IMPROVING THE INTERPRETATION OF CONFIDENCE AND CREDIBLE INTERVALS

Rink Hoekstra[1], Richard D. Morey[1,2] & Eric-Jan Wagenmakers[3]
[1]University of Groningen
[2] Cardiff University
[3]University of Amsterdam
r.hoekstra@rug.nl

Confidence intervals (CIs) are often endorsed as a useful alternative for the frequently criticized significance test. It is shown, however, that neither students nor researchers find them easy to interpret (Hoekstra et al., 2014). This may be understandable, given how complicated the interpretation of CIs can be (e.g., Morey et al., 2016), but it seems indicative of a statistical education that is suboptimal. This is underscored by an analysis of introductory statistical textbooks, which shows a striking variability of interpretations of CIs, and an alarming frequency of incorrect interpretations. Apparently, statistical education is not optimally effective. Subsequently, we discuss some constructive suggestions to improve our education, with the goal of improving students' understanding, despite the haziness in many current textbooks.

INTRODUCTION

Given the abundance of misuse or misunderstanding of NHST (e.g., Gigerenzer 2004; Haller & Kraus, 2002; Hoekstra, Finch, Kiers & Johnson, 2006; Oakes, 1986;) and given the amount of criticism of NHST (for an overview, see e.g., Kline, 2013), we seem in dire need of a useful alternative, although some defend its usefulness (e.g., Chow, 1998; Cortina & Dunlap, 1997) Many solutions have been proposed, including Bayesian analyses (e.g., Kruschke, 2013; Rouder & Morey, 2009; Wagenmakers, 2007), CIs (e.g., Cumming, 2013; Cumming & Finch, 2001; Fidler & Loftus, 2009), and changing the way we use significance testing (for example by lowering the significance level to 0.005 instead of the common 0.05, Benjamin et al., 2018). Others (Trafimow & Marks, 2015) have advocated replacing NHST by descriptive statistics only. In this paper, we will not claim that there is a single solution, but we think it cannot be disputed that if we are going to adopt one of the many proposals for moving away from NHST, it better be one that students can understand and use without running into problems yet again. Moreover, we think that the more pragmatic approach in which philosophically unsound interpretations of CIs are permitted and even endorsed is unhelpful, and should be replaced by a more principled one. If students are to learn a certain statistical technique, expecting from statistics teachers to guard them against quick-and-dirty versions seems very reasonable indeed.

Of the many suggestions for alternatives, replacing NHST by CIs or adding CIs to NHST is arguably the most frequently mentioned one. In contrast to most of the other approaches, they are also discussed in basically every contemporary introductory statistics textbook. Cohen (1994) advocated for their use because ""[e]verybody knows" that confidence intervals contain all the information to be found in significance tests and much more" (p. 1002), although he admitted that "…a magical alternative to NHST… doesn't exist" (p. 1001). As an editor of Memory and Cognition, Geoff Loftus tried to change the regulations of the journal to make CIs an integral part of every inferential analysis, which limited long-term success (Finch et al., 2004). In 1999, Leland Wilkinson, who led a taskforce that was installed to "elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives" (p. 594), presented guidelines which prominently endorsed CIs as a more useful alternative to NHST, and the suggestions of the taskforce were at least partly incorporated in later editions of the APA Manual (2004, 2009). More recently, Geoff Cumming published his book Understanding the New Statistics (2013; see also Cumming, 2014), in which he strongly advocates the use of CIs. Although his proposals are obviously far from new (CIs were developed in the 30s of the previous century), it is true that a widespread use in practice would be a substantial change to standard practice. So if we are to adopt CIs, it is pivotal that they are well understood. But are they?

Recent studies have shown that also CIs are often misinterpreted. Belia et al. (2005) showed that researchers had severe misconceptions about the relationship between CIs and

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

standard error bars. Hoekstra, Morey, Rouder and Wagenmakers (2014) showed that researchers endorsed on average more than three out of six incorrect statements about CIs, and did not clearly outperform students who had had no lectures in statistics. Although Miller and Ulrich (2016) criticized Hoekstra et al.'s approach by claiming that some of the statements could be correct under certain interpretations of the statements, Morey, Hoekstra, Rouder, Lee and Wagenmakers (2016) argued that these more lenient interpretations were not compatible with its underlying philosophy. Garcia-Pérez and Alcalá-Quintana (2016) did a follow-up of the Hoekstra et al. study, and included statements they claimed were correct. Despite this adjustment, they found very similar results, supporting the notion of a widespread misunderstanding of CIs. Recently, Kalinowski, Lai and Cumming (2018) showed that students' intuitions of CIs are basically all over the place.

