ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

International Journal of Statistics and Probability

Vol. 11, No. 6 November 2022

CANADIAN CENTER OF SCIENCE AND EDUCATION

INTERNATIONAL JOURNAL OF STATISTICS AND PROBABILITY

An International Peer-reviewed and Open Access Journal for Statistics and Probability

International Journal of Statistics and Probability (ISSN: 1927-7032; E-ISSN: 1927-7040) is an open-access, international, double-blind peer-reviewed journal published by the Canadian Center of Science and Education. This journal, published **bimonthly** (January, March, May, July, September and November) in both **print and online versions**, keeps readers up-to-date with the latest developments in all areas of statistics and probability.

The scopes of the journal:

- Computational statistics
- Design of experiments
- Sample survey
- Statistical modelling
- Statistical theory
- Probability theory

The journal is included in:

- BASE
- Google Scholar
- JournalTOCs
- LOCKSS
- SHERPA/RoMEO
- Ulrich's

Submission Policy

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the authorities responsible where the work was carried out. However, we accept submissions that have previously appeared on preprint servers (for example: arXiv, bioRxiv, Nature Precedings, Philica, Social Science Research Network, and Vixra); have previously been presented at conferences; or have previously appeared in other "non-journal" venues (for example: blogs or posters). Authors are responsible for updating the archived preprint with the journal reference (including DOI) and a link to the published articles on the appropriate journal website upon publication.

iThenticate Professional Plagiarism Prevention

The publisher and journals have a zero-tolerance plagiarism policy. We check the issue using two methods: a plagiarism prevention tool (iThenticate) and a reviewer check. All submissions will be checked by iThenticate before being sent to reviewers.

creative commons

We insist a rigorous viewpoint on the self-plagiarism. The self-plagiarism is plagiarism, as it fails to contribute to the research and science.

IJSP accepts both Online and Email submission. The online system makes readers to submit and track the status of their manuscripts conveniently. For any questions, please contact ijsp@ccsenet.org.

International Journal of Statistics and Probability; Vol. 11 No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Editorial Team

International Journal of Statistics and Probability; Vol. 11 No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Contents

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Correlation-Preserving Mean Plausible Values as a Basis for Prediction in the Context of Bayesian Structural Equation Modeling

André Beauducel[1] & Norbert Hilger[1]

[1] University of Bonn, Institute of Psychology, Bonn, Germany

Correspondence: André Beauducel, Institute of Psychology, Kaiser-Karl-Ring 9, 53111 Bonn, Germany

Received: August 28, 2022 Accepted: October 8, 2022 Online Published: October 20, 2022

doi:10.5539/ijsp.v11n6p1 URL: https://doi.org/10.5539/ijsp.v11n6p1

Abstract

Mean plausible values can be computed when Bayesian structural equation modeling (BSEM) is performed. As mean plausible values do not preserve the factor inter-correlations, they yield path coefficients that are different from the estimated path coefficients of the model. As it might be of interest to perform exactly the same predictions on the level of mean plausible values that have been estimated by BSEM, correlation-preserving mean plausible values were proposed. An example for the computation of the correlation preserving mean plausible values is given and the corresponding syntax can be found in the Appendix.

Keywords: Bayesian structural equation modeling, plausible values, factor scores, prediction

1. Introduction

1.1 The Validity of Mean Plausible Values

Bayesian structural equation modeling (BSEM) has been proposed by Muthén and Asparouhov (2012) as an alternative to conventional structural equation modeling and several improvements of BSEM have meanwhile been proposed and realized (Asparouhov & Muthén, 2021; Asparouhov, Muthén, & Morin, 2015; Zitzmann & Hecht, 2019). Advantages of BSEM are that it can be performed with relatively small samples (Bonafede, Chiorri, Azzolina, 2021) and that priors for the variability of loadings can be specified. BSEM thereby allows to overcome problems of specifying fixed zero loadings in the independent clusters model (Beauducel & Hilger, 2020) and it also allows for the specification of complex loading patterns as, for example, circumplex models (Weide, Scheuble, & Beauducel, 2021).

