Statistical Methods for Small Data

Dr Rens van de Schoot

Scientia

STATISTICAL METHODS FOR SMALL DATA

Researchers are heavily reliant on statistical techniques that are based on large sample sizes. Therefore, attempts to gain useful information from small samples can often lead to biased, or incorrect conclusions. **Dr Rens van de Schoot** at Utrecht University has shown that the limitations associated with small samples sizes can be overcome by using an alternative method – Bayesian estimation – as an all-encompassing approach to quantitative research. However, this approach comes at a price: expert knowledge must be integrated into the statistical model.

Beyond Coin Toss Statistics

In statistics, sample sizes have traditionally played a large role in the knowledge we can derive from a given field. By testing a large number of different instances of a phenomenon, we increase the likelihood that the results achieved are an accurate reflection of reality. Analogously, by flipping a coin endlessly, we can continually refine the data we have about the likelihood of an outcome.

This 'tossing a coin' approach to statistics is called the frequentist paradigm, and one that Dr Rens van de Schoot of Utrecht University has been working to overhaul. With his recent research on Bayesian methods to deal with small sample sizes, he has been working to challenge the reliance of statistics on mere numbers.

'I mostly work on small sample size problems,' Dr van de Schoot explains. 'For example, there are few children with burn wounds, that it's hard to do any representative statistical analysis on them. For those kinds of topics, I look for the appropriate statistical methods for such cases. Then I find, for example, the clinicians treating such victims, and who would benefit from such statistical methods.'

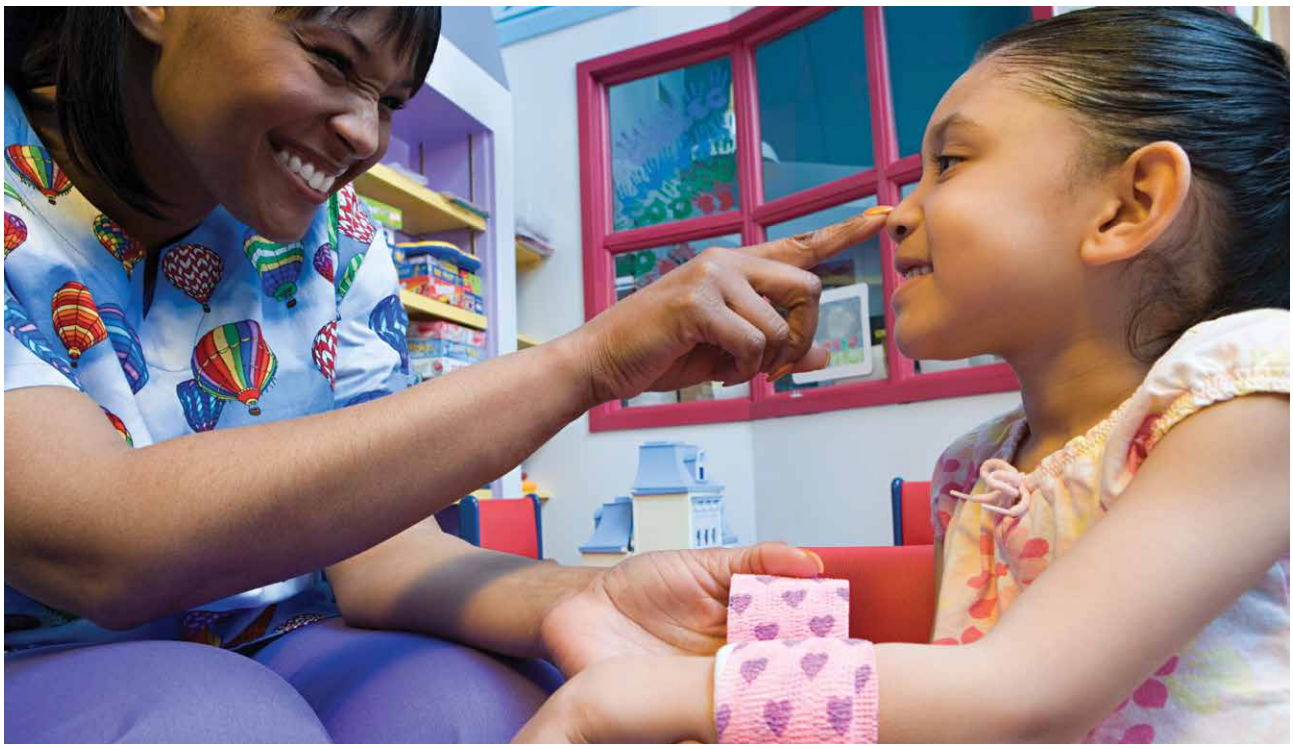
Because researchers rely so heavily on techniques based on large sample sizes, limitations associated with the samples available can restrict the usefulness of the data obtained. Dr van de Schoot explains that researchers may suffer from small sample sizes because large ones are scarce, subject to bureaucratic barriers, or costly, as with studying children with severe burn wounds, for example.

