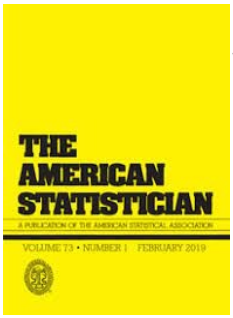

The American Statistician, ISSN 0003-1305
Volume 77, number 2 (may 2023)

Abstract
Reproducibility, the ability to reproduce the results of published papers or studies using their computer code and data, is a cornerstone of reliable scientific methodology. Studies where results cannot be reproduced by the scientific community should be treated with caution. Over the past decade, the importance of reproducible research has been frequently stressed in a wide range of scientific journals such as *Nature* and *Science* and international magazines such as *The Economist*. However, multiple studies have demonstrated that scientific results are often not reproducible across research areas such as psychology and medicine. Statistics, the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data, prides itself on its openness when it comes to sharing both computer code and data. In this article, we examine reproducibility in the field of statistics by attempting to reproduce the results in 93 published papers in prominent journals using functional magnetic resonance imaging (fMRI) data during the 2010–2021 period. Overall, from both the computer code and the data perspective, among all the 93 examined papers, we could only reproduce the results in 14 (15.1%) papers, that is, the papers provide both executable computer code (or software) with the real fMRI data, and our results matched the results in the paper. Finally, we conclude with some author-specific and journal-specific recommendations to improve the research reproducibility in statistics.

Abstract
We address the question of how to perform a paired *t*-test in situations where we do not know how to pair the data. Specifically, we discuss approaches for bounding the test statistic of the paired *t*-test in a way that allows us to recover the results of this test in some cases. We also discuss the relationship between the paired *t*-test and the independent samples *t*-test and what happens if we use the latter to approximate the former. Our results are informed by both theoretical results and a simulation study.

Abstract
Combining individual *p*-values to perform an overall test is often encountered in statistical applications. The Cauchy combination test (CCT) (*Journal of the American Statistical Association*, 2020, 115, 393–402) is a powerful and computationally efficient approach to integrate individual *p*-values under arbitrary dependence structures for sparse signals. We revisit this test to additionally show that (i) the tail probability of the CCT can be approximated just as well when more relaxed assumptions are imposed on individual *p*-values compared to those of the original test statistics; (ii) such assumptions are satisfied by six popular copula distributions; and (iii) the power of the CCT is no less than

that of the minimum *p*-value test when the number of *p*-values goes to infinity under some regularity conditions. These findings are confirmed by both simulations and applications in two real datasets, thus, further broadening the theory and applications of the CCT.

Bayesian-Frequentist Hybrid Inference in Applications with Small Sample Sizes P. 143-150

Gang Han, Thomas J. Santner, Haiqun Lin & Ao Yuan

Abstract

The Bayesian-frequentist hybrid model and associated inference can combine the advantages of both Bayesian and frequentist methods and avoid their limitations. However, except for few special cases in existing literature, the computation under the hybrid model is generally nontrivial or even unsolvable. This article develops a computation algorithm for hybrid inference under any general loss functions. Three simulation examples demonstrate that hybrid inference can improve upon frequentist inference by incorporating valuable prior information, and also improve Bayesian inference based on non-informative priors where the latter leads to biased estimates for the small sample sizes used in inference. The proposed method is illustrated in applications including a biomechanical engineering design and a surgical treatment of acral lentiginous melanoma.

Optimal and Fast Confidence Intervals for Hypergeometric Successes P. 151-159

Jay Bartroff, Gary Lorden & Lijia Wang

Abstract

We present an efficient method of calculating exact confidence intervals for the hypergeometric parameter representing the number of "successes," or "special items," in the population. The method inverts minimum-width acceptance intervals after shifting them to make their endpoints nondecreasing while preserving their level. The resulting set of confidence intervals achieves minimum possible average size, and even in comparison with confidence sets not required to be intervals it attains the minimum possible cardinality most of the time, and always within 1. The method compares favorably with existing methods not only in the size of the intervals but also in the time required to compute them. The available R package hyperMCI implements the proposed method.

Forbidden Knowledge and Specialized Training: A Versatile Solution for the Two Main Sources of Overfitting in Linear Regression P. 160-168

Chris Rohlfs

Abstract

Overfitting in linear regression is broken down into two main causes. First, the formula for the estimator includes "forbidden knowledge" about training observations' residuals, and it loses this advantage when deployed out-of-sample. Second, the estimator has "specialized training" that makes it particularly capable of explaining movements in the predictors that are idiosyncratic to the training sample. An out-of-sample counterpart is introduced to the popular "leverage" measure of training observations' importance. A new method is proposed to forecast out-of-sample fit at the time of deployment, when the values for the predictors are known but the true outcome variable is not. In Monte Carlo simulations and in an empirical application using MRI brain scans, the proposed estimator performs comparably to Predicted Residual Error Sum of Squares (PRESS) for the average out-of-sample case and unlike PRESS, also performs consistently across different test samples, even those that differ substantially from the training set.