As BSEM becomes more and more popular, the interest for score estimates of the latent variables or factors in these models may also increase. Asparouhov and Muthén (2010a, b) proposed mean plausible values as factor score estimates in the context of BSEM. Luo and Dimitrov (2018) found that even less than 500 imputations may be used in order to get mean plausible values with an appropriate validity. Moreover, Beauducel and Hilger (in press) have shown that mean plausible values of the exogenous factors $\mathbf{P}_\xi$ based on 500 imputations have nearly the same coefficient of determinacy as the best linear factor score estimate initially proposed by Thurstone (1935), which is also termed regression factor score $\mathbf{F}_\xi^R$. As $\mathbf{F}_\xi^R$ has the maximum determinacy, the result of Beauducel and Hilger (in press) implies that $\mathbf{P}_\xi$ based on more than 500 imputations is a proxy of $\mathbf{F}_\xi^R$.

Using $\mathbf{P}_\xi$ in the context of BSEM could be especially interesting in the context of the prediction of endogenous factors by exogenous factors. In applied settings, individuals might be selected according to their individual scores on exogenous factors (predictors). The selection of individuals according to their scores requires that the scores are valid indicators of the latent predictors which implies that they represent the underlying prediction model quite well. However, Skrondal and Laake (2001) have shown that using $\mathbf{F}_\xi^R$ for exogeneous factors $\boldsymbol{\xi}$ and $\mathbf{F}_\eta^R$ for endogenous factors $\boldsymbol{\eta}$ yields path coefficients that do not correspond to the path coefficients estimated by means of structural equation modeling. As $\mathbf{P}_\xi$ and $\mathbf{P}_\eta$ based on a large number of imputations are proxies of $\mathbf{F}_\xi^R$ and $\mathbf{F}_\eta^R$, it is expected that $\mathbf{P}_\xi$ and $\mathbf{P}_\eta$ yield biased path coefficients and factor inter-correlations (i.e., that the coefficients do not correspond to the coefficients obtained by means of BSEM). It could, however, be of interest to compute factor score estimates that allow for exactly the same predictions as the corresponding BSEM.

1.2 Aims of the Study

The aim of the present study was therefore to provide a method that allows to transform $\mathbf{P}_\xi$ into scores resulting in path coefficients corresponding exactly to the path coefficients estimated by means of BSEM. After some definitions we provide the transformations for the correlation preserving mean plausible values and for the computation of their

determinacy. Then, we provide an example based on a simulated data set to show the difference between conventional and correlation-preserving mean plausible values. The syntax for the transformation is given in the Appendix.

2. Method

2.1 Definitions

We use the notation of Skrondal and Laake (2001), Jöreskog (1977), and Jöreskog and Sörbom (1989) with the latent regression model

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{\Gamma}$ is the matrix of path coefficients for the prediction of the endogenous factors $\boldsymbol{\eta}$ by the exogeneous factors $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are the residuals, with $E(\boldsymbol{\zeta}\boldsymbol{\xi}') = \mathbf{0}$. The measurement model for the exogenous factors is

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2}$$

where $\mathbf{x}$ represents the observed variables of the exogenous factors, $\boldsymbol{\Lambda}_x$ is the matrix of factor loadings, and $\boldsymbol{\delta}$ are the unique factors, with $E(\boldsymbol{\delta}\boldsymbol{\delta}') = diag(E(\boldsymbol{\delta}\boldsymbol{\delta}')) = \boldsymbol{\Theta}_\delta$, $E(\boldsymbol{\delta}\boldsymbol{\xi}') = \mathbf{0}$, and $E(\boldsymbol{\xi}\boldsymbol{\xi}') = \boldsymbol{\Phi}$, with $diag(E(\boldsymbol{\Phi})) = \mathbf{I}$, so that

$$E(\mathbf{x}\mathbf{x}') = \boldsymbol{\Sigma}_x = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta. \tag{3}$$

The measurement model for the endogenous factors is

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\mathbf{y}$ represents the observed variables of the exogenous variable, $\boldsymbol{\Lambda}_y$ is the matrix of factor loadings, and $\boldsymbol{\varepsilon}$ are the unique factors, with $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = diag(E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) = \boldsymbol{\Theta}_\varepsilon$, $E(\boldsymbol{\delta}\boldsymbol{\xi}') = \mathbf{0}$, and $E(\boldsymbol{\eta}\boldsymbol{\eta}') = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$, with $diag(E(\boldsymbol{\eta}\boldsymbol{\eta}')) = \mathbf{I}$, so that

$$E(\mathbf{y}\mathbf{y}') = \boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}_y(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon. \tag{5}$$

