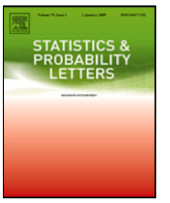
Kendall's tau estimator for bivariate zero-inflated count data

Statistics and Probability Letters 199 (2023) 109858

Contents lists available at ScienceDirect

Statistics and Probability Letters

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



Kendall's tau estimator for bivariate zero-inflated count data

Elisa Perrone *, Edwin R. van den Heuvel, Zhuozhao Zhan

Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 5, 5612 AZ, Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 8 August 2022
Received in revised form 27 March 2023
Accepted 20 April 2023
Available online 28 April 2023

Keywords:
Kendall's tau
Bivariate zero-inflated count data
Fréchet–Hoeffding bounds

ABSTRACT

This paper extends the work of Pimentel et al. (2015), presenting an estimator of Kendall's τ for bivariate zero-inflated count data. We provide achievable bounds of our proposed estimator and suggest how to estimate them, thereby making the estimator useful in practice.

1. Introduction

Zero-inflated data naturally appears in many applications such as health care and ecology (Moulton and Halsey, 1995; Arab et al., 2012). Analyzing zero-inflated data is challenging as the high amount of observations in zero invalidates standard statistical techniques. For example, assessing the level of dependence between two zero-inflated random variables becomes a difficult task as standard rank-based measures of association such as Kendall's τ and Spearman's ρ cannot be applied directly due to the large amount of tied values in zero making any tie-breaking adjustment unsatisfactory (Hollander and Wolfe, 2013). The importance of deriving consistent estimators of popular association measures, such as Kendall's τ and Spearman's ρ, for bivariate zero-inflated distributions motivated recent work on the topic. In Pimentel (2009) and Pimentel et al. (2015), the authors focused on zero-inflated continuous distributions and proposed new estimators for Kendall's τ with reduced bias. Denuit and Mesfioui (2017) derived lower and upper bounds of the newly introduced estimator, making its interpretation possible as a measure of the strength of dependence. The abundance of zero-inflated count data in practice, e.g., zero-inflated Poisson-type data, makes it crucial to define measures of dependence that can handle discreteness of the data as well as it being zero-inflated. In this paper we extend the work of Pimentel et al. (2015) and propose a new estimator of Kendall's τ for bivariate random variables with zero-inflated discrete distributions. We complete the picture by deriving the theoretical lower and upper bounds of the proposed estimator, and we compare them with the bounds obtained by Denuit and Mesfioui (2017) for Pimentel's estimator. As an illustration, we show the performance of our proposed estimator in several simulated scenarios based on zero-inflated Poisson distributions. The paper is structured as follows: Section 2 introduces the notation and basic concepts. In Section 3, we present our proposed estimator, and we discuss its attainable theoretical bounds in Section 4. In Section 5, we evaluate the performance of the estimator via a simulation study. We end with a discussion and conclusion section in Section 6.

* Corresponding author.
E-mail address: e.perrone@tue.nl (E. Perrone).

https://doi.org/10.1016/j.spl.2023.109858

2. Background and notation

We consider two independent copies $(\tilde{X}_1, \tilde{Y}_1)$ and $(\tilde{X}_2, \tilde{Y}_2)$ of the random vector (X, Y) with joint cumulative distribution function H. Kendall's τ is defined as the probability of concordance minus the probability of discordance (Kendall, 1938). For continuous random vectors, this definition results in $\tau = \mathbb{P}[(\tilde{X}_1 - \tilde{X}_2)(\tilde{Y}_1 - \tilde{Y}_2) > 0] - \mathbb{P}[(\tilde{X}_1 - \tilde{X}_2)(\tilde{Y}_1 - \tilde{Y}_2) < 0] = 2\mathbb{P}[(\tilde{X}_1 - \tilde{X}_2)(\tilde{Y}_1 - \tilde{Y}_2) > 0] - 1$. When X and Y assume values in the non-negative integers, Kendall's τ also depends on the probability of ties, i.e., $\tau = 2\mathbb{P}[(\tilde{X}_1 - \tilde{X}_2)(\tilde{Y}_1 - \tilde{Y}_2) > 0] - 1 + \mathbb{P}[\tilde{X}_1 = \tilde{X}_2 \text{ or } \tilde{Y}_1 = \tilde{Y}_2]$. The non-continuous case has extensively been studied in Denuit and Lambert (2005), Mesfioui and Tajar (2005), Nešlehová (2007), Nikoloulopoulos and Karlis (2008), and Nikoloulopoulos and Karlis (2009), where the authors also give closed-form formulas to calculate Kendall's τ for general discrete distributions when the distribution is completely known.

