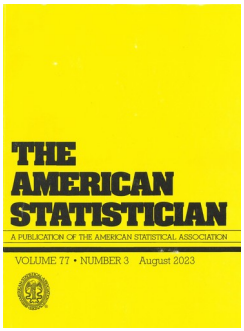

The American Statistician, ISSN 0003-1305
Volume 77, number 3 (august 2023)

Abstract
Scientific journals may counter the misuse, misreporting, and misinterpretation of statistics by providing guidance to authors. We described the nature and prevalence of statistical guidance at 15 journals (top-ranked by Impact Factor) in each of 22 scientific disciplines across five high-level domains ($N = 330$ journals). The frequency of statistical guidance varied across domains (Health & Life Sciences: 122/165 journals, 74%; Multidisciplinary: 9/15 journals, 60%; Social Sciences: 8/30 journals, 27%; Physical Sciences: 21/90 journals, 23%; Formal Sciences: 0/30 journals, 0%). In one discipline (Clinical Medicine), statistical guidance was provided by all examined journals and in two disciplines (Mathematics and Computer Science) no examined journals provided statistical guidance. Of the 160 journals providing statistical guidance, 93 had a dedicated statistics section in their author instructions. The most frequently mentioned topics were confidence intervals (90 journals) and *p*-values (88 journals). For six "hotly debated" topics (statistical significance, *p*-values, Bayesian statistics, effect sizes, confidence intervals, and sample size planning/justification) journals typically offered implicit or explicit endorsement and rarely provided opposition. The heterogeneity of statistical guidance provided by top-ranked journals within and between disciplines highlights a need for further research and debate about the role journals can play in improving statistical practice.

Abstract
Recently, several researchers have claimed that conclusions obtained from a Bayes factor (or the posterior odds) may contradict those obtained from Bayesian posterior estimation. In this article, we wish to point out that no such "contradiction" exists if one is willing to consistently define one's priors and posteriors. The key for congruence is that the (implied) prior model odds used for testing are the same as those used for estimation. Our recommendation is simple: If one reports a Bayes factor comparing two models, then one should also report posterior estimates which appropriately acknowledge the uncertainty with regards to which of the two models is correct.

Abstract
One of the challenges in clustering categorical data is the curse of dimensionality caused by the inherent sparsity of high-dimensional data, the records of which include a large number of attributes. The latent class model (LCM) assumes local independence between the variables in clusters, and is a parsimonious model-based clustering

approach that has been used to circumvent the problem. The mixture of a log-linear model is more flexible but requires more parameters to be estimated. In this research, we recognize that each categorical observation can be conceived as a network with pairwise linked nodes, which are the response levels of the observation attributes. Therefore, the categorical data for clustering is considered a finite mixture of different component layer networks with distinct patterns. We apply a penalized composite likelihood approach to a finite mixture of networks for sparse multivariate categorical data to reduce the number of parameters, implement the EM algorithm to estimate the model parameters, and show that the estimates are consistent and satisfy asymptotic normality. The performance of the proposed approach is shown to be better in comparison with the conventional methods for both synthetic and real datasets.

Quantifying the Inspection Paradox with Random Time P. 274-282

Diana Rauwolf & Udo Kamps

Abstract

The well-known inspection paradox of renewal theory states that, in expectation, the inspection interval is larger than a common renewal interval, in general. For a random inspection time, which includes the deterministic case, and a delayed renewal process, representations of the expected length of an inspection interval and related inequalities in terms of covariances are shown. Datasets of eruption times of Beehive Geyser and Riverside Geyser in Yellowstone National Park, as well as several distributional examples, illustrate the findings.

Selection Criterion of Working Correlation Structure for Spatially Correlated Data P. 283-291

Marcelo dos Santos, Fernanda De Bastiani, Miguel A. Uribe-Opazo & Manuel Galea

Abstract

To obtain regression parameter estimates in generalized estimation equation modeling, whether in longitudinal or spatially correlated data, it is necessary to specify the structure of the working correlation matrix. The regression parameter estimates can be affected by the choice of this matrix. Within spatial statistics, the correlation matrix also influences how spatial variability is modeled. Therefore, this study proposes a new method for selecting a working matrix, based on conditioning the variance-covariance matrix naive. The method performance is evaluated by an extensive simulation study, using the marginal distributions of normal, Poisson, and gamma for spatially correlated data. The correlation structure specification is based on semivariogram models, using the Wendland, Matérn, and spherical model families. The results reveal that regarding the hit rates of the true spatial correlation structure of simulated data, the proposed criterion resulted in better performance than competing criteria: quasi-likelihood under the independence model criterion QIC, correlation information criterion CIC, and the Rotnizky–Jewell criterion RJC. The application of an appropriate spatial correlation structure selection was shown using the first-semester average rainfall data of 2021 in the state of Pernambuco, Brazil.