2.2 Correlation-Preserving Mean Plausible Values

The combined matrix of all factors is $\mathbf{F} = \begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{bmatrix}$, so that

$$\mathbf{C} = E(\mathbf{F}\mathbf{F}') = E\left(\begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{bmatrix}\begin{bmatrix}\boldsymbol{\xi}' \vdots \boldsymbol{\eta}'\end{bmatrix}\right) = E\left(\begin{bmatrix}\boldsymbol{\xi}\boldsymbol{\xi}' & \boldsymbol{\xi}\boldsymbol{\eta}'\\ \boldsymbol{\eta}\boldsymbol{\xi}' & \boldsymbol{\eta}\boldsymbol{\eta}'\end{bmatrix}\right). \tag{6}$$

As all factors have unit variance, $\mathbf{C}$ is a correlation matrix. According to Equations 1 to 5 the elements of $\mathbf{C}$ can be computed from the model parameter estimates

$$\mathbf{C} = \begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\Gamma}'\\ \boldsymbol{\Gamma}\boldsymbol{\Phi} & \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\end{bmatrix}. \tag{7}$$

The regression factor score $\mathbf{F}_\xi^R$ does not preserve the correlations in $\boldsymbol{\Phi}$ which follows from the covariances $E(\mathbf{F}_\xi^R\mathbf{F}_\xi^{R'}) = \boldsymbol{\Phi}\boldsymbol{\Lambda}_x'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Phi}$ (Skrondal & Laake, 2001, Eq. 9), so that the inter-correlation of the regression factor scores is

$$\mathbf{C}_{\mathbf{F}_\xi^R\mathbf{F}_\xi^{R'}} = diag(\boldsymbol{\Phi}\boldsymbol{\Lambda}_x'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Phi}\,diag(\boldsymbol{\Phi}\boldsymbol{\Lambda}_x'\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Phi})^{-1/2}. \tag{8}$$

Inserting Equation 8 instead of $\boldsymbol{\Phi}$ into Equation 7 results in biased estimates for $E(\boldsymbol{\eta}\boldsymbol{\eta}')$ and $E(\boldsymbol{\eta}\boldsymbol{\xi}')$. As $\mathbf{P}_\xi$ and $\mathbf{P}_\eta$ based on a large number of imputations are proxies of $\mathbf{F}_\xi^R$ and $\mathbf{F}_\eta^R$, it follows that intercorrelations and path coefficients that are based on $\mathbf{P}_\xi$ and $\mathbf{P}_\eta$ will be biased.

Let $\mathbf{P} = \begin{bmatrix}diag(\mathbf{P}_\eta\mathbf{P}_\eta')^{-1/2}\mathbf{P}_\eta\\ diag(\mathbf{P}_\xi\mathbf{P}_\xi')^{-1/2}\mathbf{P}_\xi\end{bmatrix}$ and $\mathbf{C}_\mathbf{P} = E(\mathbf{P}\mathbf{P}')$ so that mean plausible values preserving the correlations in $\mathbf{C}$ can

be defined as

$$\mathbf{P}_\mathbf{C} = \mathbf{C}^{1/2}\mathbf{C}_\mathbf{P}^{-1/2}\mathbf{P}, \tag{9}$$

where "$^{1/2}$" denotes the symmetric square-root, $\mathbf{P}_\mathbf{C} = \begin{bmatrix}\mathbf{P}_{C\eta}\\ \mathbf{P}_{C\xi}\end{bmatrix}$ and $E(\mathbf{P}_\mathbf{C}\mathbf{P}_\mathbf{C}') = E(\mathbf{C}^{1/2}\mathbf{C}_\mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}'\mathbf{C}_\mathbf{P}^{-1/2}\mathbf{C}^{1/2})$

$= \mathbf{C}^{1/2}\mathbf{C}_{\mathbf{P}}^{-1/2}\mathbf{C}_{\mathbf{P}}\mathbf{C}_{\mathbf{P}}^{-1/2}\mathbf{C}^{1/2} = \mathbf{C}.$

If the mean plausible values approximate the regression factor score, Equation 9 can be computed directly from the model parameters. For convenience, this is only illustrated for the exogenous factors, although it also holds for the endogenous factors.