We denote as $\hat{\tau}$ the standard estimator of Kendall's τ computed by replacing the probability of concordance and discordance with the corresponding sample frequencies, which is by counting the number of concordant and discordant pairs and divide by the total number of pairs (Kendall, 1938). In case of ties, i.e., repeated values in the sample, there are pairs that are neither concordant nor discordant. To account for this, an adjusted version of the estimator of Kendall's τ which excludes the tied pairs from the count has been proposed (Kendall, 1945). We denote this by τ_b. To define our theoretical framework, we use a similar notation as in Pimentel et al. (2015) and Denuit and Mesfioui (2017). We consider two non-negative random variables X and Y that follow two discrete distributions (e.g., Poisson) with extra positive probability mass at zero, i.e., the cumulative distribution function (cdf) of X, F, and of Y, G, can be written as follows

$$F(s) = \begin{cases} 0, \text{ if } s < 0 \\ (1 - \pi_F) + \pi_F \cdot \tilde{F}(s), \text{ if } s \geq 0 \end{cases} \qquad G(t) = \begin{cases} 0, \text{ if } t < 0 \\ (1 - \pi_G) + \pi_G \cdot \tilde{G}(t), \text{ if } t \geq 0 \end{cases}$$

where, $\tilde{F}$ and $\tilde{G}$ are discrete distribution functions (e.g., Poisson), while for Pimentel et al. (2015) they are continuous. The probabilities $(1 - \pi_F)$ and $(1 - \pi_G)$ represent the extra zero inflation of the distribution. Since $\tilde{F}$ and $\tilde{G}$ are discrete, the total probability mass in zero is equal to $(1 - \pi_F) + \pi_F \cdot \tilde{F}(0)$ for X, and $(1 - \pi_G) + \pi_G \cdot \tilde{G}(0)$ for Y.

Following (Denuit and Lambert, 2005), for two independent copies $(\tilde{X}_1, \tilde{Y}_1)$ and $(\tilde{X}_2, \tilde{Y}_2)$ of the random vector (X, Y) with underlying copula C, we define $\mathbb{P}_C(\text{tie}) = \mathbb{P}(\tilde{X}_1 = \tilde{X}_2) + \mathbb{P}(\tilde{Y}_1 = \tilde{Y}_2) - \mathbb{P}(X_1 = X_2, \tilde{Y}_1 = \tilde{Y}_2)$. Then, we indicate with $p^U_{t_{11}} = \mathbb{P}_M(\text{tie})$ and $p^L_{t_{11}} = \mathbb{P}_W(\text{tie})$ the probability that either X_1 or Y_1 are tied when the joint distribution of (X, Y) is the upper (lower) Fréchet–Hoeffding bound M (W). We denote by X_{10} a positive random variable distributed as X given that $Y = 0$, X_{11} a positive random variable distributed as X given that $Y > 0$. Similarly, Y_{01} is a positive random variable distributed as Y given that $X = 0$, and Y_{11} a positive random variable distributed as Y given that $X > 0$. We also consider X_1 a random variable distributed as X given $X > 0$ (whose distribution is $\bar{F}$), and Y_1, analogously (with corresponding distribution $\bar{G}$). In addition, we define the following probabilities: $p_{00} = \mathbb{P}[X = 0, Y = 0]$, $p_{01} = \mathbb{P}[X = 0, Y > 0]$, $p_{10} = \mathbb{P}[X > 0, Y = 0]$, $p_{11} = \mathbb{P}[X > 0, Y > 0]$, $p^*_1 = \mathbb{P}[X_{10} > X_{11}]$, $p^*_2 = \mathbb{P}[Y_{01} > Y_{11}]$, and define τ_{11} as Kendall's τ of (X_1, Y_1), i.e., away from zero. Then, the association measure τ_H for the random vector (X, Y) considered in Pimentel et al. (2015) is given by the following formula:

