Author / title page

Running head: *Rudas-Clogg-Lindsay mixture index of fit*

Bias-corrected estimation of the Rudas-Clogg-Lindsay mixture index of fit

Jenő Reiczigel[1,*], Márton Ispány[2,3], Gábor Tusnády[3], György Michaletzky[4] and Marco Marozzi[5]

[1] University of Veterinary Medicine Budapest, Budapest, Hungary
[2] University of Debrecen, Debrecen, Hungary
[3] Alfréd Rényi Institute of Mathematics, Budapest, Hungary
[4] Eötvös Loránd University, Budapest, Hungary
[5] University of Venice, Venice, Italy

Word count (exc. figures/tables): 5060

*Requests for reprints should be addressed to Jenő Reiczigel, István u. 2, Budapest, H-1078 Hungary, (e-mail: reiczigel.jeno@gmail.com).

Acknowledgement. The authors thank two anonymous referees for their valuable suggestions and constructive criticism. Work of the first author was supported by the Hungarian National Research Fund (OTKA K 108571) and by the Research Faculty Grant 2015 of the Szent István University, Faculty of Veterinary Science.

Bias-corrected estimation of the Rudas-Clogg-Lindsay mixture index of fit

Abstract

Rudas, Clogg and Lindsay (1994) introduced the so-called mixture index of fit, also known as pi-star (π*), for quantifying goodness-of-fit of a model. It is the lowest proportion of „contamination", which, if removed from the population or from the sample, makes the fit of the model perfect. The mixture index of fit has been widely used in psychometric studies. We show that the asymptotic confidence limits proposed by Rudas et al. (1994) as well as the jackknife confidence interval by Dayton (2003) perform poorly, and propose a new bias-corrected point estimate, a bootstrap test and confidence limits for the pi-star. The proposed confidence limits have coverage probability much closer to the nominal level than the other methods do. We illustrate that the proposed method is useful in practice by presenting some practical applications to loglinear models for contingency tables.

Key words: RCL mixture index of fit, pi-star, two-point mixture model index, bootstrap, confidence interval, psychometrics

1. Introduction

Rudas, Clogg, and Lindsay (1994) proposed a measure of goodness-of-fit which is referred to as „RCL mixture index of fit" or „pi-star" (π*). It is the lowest proportion of „contamination", which, if removed from the population or from the sample, makes the fit of the model perfect. They demonstrated the properties of the new measure for the independence model in two-way contingency tables, but later the approach has been further developed, investigated, and used for addressing various other problems like testing the Rasch model, finite mixtures of item response models, minimax and logistic regression (Xi and Lindsay, 1996; Rudas, 1999; Schuster, 2002; Dayton, 2003; Verdes and Rudas, 2003; Formann, 2006; Böckenholt and Heijden, 2007; Revuelta, 2008; Ispány and Verdes, 2014).

Rudas et al. (1994) also constructed an asymptotic lower confidence limit for pi-star. They mentioned that their confidence limit may be unreliable if the true pi-star is near zero. This claim was based on theoretical reasoning, and "near zero" was not specified exactly. Now the considerably higher speed of computers allows assessment of coverage properties of confidence intervals by simulation. Our simulations revealed that the asymptotic confidence limit does not maintain its nominal level even if the true pi-star is large (even for 30% or more), and even for very large samples such as 10000 observations per cell. In practical cases of statistical modelling, pi-star is expected to be below 30% because a model is usually regarded to fit unbearably bad if 30% of the sample must be omitted to reach perfect fit. Thus, the RCL asymptotic confidence limit should be improved.

Dayton (2003) described another method for computing confidence limits for the pi-star measure. This is based on normal approximation using the jackknife estimate of the standard error (SE) of the sample pi-star. Jackknife replicates are artificial samples by removing one sample element from the sample. In this way, from a sample with *n* elements, *n* jackknife samples can be generated. If a statistic of interest, now the pi-star, is computed from each of these jackknife samples, the resulting *n* pi-star values can be used to estimate the SE of the pi-star. For more details about jackknife estimates see Efron and Tibshirani (1993). According to our experiences, in most practical cases the jackknife confidence limits for the population pi-star perform better than the asymptotic ones. However, their actual coverage often remains far below the prescribed nominal level. Medzihorsky (2013) developed an R package based on the algorithms in Rudas et al. (1994), in which the jackknife confidence limits proposed by Dayton (2003) are also implemented.

We propose a test and a confidence interval (CI) based on bootstrap for pi-star. In Section 2 we describe our methodology, in Section 3 we present a simulation study to explore the properties of the new CI and to compare it to those proposed by Rudas et al. (1994) and Dayton (2003). Section 4 describes practical applications of the method to loglinear models for contingency tables. Section 5 concludes the paper with some remarks and directions for future research. An R script for the proposed procedures is available upon request from the first author.

2. Methods

To define pi-star (π^*) in full generality, consider a statistical space (Ω, A, $\boldsymbol{P}$), where $\boldsymbol{P}$ contains all probability distributions (p.d.) of interest. For a p.d. $P \in \boldsymbol{P}$ and a statistical model $\boldsymbol{M} \subseteq \boldsymbol{P}$ define

$$\pi^* = \pi^*(P, \boldsymbol{M}) := inf\ \{\pi : P = (1 - \pi)M + \pi R\},$$

where M is a p.d. from the statistical model $\boldsymbol{M}$, and R is an arbitrary p.d. Thus, pi-star is the smallest proportion π of contamination so that the contamination-free part M of P fits exactly the model $\boldsymbol{M}$. The definition explains why pi-star is also referred to as the “two-point mixture index of fit”.