For exogenous factors Equation 9 can be written as

$$\mathbf{P}_{C\xi} = \mathbf{\Phi}^{1/2}\mathbf{C}_{P\xi}^{-1/2}\,diag(\mathbf{P}_{\xi}\mathbf{P}_{\xi}')^{-1/2}\,\mathbf{P}_{\xi}. \tag{10}$$

For $\mathbf{P}_{\xi} = \mathbf{F}_{\xi}^{R}$ it is possible to insert the right hand side of $diag(\mathbf{F}_{\xi}^{R}\mathbf{F}_{\xi}^{R'})^{-1/2}\mathbf{F}_{\xi}^{R} = diag(\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi})^{-1/2}\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}$ for $diag(\mathbf{P}_{\xi}\mathbf{P}_{\xi}')^{-1/2}\,\mathbf{P}_{\xi}$ and the right hand side of Equation 8 for $\mathbf{C}_{P\xi}^{-1/2}$ in Equation 10. This yields

$$\mathbf{P}_{C\xi} = \mathbf{\Phi}^{1/2}\,(diag(\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi})^{-1/2}\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi}\,diag(\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi})^{-1/2})^{-1/2}\,diag(\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi})^{-1/2}\mathbf{\Phi}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{x}, \tag{11}$$

so that no mean plausible values but model parameters and the measured variables are needed to compute the correlation preserving plausible values. For $\mathbf{\Phi} = \mathbf{I}$ Equation 11 can be transformed to

$$\begin{aligned}\mathbf{P}_{C\xi} &= diag(\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{1/4}(\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/2}\,diag(\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/4}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{x}\\ &= (\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/2}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{x},\end{aligned} \tag{12}$$

with $E(\mathbf{P}_{C\xi}\mathbf{P}_{C\xi}') = (\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/2}\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\,(\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/2} = \mathbf{I}.$ This is the orthogonal factor score proposed by Takeuchi, Yanai, and Mukherjee (1982). As this score is already standardized, the correlation-preserving score for $\mathbf{\Phi} \neq \mathbf{I}$ can simply be computed by a pre-multiplication of Takeuchi et al.'s factor score with $\mathbf{\Phi}^{1/2}$, so that

$$\mathbf{P}_{C\xi}^{*} = \mathbf{\Phi}^{1/2}(\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x})^{-1/2}\,\mathbf{\Lambda}_{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{x}. \tag{13}$$

Note that Equations 11 and 13 describe the relationship between mean plausible values and correlation-preserving scores for $\mathbf{P}_{\xi} = \mathbf{F}_{\xi}^{R}$, which depends on the number of imputations. For a small number of imputations $\mathbf{P}_{\xi} \neq \mathbf{F}_{\xi}^{R}$ so that $\mathbf{\Phi}^{-1/2}\mathbf{P}_{C\xi}^{*} \neq \mathbf{P}_{C\xi}$.

However, for any factor score and for any mean plausible value the determinacy should be computed. According to Beauducel and Hilger (in press, Eq. 7) the determinacy of the mean plausible values $\mathbf{P}_{C\xi}$ for exogenous factors can be estimated by means of

$$\mathbf{D}_{C\xi} = diag(\mathbf{P}_{C\xi}\mathbf{P}_{C\xi}')^{-1/2}\,\mathbf{P}_{C\xi}\mathbf{x}'\mathbf{\Sigma}_{x}^{-1}\mathbf{\Lambda}_{x}\mathbf{\Phi}, \tag{14}$$

and the determinacy of mean plausible values for endogenous factors can be estimated by means of

$$\mathbf{D}_{C\eta} = diag(\mathbf{P}_{C\eta}\mathbf{P}_{C\eta}')^{-1/2}\,\mathbf{P}_{C\eta}\mathbf{y}'\mathbf{\Sigma}_{y}^{-1}\mathbf{\Lambda}_{y}(\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}' + \mathbf{\Psi}). \tag{15}$$

3. Example

A simulated data set containing n = 10,000 cases, 15 normally distributed N(0,1) observed variables ($\mathbf{x}$) as a measurement model of three exogenous factors $\boldsymbol{\xi}$ and 10 normally distributed N(0,1) observed variables ($\mathbf{y}$) as a measurement model for two endogenous factors $\boldsymbol{\eta}$ were generated with IBM SPSS Version 26. The data file (csv) can be found in the supplement. BSEM was performed with Mplus 8.4 (Muthén & Muthén, 2019) in order to estimate the model parameters of the conceptual model presented in Figure 1 (Mplus syntax-file in Supplements).