So what *is* a CI, and how *should* it be interpreted? A CI originates from a confidence procedure (CP). An X% CP is a procedure that results in an infinite amount of CIs of which X% cover the parameter. Thus, the percentage is a property of the CP, but not of an individual CI. This entails that for a given CI, no probability claims can be made. This has consequences for the interpretability of a CI based on actual data. Before data have been collected, it makes perfect sense to claim that since the CI is a randomly drawn interval from the set of all possible CIs as defined by the CP, there is an X% probability that the CI covers the parameter. Once a CI is calculated for actual data, however, such claims can no longer be validly made. Neyman (1937), who came up with the idea underlying CIs, stated the following: "Consider now the case when a sample...is already drawn and the [confidence interval] given...Can we say that in this particular case the probability of the true value of [the parameter] falling between [the limits] is equal to [*X*%]? The answer is obviously in the negative" (p. 349). So what *can* be concluded from a given CI? The only thing that is known is that the CI results from a CP, and that this CP has X% probability of covering the parameter. Whether this CI covers the parameter is unknown, and the probability for this cannot be quantified within the frequentist framework. Another valid interpretation is based on the relation between CIs and significance testing. Given that the margin of error which defines half of the width of a CI equals the distance between the value under the null hypothesis and the critical value (the value that would just result in a significant effect), you could interpret a CI as the set of all non-rejected null hypotheses. Thus, a CI is an inverted significance test. Put differently, a CI is a summary of hypothesis tests for many effect sizes (Greenland et al., 2016).

The Bayesian equivalent of the CI is the credible interval. Like the CI, it is an interval constructed around a sample outcome. Its construction, however, depends not only on the data, but also on a prior belief. Using Bayes theorem, the prior and the information in the data are combined to form a posterior distribution. In principle, any X% portion of this posterior can be considered a X% credible interval, although typically the center X% are taken. In contrast to a CI, a credible interval can be interpreted as including the parameter with a certain percentage, assuming the particular prior that is used.

CIs have the frequentist property that they are to be used to *falsify* certain values, rather than *confirming* certain values. The Bayesian credible interval, on the other hand, is a product of the Bayesian philosophy that is aimed at confirming certain values. Although it can be argued that under certain conditions numerically the two types of intervals do not necessarily differ substantially (Albers, Kiers & van Ravenzwaaij, *under submission*), these are fundamentally different properties, which have consequences for how they can be interpreted.

The correct interpretations renders some common interpretations incorrect. Morey, et al. (2016) distinguished three common misconceptions, The first one, labeled the *Fundamental confidence fallacy*, states that "If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value." (p. 104). They called this the Fundamental fallacy because it may well be the most commonly made mistake. It could be argued that this is the interpretation that a *credible interval*, the Bayesian equivalent of a confidence interval, should have, although a Bayesian credible is conditional on a prior belief about the position of the parameter, which is not the case for a frequentist interval. The Fundamental confidence fallacy is an arguably intuitive, but unfortunately incorrect interpretation. To underscore the assumed commonness of this mistake, the rather strong challenge presented in a blog post Briggs is particularly telling: "If you can find even one

[published analysis] where the confidence interval is not interpreted as a [Bayesian] credible interval, then I will eat your hat". The second fallacy, the Precision fallacy, states that "The width of a confidence interval indicates the precision of our knowledge about the parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence respond to imprecise knowledge" (p.105). This claim seems to be true at first sight, but as we will see later this is not necessarily the case. The Likelihood fallacy states that "A confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside. This fallacy exists in several varieties, sometimes involving plausibility, credibility, or reasonableness of beliefs about the parameter". (p.106). As we have argued earlier, the falsificationist logic underlying frequentist statistics does not justify a claim on the likeliness of values. If we are to teach our students how to understand CIs correctly, we should not only make them aware of the correct interpretation, but also learn them to recognize the incorrect ones. This is not easy given that the scientific environment is apparently full of people who have problems understanding CIs, given the outcomes of the earlier mentioned studies. But what about the quality of the introductory statistics books?

Statistical textbooks are often the first source that students get their ideas about central statistical concepts from. Of course, these are not the only source of information: Even if textbooks would correctly explain what CIs are, misinterpretations could be caused by teachers' oversimplifications, the pressure of research practice to confirm to the way in which CIs are often misunderstood, or because the misinterpretations could be considered more natural than the correct interpretations. On the other hand, if CIs are explained in a confusing way at the start of the educational chain (that is, in textbooks), it is no wonder that students and researchers find them hard to interpret. So, although we don't pretend that good textbooks are a definitive solution to the problem of the magnitude of misinterpretations, incorrect textbooks would definitely be part of the problem. In this study, we will first focus on this role of textbooks. How are CIs presented? Are the misinterpretations as seen in the earlier mentioned studies on misinterpretations of CIs found in textbooks as well? Can different definitions be found within textbooks? After presenting the results of a study in which 23 textbooks were analyzed, we will discuss some constructive suggestions to improve the practice of teaching confidence intervals.