For count data, the distributions are probabilities of two- or multiway contingency tables, and the statistical model $\boldsymbol{M}$ can be the full independence model, a conditional independence model, or any other model. With regard to statistical inference, if P represents the population distribution, sampling from this P results in a sample p.d. (or empirical p.d.) $\hat{P}$. In this case the sample pi-star $\pi^*_{obs} = \pi^*\left(\hat{P}, \boldsymbol{M}\right)$ is a natural estimate of the population pi-star or true pi-star $\pi^*(P, \boldsymbol{M})$, see e.g. Xi and Lindsay (1996).

Asymptotic results

The following results show that the sample pi-star is a consistent and asymptotically unbiased estimate of the population pi-star for discrete distributions with finite support, including two- and multiway contingency tables. However, for finite samples used in typical applications (10 to 1000 per cell on average) it has a considerable upward bias, which is largest if the true pi-star is near 0 (Rudas et al., 1994; Pan and Dayton, 2011). We also found by simulation that the bias decreases with increasing true pi-star. To illustrate the magnitude of bias, for sample sizes 700 and 7000 in a 7 by 10 table with true pi-star=0, sample pi-star is on average as high as 0.29 and 0.11, respectively. The aim of our method was to correct for bias in this realistic range of sample sizes. When the sample size goes to infinity, both the bias and standard error of the sample pi-star vanish, thus the difference between our estimate and the sample pi-star converges to zero, which means that they are asymptotically equivalent (disregarding the bootstrap error, which can be kept low if the number of replications B is high). The main advantage of our method lies in the better finite-sample behaviour of the point estimate and the confidence limits (see Tables 1 to 3).

For the propositions below, let us consider the pi-star as a functional on discrete distributions with finite support (i.e. with finite number of cells with positive probability).

Proposition 1:

The pi-star is continuous on a space of discrete distributions having the same finite support.

Proof:

Let D and D' be two discrete distributions close to each other in the log-Chebyshev distance, i.e., $d(D, D'):=\max |\log D(\omega) - \log D'(\omega)|=\log \max (\max\{D(\omega), D'(\omega)\}/\min\{D(\omega), D'(\omega)\})$ $\leq \delta$ with some small $\delta > 0$, where the maximum is taken over the common support of D and D'. Note that the log-Chebyshev distance is applied in the log-Chebyshev approximation, see, e.g., Exercise 11.21 in Boyd and Vandenberghe (2004). Clearly $(1-\varepsilon) D \leq D'$ and $(1-\varepsilon) D' \leq D$ where $\varepsilon := 1-\exp\{-\delta\}$. We will show that this implies $\pi(D') \leq \pi(D)+\varepsilon$ and $\pi(D) \leq \pi(D')+\varepsilon$ and hence $| \pi(D) - \pi(D')| \leq \varepsilon$ which proves continuity since $\delta{\downarrow}0$ implies $\epsilon{\downarrow}0$.

From the definition of $\pi(D)$ follows that there exists a distribution M in the model class of interest, for which $(1-\pi(D))M \leq D$. Then $(1-\varepsilon)(1-\pi(D))M \leq D'$, which implies $(1-\varepsilon)(1-\pi(D))$ $\leq (1-\pi(D'))$, that is, $\pi(D') \leq 1-(1-\varepsilon)(1-\pi(D)) \leq \pi(D)+\varepsilon$.

Starting with $\pi(D')$ and using the inequality $(1-\varepsilon)D' \leq D$, the same calculation results in $\pi(D) \leq \pi(D')+\varepsilon$.

Proposition 2:

The sample pi-star $\pi(D_n)$ converges in probability to the population pi-star $\pi(D)$, $\pi(D_n) \overset{p}{\to} \pi(D)$, that is, the sample pi-star is a consistent estimate of the population pi-star.

Proof:

First let us note that if some cells in the population distribution D have zero probability, they will necessarily have zero counts in any sample (these are the so-called structural zeroes), thus these cells have zero probability in the empirical distribution D_n as well as in the model. So these cells do not affect the pi-star at all (neither the population pi-star, nor the sample pi-star).

It is known that the empirical distribution D_n of the sample converges in probability to the theoretical or population distribution, $D_n \overset{p}{\rightarrow} D$. From this follows that in those cells where the population distribution D has positive probability, for large enough samples zeroes occur with arbitrarily small probability even in the empirical distribution D_n. Taking into account that the pi-star is determined by those cells where the probability is positive, it follows from the continuity of pi-star (Proposition 1) that the sample pi-star converges in probability to the population pi-star, $\pi(D_n) \overset{p}{\rightarrow} \pi(D)$.

Proposition 3:

The sample pi-star is an asymptotically unbiased estimate of the population pi-star, and its standard error vanishes when the sample size goes to infinity.

