13th International Congress on Mathematical Education
Hamburg, 24-31 July 2016

RISK AS AN EXPLANATORY FACTOR FOR RESEARCHERS' INFERENTIAL INTERPRETATIONS

Rink Hoekstra

University of Groningen

Logical reasoning is crucial in science, but we know that this is not something that humans are innately good at. It becomes even harder to reason logically about data when there is uncertainty, because there is always a chance of being wrong. Dealing with uncertainty is inevitable in situations in which the evaluation of sample outcomes with respect to some population is required. Inferential statistics is a structured way of reasoning rationally about such data. One could therefore expect that using well-known statistical techniques protects its users against misinterpretations regarding uncertainty, but this does not seem to be the case. Researchers often pretend to be too certain about their effects, and data are analyzed in a selective way, which impacts the validity of the conclusions. Unwanted behavior may, however, not be as unreasonable as it seems, once the risks that researchers face are taken into account.

It is well known that human reasoning is far from optimal, and it becomes even harder when uncertainty is involved. Tversky and Kahneman (1974) showed that also when dealing with uncertainty, we often rely on a relatively small number of heuristics to come up with an answer. Gigerenzer (2008) argues that it is important to take the context in which conclusions are drawn into account, to understand why interpretational mistakes that seem illogical at first may make sense once the context is understood.

From scientific researchers, one would expect their conclusions to be justifiable based on the data at hand. A researcher who tries to answer a certain research question often has a limited amount of data to make valid inferences. When researchers want to know something about a certain group (the population), they are often restricted to information of only a small number of people from this population, which is typically referred to as a sample. This implies that the eventual outcomes are only based on a small subset of the entire group in which the researcher is interested. In these cases, uncertainty is intrinsically connected to any statement that is made about the population. But how can we make a reasonable assessment of the parameter based on the outcome, knowing how hard it is for people to reason with uncertainty? To overcome the human deficits with respect to reasoning about data with uncertainty, a systematic approach of analysing data seems necessary.

Inferential statistics are designed to provide a structure to reason rationally about sample data, in order to make an inference about the parameter. Given that statistics is a means to deal with uncertainty in a structured way, one could expect that when researchers use statistics for their inferential statements, reasoning with uncertainty is done without too many problems. The question, however, is whether this is really the case.

TWO INFERENTIAL STATISTICAL FRAMEWORKS

For reasons of clarity, it is necessary to present the two most commonly used inferential statistical frameworks in the social sciences: Frequentist statistics (which is sometimes also referred to as classical statistics, or frequentism) and Bayesian statistics. The former is by far the most frequently used framework, whereas the latter seems to gain momentum in the last couple of decades (Zyphur & Oswald, 2015).

Frequentism defines probability as a long-term relative frequency. The two frequentist techniques that are most often used are the null hypothesis significance test (hereafter referred to as NHST) and the confidence interval (CI). In the social sciences and beyond, NHST can be found in a large majority of the published papers. NHST, as it is most often used, calculates the probability of finding the outcome found in the sample or more extreme, under the assumption that a specific test-value you are interested in equals a certain value. If this probability, the *p*-value, is smaller than a certain criterion value called the significance level (which is often chosen to equal 5%), the outcome is called "significant", and if it is larger, the outcome is considered "not significant". The significance level indicates the risk of incorrectly rejecting the null hypothesis (also referred to as Type I error). Note that "significant" in this context is only a statistical term: It does not refer to the practical importance of the finding, but it merely indicates how likely the outcome or more extreme is given a certain value of the parameter. In order to say something about practical relevance, power calculations could indicate how "powerful" (strong) the procedure at hand is. This depends on, amongst others, the sample size, and the homogeneity of the group at hand. The power indicates how likely it is to find a significant effect under the assumption that in reality a certain value for the parameter were true. The complement of this probability is called the Type II error.