Researchers can circumvent small sample sizes through simplifying their hypotheses or statistical models, but the limitation inherently tends to produce biased results in analyses. The skewed results produced are therefore less in line with the greater context of the field, and don't take full advantage of human knowledge as a whole.

Maximum likelihood estimation is a method that seeks to find the most plausible parameters of a statistical model. In a 2015 study, by Dr van de Schoot and his colleagues demonstrated that this estimation technique results in biased findings if sample sizes are small. This is because it is based on the central limit theorem – it only provides reliable results with many data, just like in the coin toss example.

Dr van de Schoot and his team also showed that the reliance on a large amount of data can be overcome by using an alternative method – Bayesian estimation – as a more all-encompassing approach to quantitative research. But it comes with a price: expert knowledge must be integrated into the statistical model.

‘I am a project initiator: I jump on innovative ideas on knowledge production and build bridges between the right types of knowledge and people to make these ideas a reality. Because of my optimism and energy, I enjoy motiving colleagues to strive for a common goal.’



A New Paradigm in Statistics

Dr van de Schoot has drawn attention to the fact that there is increasing recent interest in Bayesian statistics and analysis in, for example, psychology, educational research and post-traumatic stress.

Developed in the 18th century, Bayesian statistical methods have recently become more common in social and behavioural science research. The Bayesian paradigm offers a different view of interpreting probability as a subjective result of uncertainty. This statistical approach involves the incorporation of background knowledge from previous studies, using as much previous information as possible in combination with personal judgment by experts. The combination of previous knowledge plus new findings provides an updated view of ‘plausibility’, producing a view of which strategies in the given field could move.

There are three ingredients of Bayesian statistics. The first is the background knowledge, which reflects levels of pre-existing knowledge including the uncertainty attached to it. This background knowledge is then translated into a statistical distribution, called the prior distribution. The second component is the data from the research itself – this is expressed as ‘likelihood’. The third component is posterior inference, which combines the first two aspects under the Bayesian theorem. It is a compromise, as it reflects updated knowledge but is balanced by observed data. For this reason, Dr van de Schoot argues that Bayesian statistical methods are unique in their capacity to produce a cumulative form of knowledge.

The background knowledge used in Bayesian analysis often comes from systematic reviews, meta-analyses, previous studies on similar datasets, or even experts. Dr van de Schoot argues that having background knowledge as part of the statistical model is particularly important in psychology, because prior research has established that replication is a crucial tool in the field.

Bayesian statistics allow past work to be built on in a meaningful way. The technique helps to better contextualise research data, creating results with a more all-encompassing view. Because background knowledge also contains information, it can be viewed as, loosely formulated, additional data, which is used in the estimation of the statistical model. As a result, large sample sizes are not necessarily needed in Bayesian statistics, putting less restrictions on the size of the datasets. So, Bayesian estimation can use smaller samples while demonstrating greater power of prediction due to its incorporation of prior distributions.

However, Dr van de Schoot warns that ‘researchers should not use such methods because they are lazy and don’t want to collect more data, but only if collecting data is simply not

possible for whatever reason.' This is the case for the data available on post-traumatic stress, which is one focus of Dr van de Schoot's recent research.

Dr van de Schoot's aim in his recent work into the development of post-traumatic stress symptoms after a traumatic experience is to investigate how much background information is needed to overcome small data issues, find where to get it, how to elicit it, and how to report the whole process.

Anticipating the Effects of Trauma

Previous work highlights that after traumatic events, approximately 10% of individuals develop post-traumatic stress symptoms. Some symptoms may persist for years and can have profound negative effects on a sufferer's life, including medical bills, work absence and substance abuse. When post-traumatic stress is overlooked in the early stages, it can have particularly harmful long-term effects. It is therefore of significant social importance to understand the trajectories that post-traumatic stress disorder can take.

Dr van de Schoot refers to research from 2004 by Dr George Bonanno of Columbia University, in which two relatively stable patterns of post-traumatic stress development and two more dynamic patterns were identified. *Resilient and chronic* trajectories were more simply accounted for in this study, whereas a decreasing recovery and an increasing 'delayed onset' trajectory were more variable.