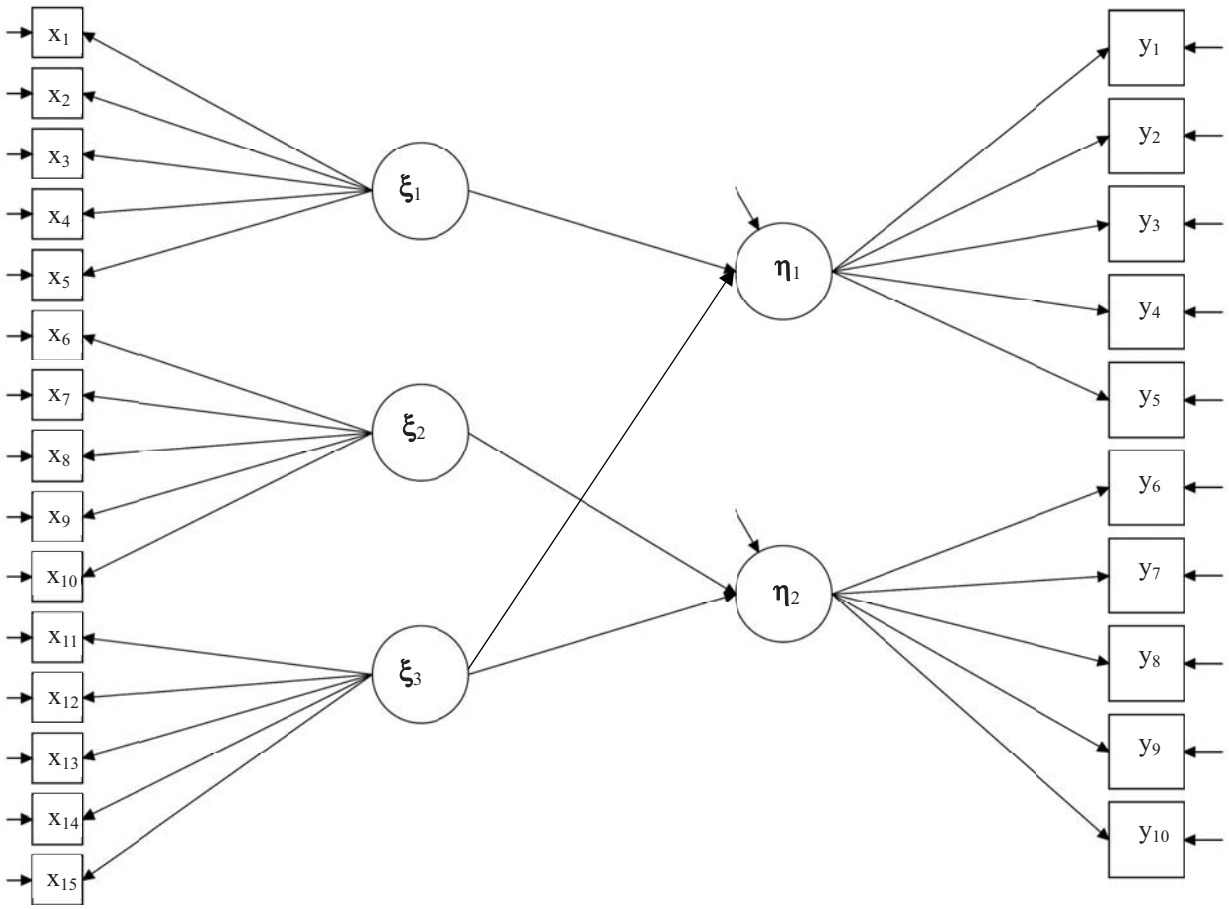

Figure 1. Conceptual model with three exogenous and two endogenous factors. The Bayesian model parameters are given in Table 1

Factor variances were fixed to one and for each factor five salient loadings were freely estimated and non-salient loadings were estimated with normally distributed priors with a zero mean and a variance of σ^2=0.01. As no model-misfit was simulated, the model fit was excellent (95th confidence-interval for difference between observed and replicated χ^2 [1670.88, 1798.32], posterior predictive p-value < 0.001, prior posterior predictive p-value < .001, RMSEA = 0.027, CFI = 0.991). The model parameter estimates are given in Table 1 (data file, Mplus input and output-file can be found in Supplements).