Estimating Knee Movement Patterns of Recreational Runners Across Training Sessions Using Multilevel Functional Regression Models P. 169-181

Marcos Matabuena, Marta Karas, Sherveen Riazati, Nick Caplan & Philip R. Hayes

Abstract

Modern wearable monitors and laboratory equipment allow the recording of high-frequency data that can be used to quantify human movement. However, currently, data analysis approaches in these domains remain limited. This article proposes a new framework to analyze biomechanical patterns in sport training data recorded across multiple training sessions using multilevel functional models. We apply the methods to subsecond-level data of knee location trajectories collected in 19 recreational runners during a medium-intensity continuous run (MICR) and a high-intensity interval training (HIIT) session, with multiple steps recorded in each participant-session. We estimate functional intra-class correlation coefficient to evaluate the reliability of recorded measurements across multiple sessions of the same training type. Furthermore, we obtained a vectorial representation of the three hierarchical levels of the data and visualize them in a low-dimensional space. Finally, we quantified the differences between genders and between two training types using functional multilevel regression models that incorporate covariate information. We provide an overview of the relevant methods and make both data and the R code for all analyses freely available online on GitHub. Thus, this work can serve as a helpful reference for practitioners and guide for a broader audience of researchers interested in modeling repeated functional measures at different resolution levels in the context of biomechanics and sports science applications.

Abstract

Conventional wisdom dispersed by fans and coaches in the stands at almost any high school track meet suggests female athletes typically peak around 10th grade or earlier (15 years of age), particularly for distance runners, and male athletes continuously improve. Given that universities in the United States typically recruit track and field athletes from high school teams, it is important to understand the age of peak performance at the high school level. Athletes are often recruited starting in their sophomore year of high school and individuals develop at different rates during adolescence; however, the individual development factor is usually not taken into account during recruitment. In this study, we curate data on event times for high school track and field athletes from the years 2011 to 2019 to determine the trajectory of fastest times for male and female athletes in the 200m, 400m, 800m, and 1600m races. We show, through visualizations and models, that, for most athletes, the sophomore peak is a myth. Performance is mostly dependent on the individual athlete. That said, the trajectories cluster into four or five types, depending on the race distance. We explain the significance of the types for future recruitment.

Abstract

When releasing record-level data containing sensitive information to the public, the data disseminator is responsible for protecting the privacy of every record in the dataset, simultaneously preserving important features of the data for users' analyses. These goals can be achieved by data synthesis, where confidential data are replaced with synthetic data that are simulated based on statistical models estimated on the confidential data. In this article, we present a data synthesis case study, where synthetic values of price and the number of available days in a sample of the New York Airbnb Open Data are created for privacy protection. One sensitive variable, the number of available days of an Airbnb listing, has a large amount of zero-valued records and also truncated at the two ends. We propose a zero-inflated truncated Poisson regression model for its synthesis. We use a sequential synthesis approach to further synthesize the sensitive price variable. The resulting synthetic data are evaluated for its utility preservation and privacy protection, the latter in the form of disclosure risks. Furthermore, we propose methods to investigate how uncertainties in intruder's knowledge would influence the identification disclosure risks of the synthetic data. In particular, we explore several realistic scenarios of uncertainties in intruder's knowledge of available information and evaluate their impacts on the resulting identification disclosure risks.

Abstract
Hierarchical clustering is one of the standard methods taught for identifying and exploring the underlying structures that may be present within a dataset. Students are shown examples in which the dendrogram, a visual representation of the hierarchical clustering, reveals a clear clustering structure. However, in practice, data analysts today frequently encounter datasets whose large scale undermines the usefulness of the dendrogram as a visualization tool. Densely packed branches obscure structure, and overlapping labels are impossible to read. In this article we present a new workflow for performing hierarchical clustering via the R package called protoshiny that aims to restore hierarchical clustering to its former role of being an effective and versatile visualization tool. Our proposal leverages interactivity combined with the ability to label internal nodes in a dendrogram with a representative data point (called a *prototype*). After presenting the workflow, we provide three case studies to demonstrate its utility.

Abstract
We provide a case study for motivating and teaching nonparametric statistical inference alongside traditional parametric approaches. The case consists of analyses by Bracht et al. who use analysis of variance (ANOVA) to assess the applicability of the human microfibrillar-associated protein 4 (MFAP4) as a biomarker for hepatic fibrosis in hepatitis C patients. We revisit their analyses and consider two nonparametric approaches: Mood's median test and the Kruskal-Wallis test. We demonstrate how this case study enables instructors to discuss critical assumptions of parametric procedures while comparing and contrasting the results of multiple approaches. Interestingly, only one of the three approaches creates groupings that match the treatment recommendations of the European Association for the Study of the Liver (EASL). We provide guidance and resources to aid instructors in directing their students through this case study at various levels, including R code and novel R shiny applications for conducting the analyses in the classroom.

Abstract
We recognize the careful reading of and thought-provoking commentary on our work by Rice and Lumley. Further, we appreciate the opportunity to respond and clarify our position regarding the three presented concerns. We address these points in three sections below and conclude with final remarks in Section 4.

Abstract
Reinforcement learning (RL) is a computational approach to reward-driven learning in sequential decision problems. It implements the discovery of optimal actions by learning from an agent interacting with an environment rather than from supervised data. We contrast and compare RL with traditional sequential design, focusing on simulation-based Bayesian sequential design (BSD). Recently, there has been an increasing interest in RL techniques for healthcare applications. We introduce two related applications as motivating examples. In both applications, the sequential nature of the decisions is restricted to sequential stopping. Rather than a comprehensive survey, the focus of the discussion is on solutions using standard tools for these two relatively simple sequential stopping problems. Both problems are inspired by adaptive clinical trial design. We use examples to explain the terminology and mathematical

background that underlie each framework and map one to the other. The implementations and results illustrate the many similarities between RL and BSD. The results motivate the discussion of the potential strengths and limitations of each approach.