$$\tau_H = p^2_{11}\tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10}) + 2p_{11}[p_{10}(1 - 2p^*_1) + p_{01}(1 - 2p^*_2)] \tag{1}$$

Pimentel et al. (2015) suggested an estimator $\hat{\tau}_H$ of Eq. (1) by replacing all probabilities of the formula with the corresponding sample frequencies, and, since no ties are expected away from zero, by substituting τ_{11} with the standard estimator $\hat{\tau}$ of Kendall's τ calculated from data on X and Y where X and Y are both positive. In addition, Denuit and Mesfioui (2017) proved that the attainable bounds of the association measure τ_H only depend on the zero-inflation probabilities $p_1 = \mathbb{P}[X = 0]$ and $p_2 = \mathbb{P}[Y = 0]$. Specifically, the bounds are given by the following formulas

$$\tau_H^{upper} = \begin{cases} 1 - p_2^2, \text{ when } p_1 \leq p_2 \\ 1 - p_1^2, \text{ when } p_1 \geq p_2 \end{cases} \tag{2}$$

$$\tau_H^{lower} = \begin{cases} -2(1 - p_1)(1 - p_2), \text{ when } 1 - p_1 - p_2 < 0 \\ (1 - p_1 - p_2)^2 - 2(1 - p_1)(1 - p_2), \text{ when } 1 - p_1 - p_2 > 0 \end{cases} \tag{3}$$

Although τ_H was not designed for zero-inflated count data, a natural question arises whether or not it is sufficient to replace the estimator of τ_{11} with τ_b in Eq. (1) to obtain an estimator of Kendall's τ that also works for such a data. The result established in the next section shows that this is not the case, and further adjustments are needed.

3. Estimator of Kendall's τ for zero-inflated count data

The association measure studied in Pimentel et al. (2015) is based on a separation of the zero-inflated part from the continuous part of the distribution. Pimentel et al. (2015)'s estimator is interesting since, on the one hand, it accounts for the ties in zero and, on the other hand, it acts as the standard estimator of Kendall's τ away from zero. Our approach is based on a similar idea of decomposing the association measure around zero and away from zero. However, due to the discrete nature of the zero-inflated count data away from zero, the estimator proposed by Pimentel et al. (2015) cannot be applied directly without further adjustments due to the non-zero probability of ties within the margins. The next result tackles this issue and establishes Kendall's τ for zero-inflated count data.

Theorem 1. *We define the probabilities of ties within the margins as* $p_1^\dagger = \mathbb{P}[X_{10} = X_{11}]$, *and* $p_2^\dagger = \mathbb{P}[Y_{01} = Y_{11}]$. *Then Kendall's* τ *is given by the following relation*

$$\tau_A = p_{11}^2\tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10}) + 2p_{11}[p_{10}(1 - 2p_1^* - p_1^\dagger) + p_{01}(1 - 2p_2^* - p_2^\dagger)]. \tag{4}$$

The proof of this theorem is straightforward and is similar to Pimentel (2009). Based on the definition of Kendall's tau, we derived the expressions for the probability of concordance and discordance respectively using the law of total probabilities. The complete proof is given in the supplementary file. As suggested in Pimentel et al. (2015), an estimator $\hat{\tau}_A$ of τ_A can be obtained by replacing probabilities with their estimates based on relative frequencies, while τ_{11} with the standard tie-corrected Kendall's τ estimator τ_b (Kendall, 1945). Moreover, consistency and asymptotic normality of the estimator $\hat{\tau}_A$ follows directly from the same arguments presented in Pimentel et al. (2015) for the estimator of τ_H.