Proof:

This follows from consistency and boundedness of the pi-star, as boundedness by [0,1] implies that all absolute moments of pi-star are finite, and this, together with the convergence in probability, provides convergence in L_r for all $0<r<\infty$ (see for example Loeve, 1977, p. 160).

The proposed bias-correction

By bootstrapping it is possible to estimate bias and standard error of estimates, and based on this, one can correct for bias and construct tests and confidence intervals (Efron and Tibshirani, 1993; Davison, Hinkley and Young, 2003). But in our case, since both the bias and SE of the sample pi-star varies with the true pi-star, it is not sufficient to bootstrap only from the observed sample. Therefore we generate resamples from the observed sample, and also from a distribution with pi-star=0, for which we use the model fitted to data. Although according to our preliminary simulations the bias is not a linear function of the true pi-star, for simplicity we use linear interpolation to estimate bias and SE based on these two points (distribution of the sample and a distribution with pi-star=0), and evaluate the finite-sample performance of this approximate solution by simulation. Even though this linear approximation is rather simple, simulation results show that it works well for typical sample sizes occurring in empirical studies. It works much better than the jackknife, which estimates the bias and SE from the sample alone, thus it cannot take into account that they vary with the true pi-star.

Assume we have a population with true pi-star π_{true} (for better readability we omit the star in the formulas), and a sample drawn from it, that is, a sample generated from that distribution. Let us denote the sample pi-star by π_{obs}. Let us draw B bootstrap samples from the sample, and another B samples from the model fitted to the data, which represents a distribution with pi-star = 0. Let m_b, m_{b0}, s_b, s_{b0} denote the mean and SD of pi-star for these bootstrap samples.

The bootstrap estimate of the bias of sample pi-star is $m_b - \pi_{obs}$ for true pi-star equal to π_{obs}, and m_{b0} for true pi-star equal to 0. From this, the linearly interpolated bias for an arbitrary true pi-star π is

$$bias_\pi = (1 - \pi / \pi_{obs})\, m_{b0} + \pi / \pi_{obs}\, (m_b - \pi_{obs}) \qquad (1)$$

The bias-corrected bootstrap estimate of pi-star can be obtained by solving the equation

$$\pi_{obs} = \pi + bias_\pi$$

which gives

$$\pi_{est} = \pi_{obs}\, (\pi_{obs} - m_{b0}) \,/\, (m_b - m_{b0}). \qquad (2)$$

By the same reasoning we obtain the following interpolated standard error of sample pi-star for an arbitrary true pi-star π

$$s_\pi = (1- \pi / \pi_{obs})\, s_{b0} + (\pi / \pi_{obs})\, s_b \tag{3}$$

If testing for H_0: $\pi_{true} = \pi$, normal approximation leads to the following acceptance region

$$\pi + bias_\pi - t_{crit}\, s_\pi \qquad \text{to} \qquad \pi + bias_\pi + t_{crit}\, s_\pi \tag{4}$$

where t_{crit} denotes the critical value of the Student-t distribution on $B-1$ degrees of freedom. For a two-sided test t_{crit} is the $(1-\alpha/2)$ quantile, while for a one-sided test it is the $(1-\alpha)$ quantile of the distribution, where α denotes the prescribed Type I error rate.

Combining the above results, we obtain for the lower endpoint of the acceptance region the following expression

$$L_\pi = \pi + (1- \pi / \pi_{obs})\, m_{b0} + \pi / \pi_{obs}\, (m_b - \pi_{obs}) - t_{crit}\, \{(1- \pi / \pi_{obs})\, s_{b0} + (\pi / \pi_{obs})\, s_b\} \tag{5}$$

and similarly for the upper endpoint

$$U_\pi = \pi + (1- \pi / \pi_{obs})\, m_{b0} + \pi / \pi_{obs}\, (m_b - \pi_{obs}) + t_{crit}\, \{(1- \pi / \pi_{obs})\, s_{b0} + (\pi / \pi_{obs})\, s_b\}. \tag{6}$$

If π_{obs} is within the acceptance region, that is, if $L_\pi \le \pi_{obs} \le U_\pi$, then we accept H_0, else reject it. A confidence interval can be obtained by inverting this test, that is, by finding the minimal and maximal π for which the inequality $L_\pi \le \pi_{obs} \le U_\pi$ holds. This can be obtained by solving the equations $L_\pi = \pi_{obs}$ and $U_\pi = \pi_{obs}$ for π. This results in the following confidence limits

$$LCL = \pi_{obs}\, (\pi_{obs} - m_{b0} - t_{crit}\, s_{b0}) \,/\, \{m_b - m_{b0} + t_{crit}\, (s_b - s_{b0})\} \tag{7}$$

and

$$UCL = \pi_{obs}\,(\pi_{obs} - m_{b0} + t_{crit}\;s_{b0}) \,/\, \{m_b - m_{b0} - t_{crit}\,(s_b - s_{b0})\}\ . \tag{8}$$

Figure 1 illustrates the proposed procedure. It demonstrates that dependence of both the bias and the standard error of the sample pi-star is approximated linearly by the procedure. Since the approximation is based on two points, on the sample and the fitted model (a distribution with pi-star equal to zero), it is expected to describe this dependence better than a method based only on the sample (like the jackknife, or bootstrapping only from the sample).