A second technique within the frequentist framework is the confidence interval. The idea behind CIs is that you construct a certain interval around the value you found in your sample in such a way, that, if an infinite number of independent samples were taken from the same population and a confidence interval was constructed for each sample, a certain percentage (for example 95%) of these intervals would include the parameter. CIs are often endorsed as an alternative for NHST, since they are assumed to give an indication of precision of the parameter estimate, and since they give a direct indication of the size of the effect (Cumming, 2012).

To complicate matters even further, the standard way of using these techniques in practice deviates from theory, and there are at least two theories about which techniques should be used, and how they should be interpreted. Historically, Fisher (e.g., Fisher, 1925) considered drawing conclusions as an essential part of inference, and according to him *p*-values are to be considered as measures of evidence. In his model, no alternative hypotheses are considered. Neyman and Pearson extended his model, but they argued that frequentist inference is to be used for decisions, and not for drawing conclusions. They also stressed the importance of keeping long-term error rates low, thus focusing on the importance of the significance level and power. The debates between these two camps were far from friendly. Ironically, the current model that is typically used seems a hybrid version of Fisher's and Neyman and Pearson's view on frequentist statistics (Gigerenzer, 1993), and it would probably have been rejected by both sides, although on different grounds (Kline, 2013). Because of its prevalence, I will focus on the hybrid version. The reader should keep in mind, however, that depending on the position one has, the interpretations that are considered correct and not correct may differ.

Probability is defined differently within the Bayesian framework. The core idea of Bayesian statistics is that it quantifies how a rational person would change his or her prior beliefs about a parameter based on the data at hand. That is, contrary to frequentist statistics, Bayesian statistics require a specification of one's belief about plausible values of the parameter before looking at the data. Within Bayes, there are several ways of analysing data. First of all, you can update the prior beliefs by means of the data. That is, given a quantification of your prior idea of the position of the parameter, Bayesian statistics tell you how you should update your belief based on the data if you were a completely rational person. An alternative way is to contrast two models of the data, and compare the likelihood of both models given the data. This ratio of likelihoods is called the Bayes Factor (Dienes, 2011).

Now that the two main inferential frameworks (frequentism and Bayesianism) have been introduced, the question is whether they are usable for researchers to draw sound conclusions. How do researchers in practice deal with uncertainty when making inferential claims? What do we know about their use of statistical methods? Do the statistical techniques they use protect them from making reasoning errors? In the sequel, it will be discussed how people use inferential statistics to reason about data. In a second step, it will be explored how the definition of "risk" as it was introduced earlier, may play an important role in explaining why people use statistics as they do.

THE USE OF INFERENTIAL TECHNIQUES IN PRACTICE

Analysing and interpreting data by means of inferential techniques involves many aspects. For the purpose of this paper, we focus on the following two questions:1) How do researchers acknowledge the uncertainty that is inextricably connected to inference in their conclusions? 2) Does the way people analyse the data justify the use of the technique at hand?

For answering the first question, I will focus on our knowledge of how people report and interpret their statistical outcomes. For the second question, I will show that given how many researchers analyse their outcomes selectively, the conclusions based on the techniques are often invalid.

Numerous articles have been written about the usability (or the lack thereof) of frequentist and Bayesian techniques (for an overview, see for example Kline, 2013). The debates between frequentists and Bayesians have been (and are to this very day) typically rather fierce, but also within both frameworks quite some disagreement can be found. Despite the attention for the usability of both techniques, few studies have focused on their use in practice.

As stated before, NHST is by far the most frequently used technique within the social sciences, medical sciences, biology and beyond (Bakker, van Dijk, & Wicherts, 2012; Fidler & Loftus, 2009; Hoekstra, Finch, Kiers, & Johnson, 2006; Kline, 2013). Unfortunately, despite the earlier mentioned "Bayesian revolution", I am not aware of any study on the use in practice of Bayesian techniques. In the sequel, the focus will therefore be on the use of frequentist statistics, which should by no means be interpreted as if frequentism were the only way to analyse data. In the absence of information about the use Bayesian methods, however, I cannot draw any conclusion about its use. In the conclusion section I will return to this by discussing to what extent Bayesian statistics might be a solution for the presented problems.