4. Attainable bounds for τ_A

Kendall's τ cannot reach the theoretical bounds ± 1 if there is a discrete component in the random vector. In light of this, knowing the attainable bounds of the estimators of Kendall's τ is crucial to assess the strength of association of the data. To make our proposed estimator τ_A of Eq. (4) useful in practice, we derive the range of admissible values for τ_A in terms of the marginal distributions of X and Y. To do so, we follow the approach of Denuit and Mesfioui (2017), which is based on the property of monotonicity of Kendall's τ with respect to the concordance order (Denuit and Lambert, 2005; Mesfioui and Tajar, 2005).

Proposition 1. *The lower and upper bounds of the association measure* τ_A *of Eq. (4) are given by*

$$\tau_A^{upper} = \begin{cases} (1-p_2^2) - (1-p_2)^2 p_{t_{11}}^U - 2(p_2 - F(\tilde{s}-1))(F(\tilde{s}) - p_2), \text{ if } p_1 \le p_2 \\ (1-p_1^2) - (1-p_1)^2 p_{t_{11}}^U - 2(p_1 - G(\tilde{t}-1))(G(\tilde{t}) - p_1), \text{ if } p_1 \ge p_2 \end{cases}$$

$$\tau_A^{lower} = \begin{cases} -2(1-p_1)(1-p_2) & \text{if } 1 - p_1 - p_2 < 0 \\ p_1^2 + p_2^2 - 1 + (1-p_1-p_2)^2 \cdot p_{t_{11}}^L + 2[(F(\tilde{s}') + p_2 - 1)(1 - p_2 - F(\tilde{s}'-1)) + & \\ \qquad + (G(\tilde{t}') + p_1 - 1)(1 - p_1 - G(\tilde{t}'-1))], & \text{if } 1 - p_1 - p_2 > 0 \end{cases}$$

where $\tilde{s}$ *is a point such that* $F(\tilde{s}) > p_2$ *and* $F(\tilde{s}-1) \le p_2$, *and* $\tilde{s}'$ *is a point such that* $F(\tilde{s}') + p_2 - 1 > 0$ *and* $F(\tilde{s}') + p_2 - 1 \le 0$ *(analogously for* $\tilde{t}$ *and* $\tilde{t}'$*).*

We notice that the points $\tilde{s}$, $\tilde{s}'$, $\tilde{t}$, $\tilde{t}'$ and the corresponding expressions are closely related to the joint probability expressed in terms of the Fréchet–Hoeffding bounds $\min\{F(x), G(y)\}$, and $\max\{F(x) + G(y) - 1, 0\}$. The complete proof of Proposition 1 is available in the supplementary file. Although the bounds reported in Proposition 1 appear to be more involved than the bounds reported in Denuit and Mesfioui (2017), they correspond to the ones established in Denuit and Mesfioui (2017) when the part away from zero is continuous (see Remark 1 in the supplementary file). Moreover, it is still possible to estimate them from the data. The probabilities of zero-inflation of X and Y, i.e., p_1 and p_2, can be replaced by the corresponding relative frequencies. The values $p_{t_{11}}^U$ and $p_{t_{11}}^L$ are both dependent on the (joint) distribution. Nevertheless, as noticed in Denuit and Lambert (2005), they can be replaced by their own lower bound $\max(\mathbb{P}[X_1 = X_2], \mathbb{P}[Y_1 = Y_2])$, which can be readily estimated from the sample without the knowledge of the distributions. Therefore, an estimator of a slightly wider range of τ_A can be constructed by substituting $p_{t_{11}}^U$ and $p_{t_{11}}^L$ with the maximum sample frequency of X_1 or Y_1 being tied. Finally, the remaining values in the formulas, i.e., $F(\tilde{s})$, $F(\tilde{s}-1)$, $F(\tilde{s}')$, $F(\tilde{s}'-1)$, and the analogous quantities for G, can be estimated via the empirical cdfs of X and Y.