Bayesian Log-Rank Test P. 292-300

Jiaqi Gu, Yan Zhang & Guosheng Yin

Abstract

Comparison of two survival curves is a fundamental problem in survival analysis. Although abundant frequentist methods have been developed for comparing survival functions, inference procedures from the Bayesian perspective are rather limited. In this article, we extract the quantity of interest from the classic log-rank test and propose its Bayesian counterpart. Monte Carlo methods, including a Gibbs sampler and a sequential importance sampling procedure, are developed to draw posterior samples of survival functions and a decision rule of hypothesis testing is constructed for making inference. Via simulations and real data analysis, the proposed Bayesian log-rank test is shown to be asymptotically equivalent to the classic one when noninformative prior distributions are used, which provides a Bayesian interpretation of the log-rank test. When using the correct prior information from historical data,

Abstract
What is the number of rolls of fair six-sided dice until the first time the total sum of all rolls is a prime? We compute the expectation and the variance of this random variable up to an additive error of less than 10^{-4}. This is a solution to a puzzle suggested by DasGupta in the Bulletin of the Institute of Mathematical Statistics, where the published solution is incomplete. The proof is simple, combining a basic dynamic programming algorithm with a quick Matlab computation and basic facts about the distribution of primes.

Abstract
In many moderate dimensional applications we have multiple response variables that are associated with a common set of predictors. When the main objective is prediction of the response variables, a natural question is: do multivariate regression models that accommodate dependency among the response variables improve prediction compared to their univariate counterparts? Note that in this article, by univariate versus multivariate regression models we refer to regression models with a single versus multiple response variables, respectively. We assume that under both scenarios, there are multiple covariates. Our question is motivated by an application in climate science, which involves the prediction of multiple metrics that measure the activity, intensity, severity etc. of a hurricane season. Average sea surface temperatures (SSTs) during the hurricane season have been used as predictors for each of these metrics, in separate univariate regression models, in the literature. Since the true SSTs are yet to be observed during prediction, typically their forecasts from multiple climate models are used as predictors. Some climate models have a few missing values so we develop Bayesian univariate/multivariate normal regression models, that can handle missing covariates and variable selection uncertainty. Whether Bayesian multivariate normal regression models improve prediction compared to their univariate counterparts is not clear from the existing literature, and in this work we try to fill this gap.

Abstract
An evolving problem in the field of spatial and ecological statistics is that of preferential sampling, where biases may be present due to a relationship between sample data locations and a response of interest. This field of research bears a striking resemblance to the longstanding problem of informative sampling within survey methodology, although with some important distinctions. With the goal of promoting collaborative effort within and between these two problem domains, we make comparisons and contrasts between the two problem statements. Specifically, we review many of the solutions available to address each of these problems, noting the important differences in modeling techniques. Additionally, we construct a series of simulation studies to examine some of the methods available for preferential sampling, as well as a comparison analyzing heavy metal biomonitoring data.

Rameela Raman, Jessica Utts, Andrew I. Cohen & Matthew J. Hayat

Abstract

Statistics education at all levels includes data collected on human subjects. Thus, statistics educators have a responsibility to educate their students about the ethical aspects related to the collection of those data. The changing statistics education landscape has seen instruction moving from being formula-based to being focused on statistical reasoning. The widely implemented Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report has paved the way for instructors to present introductory statistics to students in a way that is both approachable and engaging. However, with technological advancement and the increase in availability of real-world datasets, it is necessary that instruction also integrate the ethical aspects around data sources, such as privacy, how the data were obtained and whether participants consent to the use of their data. In this article, we propose incorporating ethics into established curricula and integrating ethics into undergraduate-level introductory statistics courses based on recommendations in the GAISE Report. We provide a few examples of how to prompt students to constructively think about their ethical responsibilities when working with data.

Consultancy Style Dissertations in Statistics and Data Science: Why and How P. 331-339

Serveh Sharifi Far, Vanda Inácio, Daniel Paulin, Miguel de Carvalho, Nicole H. Augustin, Mike Allerhand & Gail Robertson

In this article, we chronicle the development of the consultancy style dissertations of the MSc program in Statistics with Data Science at the University of Edinburgh. These dissertations are based on real-world data problems, in joint supervision with industrial and academic partners, and aim to get all students in the cohort together to develop consultancy skills and best practices, and also to promote their statistical leadership. Aligning with recently published research on statistical education suggesting the need for a greater focus on statistical consultancy skills, we summarize our experience in organizing and supervising such consultancy style dissertations, describe the logistics of implementing them, and review the students' and supervisors' feedback about these dissertations.