METHOD

Selection books. In order to come to a selection of books, we made a list of books that were used as introductory books at Dutch universities at faculties of behavioral and social sciences. Moreover, we asked publishers to come up with a list of books they thought were most often used in introductory statistics courses. In total, this resulted in 23 books.

Interpretation mistakes. We focused on the three mistakes as presented in Morey et al. (2016): The Fundamental confidence fallacy, the Precision fallacy, and the Likelihood fallacy. In case the book had an overview of definitions, we scored the definition as presented in this overview. Moreover, we scored the first time that CIs were explicitly defined in the book. In order for a statement to be categorized as a Fundamental confidence fallacy, the statement needed to combine a fixed percentage with a claim about a sample outcome. Thus, "we are 95% confident that this interval includes the parameter" would have been categorized as such, but "95% of the thus constructed intervals include the parameter" would not. For the Likelihood fallacy, a claim about a particular interval representing likely or plausible values for the parameter was categorized as this fallacy. For the Precision fallacy, the statement was required to include a claim about the width of the interval representing precision.

Analysis. For all selected statements we scored whether one or more of the fallacies was present. For the presentation of the results, we will use descriptive statistics only. Inferential results are not presented done because our sample should not be considered a random sample, and because we think that the mere occurrence of mistakes in introductory textbooks (let alone in multiple) is informative enough for our discussion.

RESULTS

Twenty-one of the twenty-three books (91%) contained at least one occurrence of one of the fallacies. In total, the Fundamental confidence fallacy was found 14 times (61%). Field (2013),

in Discovering Statistics using IBM SPSS Statistics, writes the following, which we classified as an example of this fallacy: "For a given statistic calculated for a sample of observations (e.g., the mean), the confidence interval is a range of values around that statistic that are believed to contain, with a certain probability (e.g., 95%), the true value of that statistic (i.e., the population value)" (p. 872). The Likelihood fallacy was found in 10 books (43%), and a typical example was found in Dietz & Kalof's Introduction to social statistics (2009): "[A CI is the r]ange of values that is likely to contain the true value of the population parameter, commonly we use a 90 percent, 95 percent or 99 percent confidence interval" (p. 536). The Precision fallacy was not found in our sample. Note that these numbers add up to more than 100 percent, which is due to the fact that in three books multiple different fallacies were made. In four of the books (17%), a correct definition was presented.

CONCLUSION

Our data underscore once more that confidence intervals are not easy to interpret, nor easy to explain. It was already known that researchers find them hard to understand (e.g., Belia et al., 2005; Hoekstra et al., 2014; Garcia-Pérez & Alcalá-Quintana, 2016), that students' ideas are far from stable and seldom correct (Kalinowski, Lai & Cumming, 2018), and now we see that incorrect interpretations abound in introductory textbooks. The fact that mistakes were found in almost all of the textbooks we selected is quite shocking, and the finding that we only found the correct definition in only a small minority of the books does not warrant more optimism. Haller and Krauss (2002) already presented a nice example of the confusing ways CIs can be presented in introductory textbook by showing no less than eight different interpretations in the book *Introduction to statistics for psychology and education* (Nunally, 1975), with all these statements being wrong. Our more systematic approach basically generalized this anecdotal finding.

Since the authors of introductory statistics textbooks are often established statisticians, one could wonder whether they are actually unaware of how a CI should be interpreted. Although we did not directly study the authors´ understanding, and although we know that sometimes statisticians make interpretation mistakes with regards to inference (e.g., Lecoutre, Poitevineau & Lecoutre, 2003), it seems unlikely that these textbook authors were completely unaware of these issues. Possibly, textbook writers, maybe pressured by commercial publishing houses, may be sugaring the pill when presenting complicated concepts. Howell (2011), one of the authors of the text books we studied, explicitly showed the problems he had with the interpretation: `So what it does it mean to say that the 95% confidence interval is $1{,}219 \leq \mu \leq 1{,}707$? For seven editions of each of two books I have worried and fussed about this question" (p. 193). We can only speculate what the reasons for these authors was to define CIs so badly in textbooks, but we think that is not very constructive here. A much more promising approach to tackle this problem is to come up with a didactical sound addition to these textbooks that is directly applicable and helpful for teachers, as we will present later.

Some limitations of our study should be discussed. Our sample was small, and we do not pretend they are a random sample: It only is a sample of books that are regularly used in academia. One should be careful not to overgeneralize these results: Of course there could be other books available in which no fallacies were made when presenting CIs, but it does not seem an exaggerated claim that the fact the we found fallacies in almost every book we checked is indicative of a practice in which fallacies are common.