5. Simulation study

The theoretical results developed in this work hold for arbitrary zero-inflated count data. As an illustration, we here construct a simulation study based on various parameter choices for zero-inflated Poisson distributions. In particular, to investigate the performance of our proposed estimator, we conducted a Monte-Carlo simulation study based on 1000 repetitions. Namely, we computed the values of the estimators of τ, the adjusted version of τ_H with τ_{11} estimated via τ_b due to the discrete nature of our data, and τ_A for N pairs generated from two correlated zero-inflated Poisson distributions joined through the Fréchet copula $C(u, v) = (1-\rho)uv + \rho\min(u, v)$, where $u, v, \rho \in [0, 1]$ (Nelsen, 2006). Thus, the parameters of the full distribution are five, i.e., $\pi_F, \pi_G, \lambda_F, \lambda_G, \rho$, where ρ is the copula parameter, π_F is the probability mass spread according to a Poisson distribution with mean parameter λ_F (analogously for π_G and λ_G), and $1-\pi_F$ represents the additional probability mass in zero which does not originate from the Poisson distribution of X (same for $1 - \pi_G$). We selected multiple scenarios representative of various characteristics of the samples. In particular, we considered three values for the copula parameter $\rho = 0.2, 0.5, 0.8$ depicting different strengths of association, two proportions $\pi_F = \pi_G = 0.2, 0.8$ for the probability mass associated with the Poisson distributions, and three combinations for the Poisson mean parameters $(\lambda_F, \lambda_G) = \{(2, 2); (2, 8); (8, 8)\}$ corresponding to different levels of probability of ties away from zero. We conducted our analysis in R (R Core Team, 2017) and made the code available as supplementary material. We used the standard R function `cor()` to compute the tie-corrected version of τ (namely τ_b), and we implemented the

Table 1
Comparison of the estimators' performance under various simulation scenarios and sample size $N = 150$. The reported MSE* is the standard MSE multiplied by a factor of 10^2, calculated based on the true value of τ computed according to Nikoloulopoulos and Karlis (2009).

	$\pi_F = \pi_G$	ρ	True τ	$\hat{\tau}_H$		$\hat{\tau}_A$	
				Mean	MSE*	Mean	MSE*
$\lambda_F = 2,\ \lambda_G = 2$	0.20	0.20	0.07	0.07	0.11	0.06	0.12
		0.50	0.16	0.16	0.17	0.15	0.17
		0.80	0.26	0.25	0.21	0.25	0.21
	0.80	0.20	0.15	0.24	1.16	0.15	0.38
		0.50	0.37	0.46	1.23	0.40	0.48
		0.80	0.62	0.72	1.21	0.69	0.77
$\lambda_F = 2,\ \lambda_G = 8$	0.20	0.20	0.07	0.06	0.12	0.06	0.12
		0.50	0.16	0.16	0.17	0.15	0.18
		0.80	0.26	0.25	0.21	0.25	0.21
	0.80	0.20	0.15	0.20	0.63	0.14	0.40
		0.50	0.36	0.42	0.65	0.38	0.41
		0.80	0.61	0.67	0.57	0.65	0.40
$\lambda_F = 8,\ \lambda_G = 8$	0.20	0.20	0.08	0.07	0.13	0.07	0.14
		0.50	0.18	0.18	0.19	0.17	0.19
		0.80	0.29	0.28	0.23	0.28	0.23
	0.80	0.20	0.16	0.18	0.47	0.15	0.41
		0.50	0.40	0.44	0.50	0.41	0.42
		0.80	0.69	0.73	0.39	0.72	0.33