(Figure 1 here)

If working with a small number of bootstrap replications to reduce computer time, the estimate of SE may be imprecise. Therefore, to keep the CI conservative, we can substitute the smaller one of s_{b0} and s_b with the larger one. This leads to the following confidence limits

$$LCL = \pi_{obs}\ \{\pi_{obs} - m_{b0} - t_{crit}\max(s_{b0}, s_b)\} \,/\, (m_b - m_{b0}) \tag{9}$$

and

$$UCL = \pi_{obs}\ \{\pi_{obs} - m_{b0} + t_{crit}\max(s_{b0}, s_b)\} \,/\, (m_b - m_{b0}\ )\ . \tag{10}$$

If the procedure results in confidence limits less than 0 or greater than 1, the value is truncated to 0 or 1.

3. Simulation study

The proposed test and CI are based on bootstrap bias correction, linear interpolation, and normal approximation, thus they are not exact. In order to explore how well they perform for finite samples, we carried out a simulation study. To enable a comparison with the RCL lower confidence limit and the jackknife confidence limits by Dayton (2003), we included those too in the simulation study. Similarly to Rudas et al. (1994) and Pan and Dayton (2011), we investigated the performance of the methods in the context of independence model in two-way contingency tables. We included table sizes of 3x5, 5x7, 7x10, and 10x14. For each table size, sample sizes were 10, 100, and 1000 per cell on average.

For each combination of table size and sample size we generated 2000 populations (that is, distributions) and for each of them we determined its pi-star by the iterative procedure proposed by Rudas, Clogg, and Lindsay (1994) and implemented in the R package ‘pistar’ (Medzihorsky, 2013). Then we drew a random sample from each population, determined the sample pi-star, and computed the 95% bootstrap, jackknife, and RCL confidence limits for the population pi-star, and recorded whether or not the confidence limits covered the corresponding population pi-star. In the followings, we give the details of data generation.

Each population distribution D was generated as follows.

1. First a k by m table D_0 with independent marginals was generated. Each marginal distribution was a vector of uniform random numbers from [0.1, 1], divided by their sum, and cell probabilities were obtained as the product of the corresponding marginal probabilities.

2. Then a k by m table of noise (D_1) was generated with uniform random numbers from [0.1, 1], divided by their sum.

3. The population distribution was generated by mixing D_0 and D_1 according to $D = (1–p)D_0 + pD_1$, where p was a uniform random number from [0.1, 1].

By applying this procedure, the pi-star of the generated population D is a random number that is typically less than the mixing proportion p. The reason for using this method is that no procedure is known for generating a distribution with a prescribed pi-star. In the simulation study, the range of population pi-star values extended from 0 to 0.6, what we found sufficient to draw practical conclusions. We decided to generate a mixture of model and noise because when we generated just arbitrary tables (that is, pure noise), low pi-star values occurred very rarely, unlike practical applications, where low pi-star values may also occur.

From each population one single sample was drawn. The reason for this was that, although this made the simulation slower, we wanted to include as many different populations as possible. In each simulation run with a fixed table size, 2000 populations were generated, and from each population the sample was taken according to the multinomial model (with fixed table total). Altogether 12 such runs were made (4 table sizes combined with 3 sample sizes).

In the simulation we applied the bootstrap confidence limits defined by (9) and (10) with B = 20. For the simulation we used our random search algorithm and Medzihorsky's (2013) 'pistar' package. In a small validation run we compared the two procedures and found good agreement between the estimates.

Simulation results

Table 1 shows that the bias correction was successful, the new point estimate has considerably smaller bias than the sample pi-star for all simulated table and sample sizes. At the same time, its deviation from the true population pi-star is also smaller in most cases, no matter if it is measured by the absolute or squared deviation.

(Table 1 here)

Coverage rates of the lower confidence limit and the two-sided confidence interval obtained by the proposed method are presented in Tables 2 and 3. Each item in the tables is estimated from simulation with 2000 Monte Carlo replications.

(Tables 2 and 3 here)

Table 2 shows that the coverage of the bootstrap lower confidence limit for various table sizes and sample sizes varies between 89.7% and 94.4% for 90%; between 94.9% and 97.9% for 95%; and between 98.5% and 99.6% for 99% nominal confidence level. Table 3 shows that the coverage of the two-sided bootstrap confidence interval varies between 88.4% and 93.3% for 90%; between 93.6% and 96.7% for 95%; and between 98.2% and 99.3% for 99% nominal confidence level. Thus, according to the simulation results, the coverage of the lower confidence limit is acceptable for all table sizes, even for cell counts as small as 10/cell on average. For small contingency tables also the two-sided confidence interval has acceptable coverage, but it seems that for larger tables combined with small sample sizes the coverage of the two-sided interval gets too low.

Table 4 reports the coverage of the 95% RCL lower confidence limit and the jackknife lower confidence limit for various table sizes and sample sizes, and shows that both the lower confidence limit by Rudas et al. (1994) and the jackknife limit by Dayton (2003) performs rather poorly even for samples as large as 10000/cell. In the worst cases the coverage turned out to be as low as 53.2% for the RCL, and 55.8% for the jackknife limit. Here too, coverage values are based on 2000 Monte Carlo replicates.

(Table 4 here)

4. Applications

In this section we present four practical applications with real data, including applications to social science and psychology, and we show that the proposed method is useful in practice.