As to how uncertainty is acknowledged in interpretations (the first question), it is known that the conclusions that are drawn from NHST are often not without errors. Significant results are often interpreted as if it is indisputable that the null hypothesis is false, thus ignoring the risk of making a Type I error, and non-significant results are often presented as if they reflect the null hypothesis being true, thus ignoring the risk of making a Type II error (Hoekstra et al., 2006).

When possible interpretations of *p*-values are presented to researchers, very few are able to correctly indicate which conclusions are justified (Falk & Greenbaum, 1995; Oakes, 1986). Haller and Kraus (2002) showed that even statisticians endorsed incorrect statements about *p*-values.

Confidence intervals, on the other side are seldom reported (Cumming, 2012; Fidler & Loftus, 2009; Hoekstra et al., 2006), and researchers seem to have great difficulty with interpreting them correctly (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). When the same data are presented by means of NHST or CIs in an experimental setting, it was found that both are interpreted differently (Hoekstra, Johnson, & Kiers, 2012). Since the presented techniques conveyed exactly the same information, -it was only presented differently-, logical reasoning should have led to exactly the same conclusion. Belia, Fidler, Williams, and Cumming (2005) also showed that researchers were relatively bad in showing awareness of the relation between NHST and CIs.

The second question deals with the way data are treated before the analysis that is used in the version that is eventually submitted for publication in a journal. John, Loewenstein, & Prelec, (2012) showed that when analysing data, researchers often selectively adjust their analyses based on the significance of their findings. That is, in case of a significant effect they write down their conclusions without further ado, whereas non-significant findings may lead to adjustments like increasing the sample size, or leaving out one of the conditions or dependent variables, amongst others. They dubbed such adjustments questionable research practices (QRPs). Simmons, Nelson, & Simonsohn (2011) found that using such QRPs selectively inflates the Type I error rate unacceptably. As a result, the interpretation of a *p*-value is practically impossible, and we have shown earlier that researchers already find this difficult. The most extreme QRP is fabricating data, which is obviously wrong. Some of the QRPs, however, are so generally accepted that most researchers may not even be aware that these practices are questionable. All in all, there are strong indications that QRPs are part of the routine of researchers. This has serious implications for the interpretation of the results, since the outcomes do no longer mean what they originally mean. In summary, frequentist outcomes are often interpreted incorrectly, and in papers effects or the alleged lack of an effect are often presented with too much certainty. Moreover, the apparent frequent use of QRPs impacts the interpretability of the outcomes even further. Apparently, despite the fact that inferential statistics have been designed to prevent interpretational errors from occurring, they are no guarantee for the absence of such mistakes in the published literature.

RISK AS AN EXPLANATORY FACTOR

In this paragraph, I will try to show how the previously mentioned use and interpretation of inferential outcomes can be partially explained by risk aversion behaviour. In order to assess researchers' considerations with respect to the aforementioned choices, I will use Kaplan and Garrick's (1981) risk analysis by answering the set of three questions: *What can happen?*, *How likely is it that this will happen?*, and *If it does happen, what are the consequences?*.

The idea that risk aversion may explain why the unwanted behaviours (intentionally or not) like misinterpreting statistical outcomes and some QRPs is not a completely new insight. Many have argued that the incentive structure in academia is at the core of the problem that we are having in psychology and beyond (e.g., Nosek, Spies, & Motyl, 2012; Nosek & Lakens, 2014; Open

Science Collaboration, 2012). Avoiding risks can be considered complementary to following incentives. With respect to risks, two things are of crucial importance for being a successful researcher: Having a decent amount of publications, and having a solid scientific reputation.