estimators of τ_A and τ_H, which was also not available. In this regard, the meaning of the variable X_{10} and Y_{10} (respectively X_{11} and Y_{11}) in Pimentel et al. (2015) are subject to interpretation. In particular, in Pimentel et al. (2015), X_{10} is defined as a variable with a conditional distribution of X given that $Y = 0$. Though, based on the steps of our proof and the notation chosen by the authors, we believe that X_{10} should be a *positive* random variable, i.e., only the continuous part of X, with a conditional distribution X given that $Y = 0$ to result in a correct formulation for Eq. (1). We implemented both interpretations for X_{10} and selected the one that performs the best (which in fact corresponds to X_{10} being a positive variable). We recall that, for a bivariate (zero-inflated) discrete distribution, we can calculate the true value of Kendall's τ as given in Nikoloulopoulos and Karlis (2009). Therefore, we compute the mean square error (MSE) of the estimators of τ, τ_H, and τ_A based on the true value of τ, and identify the best one through the smallest MSE. Table 1 shows the results of our simulation study for our chosen parameter values and samples of size $N = 150$. We investigated various sample sizes, i.e., $N = 150, 300, 1000$, and did not find significant differences in the behavior of the estimators. We also tested $N = 50{,}100$, and we noticed that such small sample sizes resulted in a lower number of admissible repetitions due to a higher probability of generating samples with less than two untied couples away from zero. Thus, we only present the case $N = 150$ in the paper. Given the poor performances of τ_b in all the considered scenarios, we decided not to report it in the table. As expected, the estimators of τ_H and τ_A are comparable in their performance when (1) most of the probability mass is in zero for both variables, i.e., for $\pi_F = \pi_G = 0.2$, and (2) there is a limited proportion of ties within the margins away from zero, i.e., $(\lambda_F, \lambda_G) = (8, 8)$. In the other cases, our proposed estimator $\hat{\tau}_A$ outperforms the adjusted version of $\hat{\tau}_H$. A visual representation of the performance of the estimators for selected parameter settings is presented in Fig. 1. From the boxplots of Fig. 1, we can conclude that $\hat{\tau}_A$ is generally close to the true value of τ (the constant horizontal line in the plots), while $\hat{\tau}_H$ tends to overestimate it when the zero-inflation is mild (e.g., $\pi_F = \pi_G = 0.8$). Collectively, our analysis demonstrates that replacing τ_{11} in Eq. (1) by τ_b is not enough to ensure accurate performances, and our adjustment accounting for the probability of ties within the margins is needed when dealing with zero-inflated count data. Besides looking at the performance of the estimators, we also investigated their attainable bounds. We estimated the range of τ_H as suggested in Denuit and Mesfioui (2017) and the bounds of τ_A as described in Section 4. For comparison, we applied Proposition 1 and find the theoretical bounds of τ_A for our specific marginal distributions (Zero-Inflated Poisson). The results are reported in Table 2. The estimated ranges for τ_H and τ_A are very similar to each other and close to the theoretical bounds of τ_A when a small probability mass is spread away from zero. When zero-inflation is limited, our estimated bounds are sharper than the ones derived in Denuit and Mesfioui (2017), as expected. We observe that the bounds get closer to the theoretical ones as the number of ties away from zero decreases (i.e., when λ_F and λ_G increase). This can be explained by noticing that we are not computing exact estimates of the bounds of Proposition 1, as we are using a looser approximation for the attainable bounds of τ_{11}. Nevertheless, such non-parametric estimators of the bounds are close enough to the theoretical bounds, and can certainly be used in practice to interpret the strength of association of an estimate of τ_A without the need to make assumptions on the underlying distributions.