Another potential limitation is our strictness regarding the correct interpretation of CI. We are aware that there are people who think that we are too strict with what we consider correct interpretations, as was shown by our exchange with Miller and Ulrich (Miller & Ulrich, 2016; Morey et al, 2016a). As we have presented elsewhere (Morey et al, 2016b), we think there are good reasons for our strictness, and we have shown that what we consider fallacies are clearly at odds with Neyman's description of CIs. When we are considered to be too rigid, this should also hold for the one who stood at the basis of CIs.

A last source of criticism could be the arguably limited role of text books in academia. In this rather cynical view on how students prepare for their tests, they hardly open their books but merely listen to how their teachers explain the course material instead. Although this may hold for some students, we think that the books may impact the teaching of the teachers quite a bit, so the

student might be either directly (via the book) or indirectly (via the teacher) affected by how things are written up. There may be teachers who go against the literal interpretation in the book, but according to us, they have limited ways of practicing the correct interpretation of CIs. In the next paragraph, we will present some constructive suggestions to support those teachers.

DISCUSSION

If CIs are the alternative that needs to replace the uncritical application of NHST that has been criticized for decades, it better be well-understood and correctly taught. If teachers cannot rely on many of the textbooks for the interpretation, we think teachers should be presented useful suggestions and tools that can help them with explaining the essential conceptual ideas underlying CIs. Resonating Haller and Krauss (2002), one of the promising ways to go about this seems by contrasting CIs with Bayesian credible intervals, rather than mixing ingredients from both approaches, as we have done consciously or unconsciously for decades. Gigerenzer and Marewski (2015) explicitly criticize textbook authors for this: "textbook writers in the social sciences have transformed rivaling statistical systems into an apparently monolithic method that could be used mechanically", thus suggesting that different philosophical approaches are intentionally mixed. If we want to introduce our students to both frequentist and Bayesian philosophies, intervals seem a nice way to start explaining the similarities and differences between both philosophies. The alternative, -contrasting p-values with Bayesian testing-, seems more difficult on a technical and on a philosophical level, since with *p*-values only one model is evaluated, whereas with Bayesian testing two models are contrasted. Credible intervals and CIs, however, have similar starting positions, and under some conditions they even coincide numerically (e.g., Albers, Kiers & van Ravenzwaaij, *under submission*), although the interpretation of both intervals is quite distinct.

We want to add a couple of other suggestions we consider important:

- When introducing CIs and credible intervals, start with explaining the philosophical underpinnings of the technique at hand. The pragmatic approach, in which philosophical matters were typically ignored, has utterly failed
- When introducing CIs, talk about the CP as well. This can be elegantly done by showing for example a forest plot with multiple results.
- Explain and discuss extensively what inference is, and why we need it in the first place. Subsequently, the more confirmationist Bayesian and the falsificationist logic of the frequentist techniques can be introduced.
- Talking about philosophy, inference and statistical concepts can be very abstract. Make it concrete by presenting that are easy to understand fort students.

REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389-396.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Chow, S. L. (1998). Statistical significance: Rationale, validity, and utility. *Behavioural and Brain Sciences*, *21*, 169-238.

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

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

Cumming G, Finch, S. (2001) A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–574. doi: 10.1177/0013164401614002

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. (2004). Reform of statistical inference in psychology: The case of memory and cognition. *Behavior Research Methods, Instruments, & Computers, 36,* 312–324.

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The Interpretation of Scholars' Interpretations of Confidence Intervals: Criticism, Replication, and Extension of Hoekstra et al.(2014). *Frontiers in psychology*, *7*, 1042.

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33, 587–606.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421-440.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online* [On-line serial], 7, 120. Retrieved February 25, 2018, from www2.uni-jena.de/svw/metheval/lehre/0405-ws/evaluationuebung/haller.pdf.

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p-values. *Psychonomic Bulletin & Review*, *13*, 1033-1037.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Kalinowski, P., Lai, J., & Cumming, G. (2018). A Cross-sectional Analysis of Students' Intuitions when Interpreting CIs. *Frontiers in Psychology*, *9*, 112.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*. American Psychological Association.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573.

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

Miller, J., & Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic bulletin & review*, *23*(1), 124-130.

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E. J. (2016). Continued Misinterpretation of confidence intervals: response to Miller and Ulrich. *Psychonomic Bulletin & Review*, *23*, 131-140.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*, 103-123.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theoryof probability. Philosophical Transactions of the Royal Society of London. Series A. *Mathematical and Physical Sciences*, *236*, 333-380.

Nunally, J.C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.

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

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*(3), 655.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37,* 1–2. doi:10.1080/01973533.2015.1012991

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *1*, 779-804.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.