Let us consider first the two examples analysed in Rudas et al. (1994). In both cases, the model of interest is the independence model in a two-way contingency table. The first example, originally considered by Snee (1974), involves a 4 by 4 table cross-classifying eye colour and hair colour of 592 subjects. The proposed bootstrap procedure with $B = 50$ (that is, with 50 bootstrap replications from the sample and 50 replications from the model) results in a sample pi-star of 0.296, which is fairly close to 0.298, the value obtained by Rudas et al. (1994), and is equal to that calculated by Ispány and Verdes (2014). This value was obtained by our random search algorithm as well as by Medzihorsky's (2013) 'pistar' package. The estimated population pi-star, after correcting the bias of the sample pi-star, is 0.262. Our 90% bootstrap confidence interval is from 0.199 to 0.325, the lower endpoint of which can be regarded as a 95% lower confidence limit for the population pi-star. This means that the data deviate considerably from the independence model, which is in accordance with the chi-square test ($\chi^2 = 138.3$, df = 9, $p<0.0001$) but the pi-star analysis reveals more details of their relation. The bias-corrected pi-star is markedly lower than those by Rudas et al. (1994) and Dayton (2003), which are 0.236 and 0.230, respectively. Note that the upper confidence limit, 0.325 also has a clear meaning: it is not reasonable to think that the population pi-star is greater than this value because from a population with pi-star higher than 0.325 one would expect a sample with larger sample pi-star.

The second example, which Rudas et al. (1994) took from Cramer (1946), is a 5x4 table cross-classifying the number of children by grouped annual income level of 25263 households. For this contingency table the proposed bootstrap procedure (again with $B = 50$) results in a sample pi-star of 0.102, and in a bias-corrected estimate of the population pi-star of 0.099. Our 90% confidence interval is from 0.087 to 0.112. Here too, the sample pi-star is fairly close to 0.104, that obtained by Rudas et al. (1994), but the lower confidence limit is again a little below that obtained by them (0.091). Surprisingly, the jackknife lower confidence limit is here higher (0.093) than the asymptotic one. Here again, although the results of the classical association analysis are similar to that of the pi-star analysis (Cramer's V = 0.087, $\chi^2 = 568.6$, df = 12, p<0.0001), the latter quantifies the deviation in a more palpable way.

Our third example, taken from Richburg et al. (2009), addresses an important issue in higher education learning: the relationship between student gender and course grades. The data are displayed in Table 5, summarizing gender and grade of 778 students in 30 different classes at a small regional university. The literature on the relationship between student characteristics and academic achievement in higher education is rather vast. However, most studies measure academic achievement through enrolment figures, degree attainment figures and grade point averages, rather than course grades, what makes the study by Richburg et al. (2009) remarkable.

(Table 5 here)

For this table, the sample pi-star is 0.076 and the estimated population pi-star after bias correction is 0.040. Our 90% bootstrap confidence interval is from 0 to 0.108. This estimate means that independence between gender and grade can be assumed for 96% of the student population. As the lower confidence limit is 0, the sample does not provide strong evidence against independence of gender and grade, given the deviation is measured by the mixture index pi-star. At the same time, the asymptotic and jackknife lower confidence limits are 0.029 and 0.016, respectively. In the light of the above simulation results, presumably both of them are biased, in particular the asymptotic one. Classical methods for the analysis of association detected a weak but statistically significant association between gender and grade. Cramer's V coefficient was 0.126, and chi-square test (as well as Fisher's exact test) resulted in p=0.0152. However, stating that this association affects about 4% of the students (and presumably not more than 11% of them) is much easier to interpret. The lower confidence limit being 0, the pi-star analysis concludes that these data are no strong evidence against independence. However, thinking of the critics against the chi-square test that it rejects the null hypothesis too often for large samples (Diaconis and Efron, 1985), this may even be considered as an advantage.

In the last application we analyze the data set in Agresti (2002), page 322, representing the results of a survey in which 2276 high school students were asked in their last year in school whether they had ever used alcohol, cigarettes, or marijuana. The data set has been re-analyzed several times and is also available in the R package vcdExtra (Friendly, 2015). First we fit the independence model that assumes full independence of the three factors. For this model, the sample pi-star is 0.330, and the 95% asymptotic and jackknife lower confidence limits are equal, namely 0.314. Our method (with B=50 replications) results in a bias-corrected estimate of the population pi-star 0.304. The 90% CI (whose lower endpoint is a 95% lower confidence limit) is [0.127, 0.481]. These results indicate that the fit of the independence model is rather poor. This is in agreement with the result of chi-square test for independence ($\chi^2 = 2676$, df = 7, p<0.0001), but expresses it more specifically. It should also be emphasized that the RCL as well as the jackknife lower confidence limits are much higher than the lower endpoint of the bootstrap CI. This is in accordance with our simulation results that the RCL limits may not maintain the nominal confidence level.