Most studies that have been conducted since then support these four distinctive trajectories. However, Dr van de Schoot and colleagues point out that the 34 papers produced since the cornerstone 2004 study all employed a technique that is not statistically equipped to detect smaller trajectories such as delayed onset. They state that this does not produce a sufficiently comprehensive picture of the numbers suffering from certain conditions.

The chronic trajectory and delayed onset trajectory only exist to a small extent in the data results. Because delayed onset trajectory sufferers generally present symptoms only after the limited period in which resources such as hospital treatment and legal support are available, their condition may not be identified or treated. Additionally, these symptom patterns often reach a high level of severity. This is why, Dr van de Schoot and his colleagues argue, it is necessary to develop models using statistical methods that are sensitive to smaller quantities of data (i.e. Bayesian statistics with background information incorporated into the analyses). Only these would have the sophistication to predict the development of such symptoms. 'The awareness that small but clinically relevant trajectories may appear, even beyond the acute phase, may help clinicians to develop efficient follow-up programs and to provide these individuals with help when indicated,' says Dr van de Schoot.

Statistics for the Future

The 2004 study in which four distinct trajectories were found for post-traumatic stress symptoms after a traumatic experience is a site that Dr van de Schoot feels is necessary to re-examine under the lens of Bayesian statistics. He aims to re-analyse the datasets created by the 34 studies that have been published since. By creating new and more macrocosmic data, he hopes to institutionalise a more broadly encompassing model with which further studies can be undertaken.

Dr van de Schoot sees the future of this statistical technique as only being enhanced as the Information Age progresses. He notes that Bayesian computational methods are increasingly available in free and proprietary software, and argues that this means researchers using statistics should not have reservations about taking advantage of this new paradigm. 'I want to create a workflow between field experts and statistical experts to help each other get more understanding,' Dr van de Schoot states. 'What gives me most joy is that society as a whole eventually benefits from this exchange of knowledge.'

In time, Dr van de Schoot hopes that experts, data scientists, professionals and individual researchers will no longer need to be rendered separate from each other through isolated research practices, and that information can be synergised in a meaningful way through which ordinary members of society can directly benefit. With these new techniques at researchers' disposal, he believes in the power of Bayesian analysis to provide them with more powerful tools for interdisciplinary action that benefits previously discrete fields mutually. This could lead to conclusions and understandings that existing methods have not readily accommodated.

Meet the researcher

Dr Rens van de Schoot
Utrecht University
Utrecht
The Netherlands

Dr Rens van de Schoot obtained his PhD from the Methodology and Statistics Department at Utrecht University in 2010, graduating cum laude and winning the APA award for best dissertation of the division of Qualitative and Quantitative Methods. Directly after achieving his PhD, he was appointed as Assistant Professor at Utrecht University, where he is now an Associate Professor. Here, Dr van de Schoot primarily conducts research into Statistics for Small Data Sets, Bayesian statistics, responsible research practices and posttraumatic stress disorder. He is a member of the Young Academy of The Royal Netherlands Academy of Arts and Sciences (KNAW), a member of the program board of the think-tank called the FD young circle, and has recently become a member of The Society of Multivariate Experimental Psychology (SMEP). Dr van de Schoot also works as a statistical research consultant for the many organisations such as the Association of Dutch Burns Centres (VSBN), and the Netherlands Centre for Graduate and Research Schools in The Netherlands.

CONTACT

E: A.G.J.vandeSchoot@uu.nl
W: https://www.rensvandeschoot.com

KEY COLLABORATORS ON THIS PROJECT

Nancy E. van Loey, Utrecht University & Association of Dutch Burns Centres
Sarah Depaoli, University of California, Merced
Marit Sijbrandij, VU University
Sonja D. Winter, University of California, Merced
Miranda Olff, University of Amsterdam & Arq Psychotrauma Expert Group

FUNDING

This work was supported by Grant NWO-VIDI-452-14-006 from the Netherlands Organisation for Scientific Research.

FURTHER READING

R van de Schoot, M Sijbrandij, S Depaoli, SD Winter, M Olff, NE van Loey, Bayesian PTSS-Trajectory Analysis with Informed Priors Based on a Systematic Literature Search and Expert Elicitation, Multivariate Behavioural Research, 2018, 1–25.

R van de Schoot, D Kaplan, J Denissen, JB Asendorpf, FJ Neyer, F Schiller, MAG van Aken, A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research, In Child Development, 2014, 85, 842–860.

R van de Schoot, JJ Broere, KH Perryck, Ml Zondervan-Zwijnenburg, NE van Loey, Analysing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors, In European Journal of Psychotraumatology, 2015, 6, 25216.

Universiteit Utrecht