Let's first focus on the risks involved with one of the types of unwanted behaviour discussed in this paper: QRPs. In the current scientific publication system (and assuming the use of frequentist statistics), a significant finding seems a requisite for a publishable paper. When effects happen to be non-significant, QRPs may be tempting. When *not* using them in such cases, the risk of *not* publishing the article is believed to be high. Nevertheless, non-significant findings could be indicative of interesting findings. If papers with non-significant effects are not published, however, we will end up with a so-called *file-drawer problem* (non-significant findings disappear in a file-drawer), which makes it hard to get a good estimate of the parameter. QRPs increase the probability of an article getting published, since the probability of finding a significant effect is artificially increased. Thus, avoiding QRPs might lead to a paper not being published, and it consequently might hamper a researcher's career and reputation. The probability of that reputation being hampered in case a researcher decides to apply QRPs, on the other hand, seems negligible, because the probability of getting caught when doing so is practically zero: Typically, a researcher is not observed when analysing the data. And even if someone happens to get caught, the reputational damage is limited: The argument "Everybody else is doing it, so why can't I?" can easily be used as a defence for most QRPs.

A second choice a researcher has to make considers acknowledging uncertainty. Within the context of NHST, ignoring uncertainty results in interpreting a significant effect as definite proof that the null-hypothesis is untrue (thus ignoring the possibility of a Type I error), and a non-significant effect as if there is definite proof that the null-hypothesis is true (thus ignoring the possibility of a Type II error). A pragmatic researcher would base this choice at least partly on the impact it might have on the publishability of the paper and on his or her reputation. If uncertainty is acknowledged, the paper is somewhat more honest, but also harder to read, and arguably less convincing. The culture in academia seems to require novel and ground-breaking research, and questioning oneself could be seen as a sign of weakness. In that sense, acknowledging uncertainty can decrease the probability of a paper being published, and if it is published, it can be seen as if the results are not very strong, which can impact the reputation of the researcher negatively. Presenting the results in a way that implies more certainty than is justified, on the other hand, does not seem risky at all. The likeliness of being accused of being too certain is small, since it is common practice among behavioural scientists.

In summary, we saw earlier that mistreating data (QRPs) and a misinterpretation of outcomes (ignoring uncertainty) is observed regularly. Although unwanted, this behaviour can be considered rational once the scientific context is taken into account. The risks of not publishing an article or of a damaged reputation seem higher when QRPs are *not* applied, and, arguably, the risks are also higher when uncertainty is acknowledged.

CONCLUSION AND DISCUSSION

Researchers have difficulties to reason soundly about their inferential outcomes. Although we know that people are generally bad in reasoning with uncertainty (see e.g., Gigerenzer, 2008; Tversky & Kahneman, 1974), one would hope that inferential techniques, which are designed to assist people in reasoning in a structured way, would prevent these problems, but clearly that is not the case. Not only do researchers often seem to ignore the amount of uncertainty that is inextricably connected to every conclusion that is drawn about a population based on a sample outcome, but QRPs make it almost impossible to draw valid conclusions from the frequentist outcomes that are typically used. QRPs and pretending certainty are two different issues, but they are connected: With QRPs the probability of finding a Type I error are artificially inflated, and when uncertainty is not acknowledged Type I and Type II errors are basically ignored.

The findings that QRPs are regularly used and that uncertainty is often not acknowledged are non-trivial outcomes: In some papers the current status of the use of inferential methods in psychology is referred to as a “crisis” (e.g., Pashler & Wagenmakers, 2012), and replication studies show that many significant outcomes are not significant in exact replications of the studies (e.g., Galak, LeBoeuf, Nelson, & Simmons, 2012), which could be indicative of a regular use of QRPs in the original articles. Unfortunately, knowing the existence of QRPs also impacts the trustworthiness of those papers for which they were not used, since it is almost impossible for reader to distinguish between the two.

In this paper the focus is on the role of risk in the choices that individual researchers seem to make. From this perspective, the unwanted behaviour may not be as irrational as it seems. If one wants to minimize both the risk of a hampered reputation and the risk of not publishing a sufficient amount of papers, presenting outcomes as clear as possible, -even if this impacts the formally correct interpretation somewhat-, and showing some QRPs seem in fact rational choices.