Next we consider the conditional independence model assuming that conditioned on the smoking status, alcohol and marihuana use are independent. For this model, the sample pi-star is 0.036, and the 95% asymptotic and jackknife lower confidence limits are equal again, 0.029. Applying our method to this model, it turns out that in this case the bias of the sample estimate is negligible. The bias-corrected estimate of the population pi-star is also 0.036, with a 90% CI of [0.023, 0.048]. Here too, the asymptotic and jackknife lower confidence limits are somewhat higher than the lower endpoint of the new CI, but now the difference is rather small. Chi-square analysis results in a rather bad fit of the model ($p<0.0001$) which demonstrates again sensitivity of the goodness-of-fit tests to sample size (Diaconis and Efron, 1985). Although pi-star analysis does not support model fit either, it quantifies the deviation in a more specific and interpretable way, namely that this affects about 3% (at least 2%, at most 5%) of the population.

5. Discussion

Although computation of pi-star is very time-consuming, and procedures for this are not included in the most widely used commercial software packages, its applications have appeared from time to time in various research fields, among others in psychometrics (Schuster, 2002; Dayton, 2003; Formann, 2006; Revuelta, 2008).

By now, the increasing speed of computers made it possible to examine the finite sample properties of the RCL asymptotic lower confidence limit and the jackknife confidence limit, and to compare the proposed new confidence interval to them. Our simulations revealed that both the asymptotic and the jackknife confidence limits have rather low coverage probabilities even with as large samples as 10000 per cell on average (Table 4). At the same time, the new CI has good coverage in most of the examined cases (Tables 2 and 3). It should be noted that our study has the limitation that we investigated the performance of the procedures only for two-way contingency tables.

At first sight it may seem strange that the jackknife performs so poorly for the largest sample sizes. The reason for this is that in case of a very large sample removing one single sample element causes just a minuscule change in the pi-star that is below the computational precision, which leads to numerically indistinguishable jackknife pi-star values, that is, to zero jackknife standard error, and as a result, the confidence interval fails.

Xi and Lindsay (1996) noticed that the EM algorithm proposed by Rudas et al. (1994) may result in an infeasible solution, that is, in which the model counts are greater than the observed ones, while the algorithm proposed by Xi and Lindsay does not suffer from this. We also experienced this when using the R package by Medzihorsky (2013), although in most cases it affected just the third decimal of the resulting pi-star value.

Our ad hoc method uses the bootstrap to estimate the bias and standard error of the sample pi-star, then linear interpolation to correct for the bias, and normal approximation to compute the confidence limits, of which linearity is just a crude approximation. It is a future task to find a more realistic curve for describing the relation between population and sample pi-star, which may reduce length of CI.

The question arises whether pure bootstrap would produce comparable results with this hybrid procedure. To explore this, we made simulation experiments with the two best bootstrap confidence intervals, the percentile interval and the bias-corrected accelerated (BCa) interval (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996) applying them to 3 by 5 tables with n=15000. In spite of the much higher number of bootstrap replications (B=500, 12.5-fold of that used in the simulation with our method), the experiment resulted in disappointingly low coverage probabilities (88% and 77% at a nominal level of 95%).

Although Rudas et al. (1994) focused on the lower confidence limit (most probably for computational reasons), it is meaningful to calculate an upper limit as well. As pi-star is a functional of a distribution, and it has a sampling distribution, it is also reasonable to search for two-sided confidence intervals. Even from the practical point of view, it is clear that one does not expect excellent model fit for a sample drawn from a population which is rather poorly described by that model. The two-sided CI represents the range of population pi-star values which are in accordance with the observed sample. It is plausible to reject too small as well as too large population pi-star values based on a sample from the population.

The coverage of the two-sided bootstrap CI is not as good as that of the lower confidence limit. The reason for this is the linear approximation based on a population with pi-star=0. Simulations confirmed the expectation that while the lower limit is conservative, the upper limit is too liberal. A possible direction for future research is to apply double bootstrap around the observed pi-star to obtain a better solution (Booth and Hall, 1994). Unfortunately, this would increase the computing time so much that it would prevent from exploring the finite sample properties of the method by simulation, at least with the current computing facilities (Booth and Presnell, 1998). Another way to improve the coverage would be to find a non-linear function that describes better the relation between population and sample pi-star than the linear approximation applied here.

It is important to recognize that the numerical precision of the pi-star calculation may have a non-negligible influence on the estimates. Especially with very large samples, when the magnitude of the numerical and sampling errors are comparable, the numerical imprecision may seriously affect the point estimate as well as the jackknife or bootstrap standard error. Therefore, as a general rule, the numerical error should be kept by one magnitude lower than the expected sampling error at the particular sample size. The algorithms allow this at the expense of longer computing time.

References

Agresti, A. (2002). Categorical Data Analysis, 2nd ed. John Wiley & Sons, Hoboken, New Jersey, USA.

Booth, J. G., Hall, P. (1994). Monte Carlo approximation and the iterated bootstrap. *Biometrika*, 81, 331-340.

Booth, J., Presnell, B. (1998). Allocation of Monte Carlo Resources for the Iterated Bootstrap. *Journal of Computational and Graphical Statistics,* 7, 92-112.

Boyd, S., Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press, Cambridge, UK.

Böckenholt, U., van den Heijden, P. G. M. (2007). Item randomized-response models for measuring noncompliance: risk-return perceptions, social influences, and self-protective responses. *Psychometrika,* 72, 245-262.

Cramer, H. (1946). Mathematical Methods in Statistics. Princeton University Press, Princeton, USA.

