 Operational definitions. As far as selection of analysis outcomes is concerned, our two main questions were: “Is effect size reported when interpreting results?” and “How do researchers deal with the consequences of sampling variations?”. In order to examine whether effect size, NHST or CIs were used for interpretation of the results, the following was required. Effect size was considered to be reported whenever there was some kind of reference to the means, the differences between the means, or the standardised differences between the means. NHST was considered to be
used whenever there was a reference to the $p$-value, or to the significance of the test. A CI was considered to be used whenever one was mentioned, even if it was not given any further interpretation.

3.3.3.2 Results. Effect size according to our definition was found in only 12% (CI = [8,18]) of the conclusions. None of the 180 conclusions contained a CI, and 174 contained NHST results, again stressing the central position of this technique. Note that in the other six conclusions, the participants decided not to make an inferential statement at all because of violation of one of the assumptions. Thus effectively in all cases where inferential conclusions were found acceptable, these were based on NHST. In 30% (CI = [24%,37%]) of the conclusions only the significance of the results was mentioned, without either a $p$-value or a reference to effect size.

3.3.4 Indicating Uncertainty when Presenting Conclusions

For getting information on whether oversimplified interpretation of the results had taken place, we assessed whether the following two misinterpretations were made: accepting the $H_0$, and accepting the alternative hypothesis. With the latter, we refer to the conclusion of being certain of the existence of a population effect, based on a significant finding.

3.3.4.1 Operational definitions. Accepting the $H_0$ was said to have occurred whenever the summary contained a phrase indicating the absence of an effect. For example, sentences like “There is no effect”, or “The groups do not differ as far as means are concerned” are considered occurrences of this misinterpretation. Occurrence of accepting the alternative hypothesis was said to have taken place when a statement was made about the population parameter without reflecting any doubt or uncertainty. “The groups differ”, “There is a real difference” or “It is proved that there is a
difference” were considered a misinterpretation of this kind, whereas “The groups probably differ”, or “The sample means differ” were not. Note that since participants were explicitly asked to make a statement about the population based on the sample results, a statement like “the groups differ” is either unwarranted or completely trivial, because, in real-life situations, all groups differ to some extent.

3.3.4.2 Results. In 70% (CI = [57, 80]) of the instances in which a non-significant effect was found, the $H_0$ was accepted, according to our definition of this misinterpretation. In most of these cases, it was said that there was no effect. In 50% (CI = [38, 62]) of the instances that a significant effect was found, a phrase indicating certainty about the existence of the effect in the population was used. It can therefore be concluded that occurrences of oversimplified interpretations of the data sets were found relatively frequently.

3.4 Conclusions

In the present study, it was found that often the data were not visualised, possible violations of assumptions were seldom checked, effect size was often not taken into consideration when drawing conclusions, CIs were never constructed to interpret the results, $p$-values were always used as the main outcome to base the conclusions on, and these conclusions often contained interpretation errors. So, despite the fact that the techniques had been used before and were said to be well-known, and despite the fact that the data sets were relatively small and unproblematic (i.e., no missing values and no data points with unrealistic values), the task behaviour was in general insufficient.

We think that these outcomes are representative for the way researchers analyse data in their working environment. Of course, they are
usually not being observed during analyses, but we expect that this would improve rather than worsen their performance.

3.5 Appendix

Research question descriptions. In this Appendix, the six research question descriptions are presented in translated form. Descriptions 1 and 2 were supposed to be answered by means of a t-test, Descriptions 3 and 4 by means of regression analysis, and Descriptions 5 and 6 by means of ANOVA.

1. A researcher is interested in the extent to which group A and group B differ in cognitive transcentivity. He has scores of 25 randomly selected participants from each of the two groups on a cognitive transcentivity test (with the range of possible scores from 0 to 25). In column 1 of the SPSS file the scores of the participants on the test are given, and in column 2 the group membership (group A or B) is given.

2. A researcher is interested in the extent to which group C and group D differ in cognitive transcentivity. He has scores of 25 randomly selected participants from each of the two groups on a cognitive transcentivity test (with the range of possible scores from 0 to 25). In column 1 of the SPSS file the scores of the participants on the test are given, and in column 2 the group membership (group C or D) is given.

3. A researcher is interested to what extent the weight of men can predict their self esteem. She expects a linear relationship between weight and self esteem. To study the relationship, she takes a random sample of 100 men, and administers a questionnaire to them to measure their self esteem (on a scale form 0 to 50), and the participants’ weight is measured. In column 1 of
the SPSS file the scores on the self esteem questionnaire are given. The second column shows the weights of the men, measured in kilograms.

4. A researcher is interested to what extent the weight of women can predict their self esteem. She expects a linear relationship between weight and self esteem. To study the relationship, she takes a random sample of 100 women, and administers a questionnaire to them to measure their self esteem (on a scale form 0 to 50), and the participants’ weight is measured. In column 1 of the SPSS file the scores on the self esteem questionnaire are given. The second column shows the weights of the men, measured in kilograms.

5. A researcher is interested to what extent young people of three nationalities differ with respect to the time in which they can run the 100 meters. To study this, 20 persons between 20 and 30 years of age per nationality are randomly selected, and the times in which they run the 100 meters is measured. In column 1 of the SPSS file their times are given in seconds. The numbers “1, “2”, and “3” in column 2 represent the three different nationalities.

6. A researcher is interested to what extent young people of three other nationalities differ with respect to time in which they can run the 100 meters. To study this, 20 persons between 20 and 30 years of age per nationality are randomly selected, and the times in which they run the 100 meters is measured. In column 1 of the SPSS file their times are given in seconds. The numbers “1, “2”, and “3” in column 2 represent the three different nationalities.