If we, as a scientific community, consider such behaviour unacceptable, a viable solution should lead to an increased risk of not publishing the paper at hand, or reputational damage in case such behaviour is identified. As long as these risks remain low, the chances for a rigorous change might be slim. Below, a few possible solutions will be discussed with respect to the implications for risk. First, a relatively easy solution (e.g., Nosek et al., 2012; Nosek & Lakens, 2014) is to install a system of preregistration. In a version of such a system, studies are submitted before they are conducted, and only the theoretical part and a description of the method are reviewed, and the decision whether to publish the article is conditional on the quality of these parts only. Thus, accepting or rejecting a paper does no longer depend on the results. If such a system would be implemented, there is no longer a direct need for QRPs or presenting the outcomes with certainty. It could be argued, however, that even when preregistration would be implemented, the risks of using QRPs or presenting data with certainty for hampering the publication or the researchers’ reputation would remain small. It is not inconceivable that researchers would keep doing displaying those, because this was common practice for a long time, and habits are known to be hard to change. Therefore, preregistration alone may not be sufficient to unlearn such behaviour. A second solution is to require authors to make their data publically available (Wicherts, Bakker, & Molenaar, 2011). If *all* data are shared after publishing an article, some of the QRPs are detectable, and thus the likelihood of reputational damage in case of the use of QRPs may increase. I expect that making all the data available, maybe in combination with preregistering the research questions would increase the risks of using QRPs substantially. Preregistration and making data available would, however, not necessarily prevent a too rigid interpretation of NHST.

A third solution is to require researchers to use Bayesian instead of frequentist techniques. Philosophical debates about the usability of Bayesian techniques and frequentist techniques have been going on for decades, in which both sides claim that the framework they adhere to has clear advantages over the other (see e.g., Kline, 2013; Wagenmakers, 2007). Whether one is better than the other with respect to dealing with uncertainty is beyond the scope of this paper. I do want to stress, however, that the rigid frequentist cut-off values for NHST (and it could be argued that the clear limits of CIs can be used equally rigid (Abelson, 1997)) are not intrinsically connected to Bayesian techniques. In that sense, applying QRPs in order to reach a certain outcome does not make sense since no values need to be reached in order to have a publishable paper. With respect to acknowledging uncertainty, Bayesian techniques explicitly deal with what a rational person should believe, assuming a certain starting position. Thus, pretending more certainty than is justified is taking an enormous risk, because every reader who understands the techniques at hand can see that the claim is unwarranted.

If we want to solve the problems that are discussed, systemic changes are inevitable. Nevertheless, individual researchers do have a responsibility as well, and given the amount of attention for this issue it gets more and more dubious to maintain a waiting position. As Gigerenzer (2008) puts it: “It takes some measure of courage to cease playing along in this embarrassing game. This may cause friction with editors and colleagues, but in the end it will help them enter the dawn of statistical thinking” (p. 171). When researchers would take more risks, and the scientific community would take the risks researchers face into account, inferential techniques may become what they were originally designed for: Tools to reason about uncertainty, instead of tools to increase the chances of publishing suboptimal papers.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

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

Buchanan, M. (2007). Statistics: Conviction by numbers. *Nature*, *445*, 254-255.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY, Routledge.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift Für Psychologie/Journal of Psychology*, *217*, 27-37.

Fisher, R. A. (1925). *Statistical methods for research workers.* (11th ed. rev.). Edinburgh: Oliver and Boyd.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*, 933.

Gigerenzer, G., 1993. The superego, the ego, and the id in statistical reasoning. In Keren, G., Lewis, C. (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care, 10*, 197-211.

Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*, 1-20.

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., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72*, 1039-1052.

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532.

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis*, *1*, 11-27.

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

Washington, DC: APA Books.

Manktelow, K. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making.* Hove: Psychology Press.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615-631.

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

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657-660.

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence? *Perspectives on Psychological Science*, *7*, 528-530.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, *6*, e26828.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference A User's guide. *Journal of Management*, *41*, 390-320. doi: 10.1177/0149206313501200.