Table 1. BSEM Model parameter estimates (completely standardized solution)

x		**Λ_x**		**y**	**Λ_y**	
x_1	**0.750**	0.066	0.025	y_1	0.160	0.251
x_2	**0.845**	0.049	0.002	y_2	0.160	0.251
x_3	**0.938**	0.031	-0.021	y_3	**0.999**	-0.041
x_4	**0.845**	0.049	0.002	y_4	**0.999**	-0.041
x_5	**0.845**	0.049	0.002	y_5	**0.999**	-0.041
x_6	0.031	**0.762**	0.023	y_6	-0.038	**0.534**
x_7	0.008	**0.858**	0.001	y_7	-0.038	**0.534**
x_8	-0.015	**0.953**	-0.022	y_8	-0.037	**0.533**
x_9	0.008	**0.859**	0.000	y_9	-0.038	**0.533**
x_{10}	0.008	**0.858**	0.001	y_{10}	-0.038	**0.534**
x_{11}	0.064	0.027	**0.749**		$E(\eta\eta')$	
x_{12}	0.046	0.002	**0.846**		1.000	**0.513**
x_{13}	0.029	-0.024	**0.942**		**0.513**	1.000
x_{14}	0.047	0.002	**0.846**			
x_{15}	0.047	0.002	**0.846**		**Γ**	
	Φ				η_1	η_1
	1.000	0.275	0.270	ξ_1	0.270	0.000
	0.275	1.000	0.324	ξ_2	0.000	0.037
	0.270	0.324	1.000	ξ_3	0.016	**0.447**

Note. Model parameters greater .40 are given in bold face

The path coefficients for the prediction of the mean plausible values of the endogenous factors by the mean plausible values of the latent exogenous factors are given in Table 2, together with the path coefficients for the correlation-preserving mean plausible values. The SPSS Syntax for the computation of the mean plausible values and the corresponding regression analyses are given in the Appendix.

Table 2. Standardized path coefficients (beta) based on mean plausible values and on the basis of correlation-preserving mean plausible values

	$P_{\eta 1}$	$P_{\eta 2}$		$P_{C\eta 1}$	$P_{C\eta 2}$
$P_{\xi 1}$	0.275	-0.038	$P_{C\xi 1}$	0.270	0.000
$P_{\xi 2}$	-0.079	0.005	$P_{C\xi 2}$	0.000	0.037
$P_{\xi 3}$	0.053	0.549	$P_{C\xi 3}$	0.016	0.447

Only for the correlation-preserving mean plausible values the path coefficients are identical to the path coefficients of the model (Table 1). The coefficients of determinacy are given in Table 3. They are very similar for the mean plausible values and for the correlation-preserving mean plausible values.

Table 3. Coefficients of determinacy for mean plausible values and for correlation-preserving mean plausible values

$P_{\xi 1}$	$P_{\xi 2}$	$P_{\xi 3}$	$P_{\eta 1}$	$P_{\eta 2}$
.97	.97	.97	.97	.85
$P_{C\xi 1}$	$P_{C\xi 2}$	$P_{C\xi 3}$	$P_{C\eta 1}$	$P_{C\eta 2}$
.97	.97	.97	.99	.82

4. Discussion

As mean plausible values based on 500 or more imputations are a proxy of the regression factor score, which does not preserve the inter-correlations of the factors (Skrondal & Laake, 2001), it was concluded that mean plausible values are not correlation-preserving. It might, however, be of interest in the context of BSEM to compute correlation-preserving mean plausible values with the same inter-correlations as the factors. Only correlation-preserving mean-plausible values will result in the same path coefficients from exogenous factors to endogenous factors as the model estimates. Therefore, correlation-preserving mean plausible values were proposed. An example demonstrates how correlation-preserving mean plausible values can be computed from mean plausible values. It is also shown that only the correlation-preserving mean plausible values yield the path coefficients that estimated by BSEM.