Davison, A.C., Hinkley, D.V., Young, G.A. (2003). Recent Developments in Bootstrap Methodology. S*tatistical Science*, 18, 141-157.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology,* 56, 1-13.

Diaconis, P., Efron, B. (1985) Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals of Statistics,* 13, 845-874.

DiCiccio, T. J., Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science,* 11, 189-228.

Efron, B., Tibshirani, R. (1993). An introduction to the bootstrap. Chapman & Hall, London, U.K.

Formann, A.K. (2006). Testing the Rasch model by means of the mixture fit index. *British Journal of Mathematical and Statistical Psychology,* 59, 89-95.

Friendly, M. (2015). vcdExtra: 'vcd' Extensions and Additions. R package version 0.6-8. http://CRAN.R-project.org/package=vcdExtra (assessed on 2 June 2016)

Ispány, M., Verdes, E. (2014). On the robustness of mixture index of fit. *Journal of Mathematical Sciences,* 200, 4, 432–440.

Loeve, M. (1977). Probability Theory I, 4th ed. Springer-Verlag, New York, USA.

Medzihorsky, J. (2013) Manual of the R Package "pistar", version 0.5.2, https://github.com/jmedzihorsky/pistar/blob/master/pistar_0.5.2_manual.pdf (assessed on 29 April 2016)

Pan, X., Dayton, C. M. (2011). Factors Influencing the Mixture Index of Model Fit in Contingency Tables Showing Independence. *Journal of Modern Applied Statistical Methods,* Volume 10, Issue 1, Article 16.

Revuelta, J. (2008). Estimating the pi* goodness of fit index for finite mixtures of item response models. *British Journal of Mathematical and Statistical Psychology,* 61, 93-113.

Richburg, K., Lewis, W., Barbour, K., Caines, W. R. (2009). Academic grades: Does race or gender matter? *The Journal of Learning in Higher Education,* 5, 13-20.

Rudas, T. (1999). The mixture index of fit and minimax regression. *Metrika,* 50, 163-172.

Rudas, T., Clogg, C. C., & Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B,* 56, 623–639.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology,* 55, 289-303.

Snee, R. (1974). Graphical display of two-way contingency tables. *The American Statistician,* 38, 9-12.

Verdes, E., Rudas, T. (2003). The π* Index as a New Alternative for Assessing Goodness of Fit of Logistic Regression. In: Haitovsky, Lerche, Ritov (eds.), *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000,* Springer Physica Verlag, Heidelberg.

Xi, L. & Lindsay, B. G. (1996). A note on calculating the π* index of fit for the analysis of contingency tables. *Sociological Methods and Research,* 25, 248–259.

Table 1. Comparison of the new point estimate to the sample pi-star with respect to bias, mean absolute deviation (MAD) and root mean squared deviation (RMSD) from the true population pi-star. Each entry is calculated from 2000 simulated populations and samples.

		Sample pi-star			Bias-corrected pi-star		
		Bias	MAD	RMSD	Bias	MAD	RMSD
3x5,	*n*=150	0.097	0.104	0.123	0.001	0.086	0.107
	n=1500	0.016	0.027	0.034	-0.005	0.027	0.035
	n=15000	0.002	0.008	0.010	-0.001	0.008	0.011
5x7,	*n*=350	0.133	0.135	0.154	-0.007	0.082	0.103
	n=3500	0.023	0.029	0.037	-0.006	0.024	0.031
	n=35000	0.003	0.008	0.010	-0.002	0.007	0.010
7x10,	*n*=700	0.156	0.156	0.175	-0.010	0.073	0.092
	n=7000	0.028	0.032	0.040	-0.008	0.021	0.027
	n=70000	0.004	0.007	0.010	-0.001	0.006	0.008
10x14,	*n*=1400	0.184	0.184	0.203	-0.006	0.063	0.080
	n=14000	0.034	0.035	0.045	-0.008	0.019	0.024
	n=140000	0.005	0.008	0.010	-0.001	0.006	0.007

Table 2. Coverage of the bootstrap lower confidence limit for various table sizes and sample sizes.

Table size	*n* (*n*/cell)	Coverage of the 90% lower CL	Coverage of the 95% lower CL	Coverage of the 99% lower CL
3x5	150 (10/cell)	0.912	0.957	0.989
	1500 (100/cell)	0.927	0.964	0.994
	15000 (1000/cell)	0.913	0.960	0.992
5x7	350 (10/cell)	0.910	0.955	0.993
	3500 (100/cell)	0.933	0.969	0.994
	35000 (1000/cell)	0.915	0.958	0.990
7x10	700 (10/cell)	0.903	0.949	0.989
	7000 (100/cell)	0.944	0.974	0.994
	70000 (1000/cell)	0.919	0.958	0.989
10x14	1400 (10/cell)	0.897	0.951	0.985
	14000 (100/cell)	0.949	0.979	0.996
	140000 (1000/cell)	0.911	0.952	0.988

Table 3. Coverage of the two-sided bootstrap confidence interval for various table sizes and sample sizes.