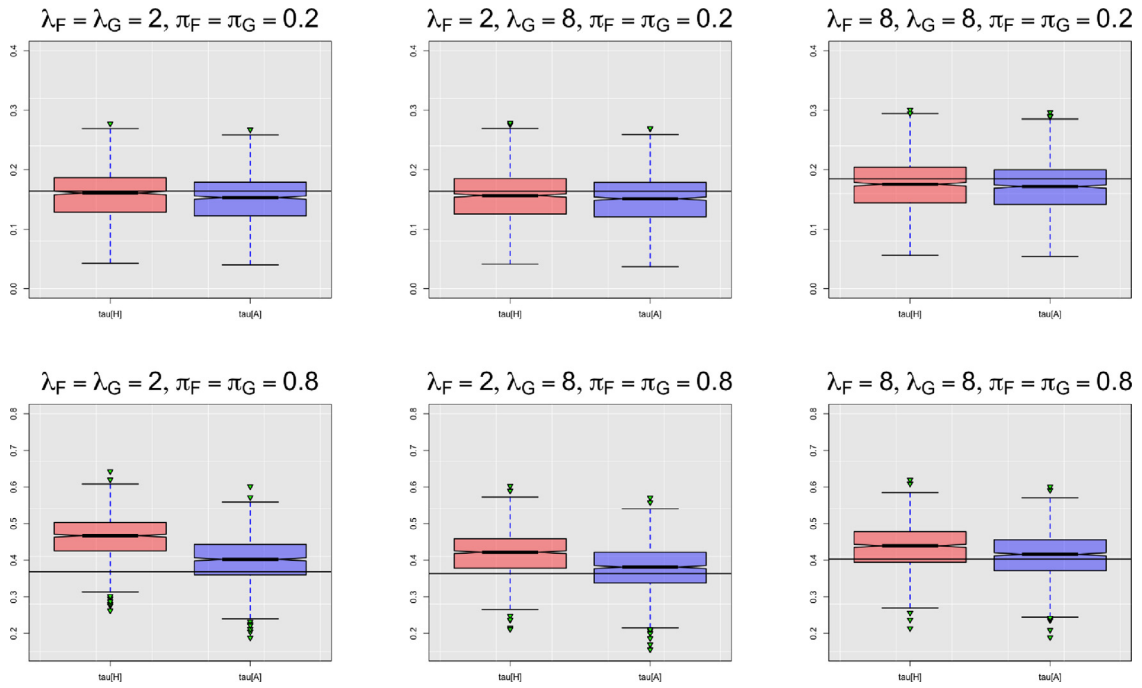


Fig. 1. Boxplots of τ_H (red) and τ_A (blue) over 1000 simulations for six different parameter settings and fixed $\rho = 0.5$. The constant horizontal line in the plots represents the true value of τ.

Table 2
Estimates of the lower and upper bounds of τ_H and τ_A for the simulated scenarios, sample size $N = 150$, and averaged across 1000 runs.

	$\pi_F = \pi_G$	Bounds τ_H	Bounds τ_A	Theoretical bounds τ_A
$\lambda_F = 2,\ \lambda_G = 2$	0.20	[−0.06, 0.29]	[−0.06, 0.29]	[−0.06, 0.31]
	0.80	[−0.81, 0.90]	[−0.76, 0.84]	[−0.75, 0.78]
$\lambda_F = 2,\ \lambda_G = 8$	0.20	[−0.07, 0.32]	[−0.07, 0.32]	[−0.07, 0.31]
	0.80	[−0.86, 0.90]	[−0.82, 0.85]	[−0.80, 0.77]
$\lambda_F = 8,\ \lambda_G = 8$	0.20	[−0.08, 0.35]	[−0.08, 0.35]	[−0.08, 0.36]
	0.80	[−0.92, 0.96]	[−0.89, 0.92]	[−0.87, 0.90]

6. Conclusion

In this paper, we built on previous results presented in Pimentel et al. (2015) by proposing an adjusted estimator of Kendall's τ that can tackle both zero-inflated continuous and count data. We made the proposed estimator interpretable and useful in practice by deriving its theoretical attainable bounds and suggesting a way to estimate them. Our theoretical results were paired with a simulation study, where we analyzed the estimators' performance in various settings based on zero-inflated Poisson distributions. A more extensive simulation study with other discrete distributions can be done by adjusting the simulation code provided as supplementary material or made upon request. Overall, our proposed estimator is more flexible and preferable in practice since it coincides with the estimator proposed by Pimentel et al. (2015) if there are no ties within the margins, while it outperforms it if the zero-inflation is mild.

