Computational Statistics and Data Analysis 103 (2016) 170–192

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Confidence intervals for an ordinal effect size measure based on partially validated series

Shi-Fang Qiu[a,*], Wai-Yin Poon[b], Man-Lai Tang[c]

[a] Department of Statistics, Chongqing University of Technology, Chongqing 400054, China
[b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China
[c] Department of Mathematics and Statistics, Hang Seng Management College, Hong Kong, China

ARTICLE INFO

Article history:
Received 25 July 2015
Received in revised form 17 April 2016
Accepted 9 May 2016
Available online 19 May 2016

Keywords:
Bayesian credible interval
Bootstrap-resampling
Ordinal data with misclassification
Partially validated series

ABSTRACT

An ordinal effect size measure is used to assess whether one variable is stochastically larger than the other; therefore, this measure is a useful means by which to describe the difference between two ordinal categorical distributions. In practical analysis, it is desirable to obtain data only by a gold standard test, but such tests are often limited due to high costs or ethical considerations, especially in medical research applications. However, misclassification can arise when a cheaper, faster, and/or non-invasive, but fallible test is used to collect data. The use of partially validated data obtained by double sampling has become a popular compromise between these two approaches. In this study, we develop twelve estimators of the confidence interval (CI) for an ordinal effect size measure based on partially validated data. The performance of the proposed CIs are evaluated by simulation studies in terms of the empirical coverage probability, the empirical coverage width, and the ratio of the mesial non-coverage probability and non-coverage probability. Simulation results show that the Wald CI on logit scale, the Bootstrap percentile CI and the Bias-corrected Bootstrap normal CI have outstanding performance even in small sample designs. When sample sizes are moderate, all CIs except the Wald, Bias-corrected Bootstrap percentile and logit-transformation-based Bootstrap percentile-t CIs demonstrate good coverage properties. Moreover, all CIs perform well when sample sizes are large. All methods are illustrated by analyzing a real data set from a research study of highway safety on automobile accidents.

1. Introduction

An ordinal effect size measure is often used to compare two or more groups when ordered categorical variables are involved. Specifically, let Y_1 and Y_2 be two ordinal random variables with r outcome categories. An ordinal effect size measure for the categories is defined as $\theta = P(Y_1 < Y_2) + 0.5P(Y_1 = Y_2)$ (see Klotz, 1966). This measure is also known as a measure of stochastic superiority of Y_2 over Y_1 (see Vargha and Delaney, 1998). It is obvious that θ summarizes the relative size of the two ordinal random variables, that is, the probability that an outcome from one variable falls above the outcome from the other. Klotz (1966) proposed a method which tests the equivalence of the distributions of two groups on the basis of this ordinal effect size measure. Hochberg (1981) considered the confidence interval construction for $P(Y_1 < Y_2) - P(Y_1 > Y_2)$ (which is $2\theta - 1$) via the delta method and the U-statistic, and his proposed methods can be readily used to construct a confidence interval for θ. Halperin et al. (1989) investigated distribution-free confidence intervals for θ on the basis of a

* Corresponding author.
E-mail address: sfqiu@cqut.edu.cn (S.-F. Qiu).

http://dx.doi.org/10.1016/j.csda.2016.05.006
0167-9473/

Table 1
Automobile accident data.

(Male drivers)					
For male drivers	Validated series	Fallible classifier (police report)			
		No injury	Minor injuries	Severe injuries	Total
True classifier (Intensive investigation)	No injury	875	7	16	898
	Minor injuries	59	39	0	98
	Severe injuries	75	0	132	207
	Total	1 009	46	148	1203
		Classified by police report alone			
	Unvalidated series	45 173	1886	7329	54 388
Grand total		46 182	1932	7477	55 591
(Female drivers)					
For female drivers	Validated series	Fallible classifier (police report)			
		No injury	Minor injuries	Severe injuries	Total
True classifier (Intensive investigation)	No injury	359	7	6	372
	Minor injuries	43	30	0	73
	Severe injuries	61	0	87	148
Total		463	37	93	593
		Classified by police report alone			
	Unvalidated series	20 153	1539	4004	25 696
Grand total		20 616	1576	4097	26 289

pivotal quantity. Treating the distributions of Y_1 and Y_2 as continuous, Newcombe (2006a) investigated some issues related to the Mann–Whitney statistic divided by the product of the two sample sizes and developed a tail-area-based confidence interval that can be applied to very small samples or extreme outcomes. Newcombe (2006b) proposed eight asymptotic confidence intervals for θ and concluded that these methods can be applied to ordinal categorical data. Recently, Ryu and Agresti (2008) presented statistical model and inference for the ordinal effect size measure in which they developed and compared several confidence intervals for the measure. Ryu (2009) proposed a simultaneous confidence interval for θ that combines the Studentized range distribution with the score test statistic.

All the aforementioned work assume that the responses have no misclassification errors, that is, a sample of ordinal data is obtained via some "true classification mechanism" that has no error. It has long been accepted that obtaining accurate data solely via some true classification mechanism could be expensive, time-consuming, and/or invasive. On the other hand, obtaining data via a "fallible classification mechanism" that is usually cheaper, faster, and/or non-invasive could induce misclassification errors. To resolve the difficulties involved in the latter situation, Tenenbein (1970, 1972) proposed the well-known double-sampling design which is the utilization of an additional sample. Under the double-sampling scheme, a random sample of $N-n$ subjects is drawn from the target population, and each of the subjects is classified by a fallible classifier. The resultant data series is an "unvalidated" series. Another random sample of n $(n < N)$ subjects is also drawn from the target population, and each of the subjects is classified by a gold-standard test (true or infallible classifier). The resultant data series is a "validated" series. The entire data set obtained via such double-sampling scheme is also known as partially validated series (see for example, Tang et al., 2012). Hochberg (1977) reported a data set regarding automobile accidents. Specially, the original data set consisted of 81,880 automobile accident cases, and each was classified into one of three categories (i.e., no injury, minor injuries and severe injuries) by the North Carolina State Police. This is considered the fallible classifier since the data set from the police reports contained systematic misclassification errors with respect to occupant injuries. Of the 81,880 cases, 1796 cases were again classified by the intensive investigation (i.e., the true classifier). The purpose of our study is to investigate whether a significant difference exists in the injuries between male and female drivers in automobile accidents, and the resultant data set is shown in Table 1. (See also Yiu and Poon, 2008 for details.)

Obviously, the data obtained by infallible classifiers reflect the true state and can hence provide information on the manner in which the misclassification information results can properly be used. On the basis of the partially validated series, statistical inference for ordinal categorical data with misclassification has been investigated. For example, Yiu and Poon (2008) considered the estimation of the polychoric correlation with data obtained by a double-sampling scheme. Poon and Wang (2010a) then extended it to a multivariate model and proposed an expectation–maximization-type (EM-type) algorithm to analyze the model. Poon and Wang (2010b) developed a new class of multivariate probit models that enables the analysis of the information obtained via surrogate response variables and surrogate covariates under the Bayesian framework. However, comparison between two groups by means of an ordinal effect size measure on the basis of the partially validated series has not yet been studied.