Table size	*n* (*n*/cell)	Coverage of the 90% two-sided CI	Coverage of the 95% two-sided CI	Coverage of the 99% two-sided CI
3x5	150 (10/cell)	0.933	0.967	0.990
	1500 (100/cell)	0.918	0.961	0.991
	15000 (1000/cell)	0.908	0.951	0.987
5x7	350 (10/cell)	0.913	0.960	0.993
	3500 (100/cell)	0.917	0.958	0.992
	35000 (1000/cell)	0.894	0.947	0.986
7x10	700 (10/cell)	0.888	0.947	0.991
	7000 (100/cell)	0.913	0.960	0.993
	70000 (1000/cell)	0.903	0.946	0.987
10x14	1400 (10/cell)	0.876	0.936	0.985
	14000 (100/cell)	0.885	0.942	0.989
	140000 (1000/cell)	0.884	0.939	0.982

Table 4. Coverage of the 95% RCL lower confidence limit and the jackknife lower confidence limit for various table sizes and sample sizes.

Table size	Average sample size /cell					
	100		1000		10000	
	RCL	jackknife	RCL	jackknife	RCL	jackknife
3x5	0.870	0.892	0.904	0.924	0.938	0.794
5x7	0.740	0.784	0.878	0.858	0.926	0.758
7x10	0.619	0.686	0.826	0.732	0.906	0.716
10x14	0.532	0.558	0.771	0.740	0.923	0.688

Table 5. Contingency table of grade by gender of 778 university students, as reported by Richburg et al. (2009).

	Grade				
Gender	A	B	C	D	F
Female	76	128	130	45	31
Male	51	104	121	38	54

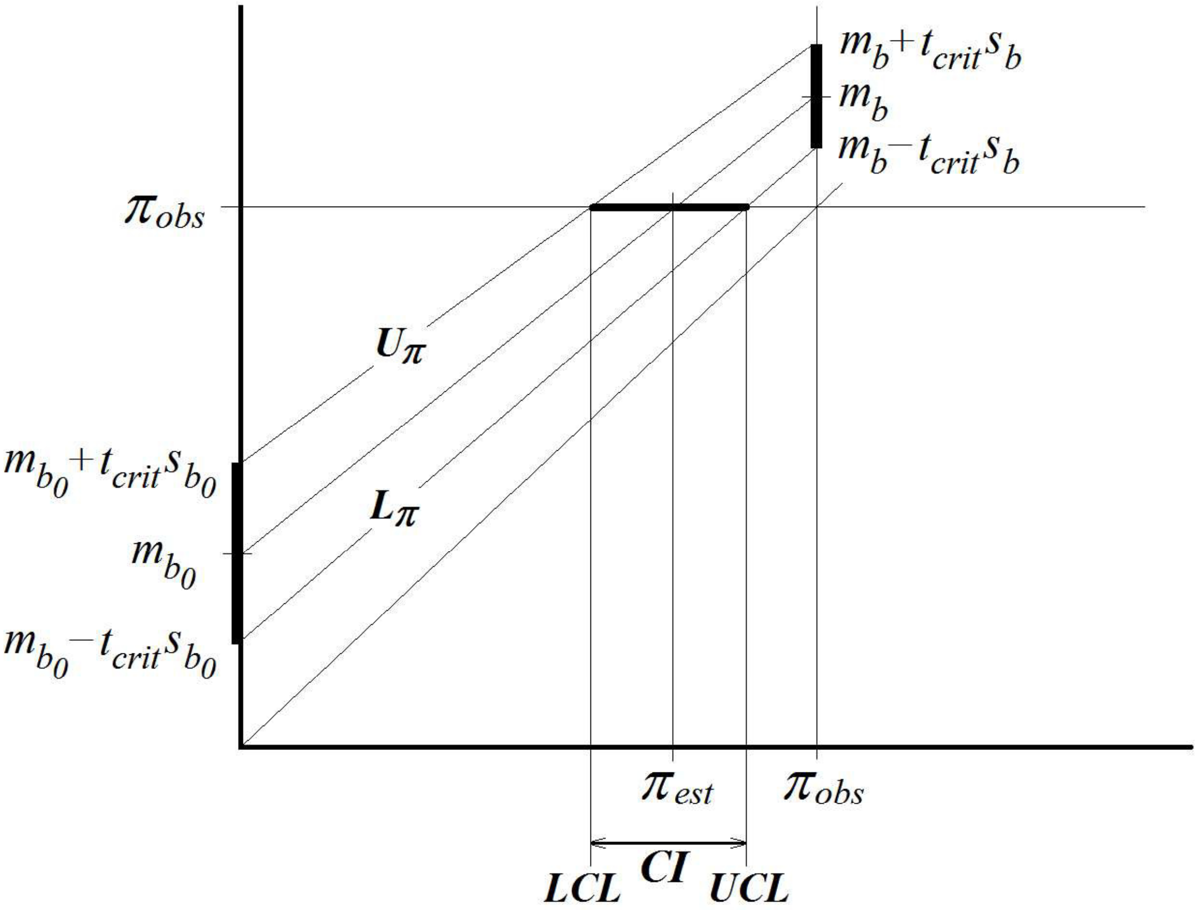

Figure 1. Illustration of the linear interpolation of the bias and SE of the sample pi-star, and construction of the confidence interval. The *x* axis represents the pi-star of the distribution from which the resamples are generated, the *y* axis represents the pi-star of the resamples