This paper was motivated by the need for *ad hoc* statistical methods to quantify association between zero-inflated count data. In this work, we mostly focus on point estimates. In a follow-up paper, we plan to derive confidence intervals and coverage probabilities to investigate the convergence rate to normality. We will also apply the proposed estimator to real datasets. Another natural follow-up of the work presented here would be to derive a suitable estimator for Spearman's ρ in case of zero-inflated count data. Preliminary results on the topic for the continuous zero-inflated case have been presented in Pimentel (2009) and Mesfioui and Trufin (2022), while Mesfioui et al. (2022) recently analyzed the attainable bounds of Spearman's ρ when at least one variable is discrete. However, more research and investigations are needed to derive for Spearman's ρ the same tools now available for Kendall's τ.

Data availability

We have attached the R code for the simulation study as a supplementary file.

Acknowledgments

We are grateful to two anonymous reviewers and the editor for their careful reading and thoughtful comments and suggestions on an earlier version of the paper.

Appendix A. Supplementary data

The complete proofs of Theorem 1 and Proposition 1 are available online.
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2023.109858.

References

Arab, A., Holan, S.H., Wikle, C.K., Wildhaber, M.L., 2012. Semiparametric bivariate zero-inflated Poisson models with application to studies of abundance for multiple species. Environmetrics 23 (2), 183–196. http://dx.doi.org/10.1002/env.1142.

Denuit, M., Lambert, P., 2005. Constraints on concordance measures in bivariate discrete data. J. Multivariate Anal. 93 (1), 40–57. http://dx.doi.org/10.1016/j.jmva.2004.01.004.

Denuit, M.M., Mesfioui, M., 2017. Bounds on Kendall's tau for zero-inflated continuous variables. Statist. Probab. Lett. 126, 173–178. http://dx.doi.org/10.1016/j.spl.2017.03.005.

Hollander, M., Wolfe, D., 2013. Nonparametric Statistical Methods, third ed. Wiley, New York.

Kendall, M., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

Kendall, M., 1945. The treatment of ties in ranking problems. Biometrika 33, 239–251.

Mesfioui, M., Tajar, A., 2005. On the properties of some nonparametric concordance measures in the discrete case. J. Nonparametr. Stat. 17 (5), 541–554. http://dx.doi.org/10.1080/10485250500038967.

Mesfioui, M., Trufin, J., 2022. Bounds on multivariate Kendall's tau and Spearman's rho for zero-inflated continuous variables and their application to insurance. Methodol. Comput. Appl. Probab. 24, 1051–1059. http://dx.doi.org/10.1007/s11009-021-09869-3.

Mesfioui, M., Trufin, J., Zuyderhoff, P., 2022. Bounds on Spearman's rho when at least one random variable is discrete. Eur. Actuar. J. 12, 321–348. http://dx.doi.org/10.1007/s13385-021-00289-8.

Moulton, L.H., Halsey, N.A., 1995. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics 51 (4), 1570–1578.

Nelsen, R.B., 2006. An Introduction To Copulas, second ed. In: Springer Series in Statistics, Springer.

Nešlehová, J., 2007. On rank correlation measures for non-continuous random variables. J. Multivariate Anal. 98 (3), 544–567. http://dx.doi.org/10.1016/j.jmva.2005.11.007.

Nikoloulopoulos, A.K., Karlis, D., 2008. Multivariate logit copula model with an application to dental data. Stat. Med. 27 (30), 6393–6406. http://dx.doi.org/10.1002/sim.3449.

Nikoloulopoulos, A.K., Karlis, D., 2009. Modeling multivariate count data using copulas. Comm. Statist. Simulation Comput. 39 (1), 172–187. http://dx.doi.org/10.1080/03610910903391262.

Pimentel, R.S., 2009. Kendall's Tau and Spearman's Rho for Zero-Inflated Data (Ph.D. thesis). Western Michigan University, Michigan.

Pimentel, R.S., Niewiadomska-Bugaj, M., Wang, J.C., 2015. Association of zero-inflated continuous variables. Statist. Probab. Lett. 96, 61–67. http://dx.doi.org/10.1016/j.spl.2014.09.002.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.