Factor score determinacies of mean plausible values and correlation-preserving mean plausible values were similar. It should, however, be noted that the determinacy of correlation-preserving mean plausible values will typically be smaller than the determinacy of the mean plausible values. This follows from the mean plausible value being a proxy of the regression factor score, which has the largest possible determinacy in a given data set. However, as one can see in the example, the prediction of the endogenous factors is different for the mean plausible values and for the correlation-preserving mean plausible values. This may result in a larger determinacy of the correlation-preserving mean plausible values for the endogenous factors. Further research may explore the conditions for the higher determinacy of correlation-preserving mean plausible values for endogenous factors systematically.

When mean plausible values are equivalent to the regression factor score, the correlation-preserving mean plausible values are equivalent to a correlation-preserving version of Takeuchi et al.'s (1982) orthogonal factor score. As Takeuchi et al.'s factor score has been shown to be identical to Anderson-Rubin's factor score (Anderson & Rubin, 1956; Beauducel, 2015), this also implies that McDonald's (1981) correlation-preserving factor score, will be equivalent to the correlation-preserving mean plausible values under this condition. When equivalence of mean plausible values and regression factor scores is obtained, mean plausible values need not to be computed and the scores can directly be computed from the model parameters and the observed variables. A limitation of the present study is the focus on the mean plausible values, whereas the median of the plausible values might also be of interest. Whether a correlation-preserving version of the median plausible value might be of interest, especially with small samples, might be explored in further studies.

In order to compute the correlation-preserving mean plausible values, the inter-factor correlations estimated by means of BSEM and the mean plausible values should be entered into the example syntax provided in the Appendix. Therefore,

the procedure proposed here can also be applied to mean plausible values that are based on a small number of imputations. When the determinacies are to be computed, which is recommended, the loadings and inter-correlations of the exogenous factors, the loadings and inter-correlations of endogenous factors can also be inserted into the example syntax.

Acknowledgments

This study was funded by the German Research Foundation (DFG), BE 2443/18-1.

References

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. Proc. Third. Berkeley. *Symposium of Mathematical Statistics and Probability, 5,* 111–150.

Asparouhov, T., & Muthén B. (2010a). *Plausible values for latent variables using Mplus.* Technical Report. Retrieved from www.statmodel.com/download/Plausible.pdf

Asparouhov, T., & Muthén B. (2010b). *Multiple imputation with Mplus.* Technical Report. Retrieved from www.statmodel.com/download/Imputations7.pdf

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management, 41*, 1561-1577. Beauducel, A., & Hilger, N. (2020). Overcoming limitations of the independent clusters model for CFA by means of Bayes-estimation and buffered simple structure. *International Journal of Statistics and Probability, 9,* 62-77. https://doi.org/10.1177/0149206315591075

Beauducel, A. (2015). A note on equality and inequality of some factor score predictors. *Communications in Statistics—Theory and Methods, 44*, 3653–3657. https://doi.org/10.1080/03610926.2013.839040

Beauducel, A., & Hilger, N. (in press). Coefficients of factor score determinacy for mean plausible values of Bayesian factor analysis. *Educational and Psychological Measurement*, 2022. https://doi.org/10.1177/00131644221078960

Bonafede, M., Chiorri, C., & Azzolina, D., et al. (2021). Preliminary validation of a questionnaire assessing psychological distress in caregivers of patients with malignant mesothelioma: Mesothelioma Psychological Distress Tool-Caregivers. *Psycho-Oncology*, 1-8. https://doi.org/10.1002/pon.5789

Jöreskog, K. G. (1977). *Structural equation models in the social sciences: Specification, estimation, and testing.* In P. R. Krishnaiah (Ed.), Applications of Statistics (pp. 265-286). Amsterdam: North-Holland.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL7: A guide to the program and applications.* Chicago, IL: SPSS.

Luo, Y., & Dimitrov, D. M. (2018). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: How many plausible values are needed? *Educational and Psychological Measurement, 79*, 272-287. https://doi.org/10.1177/0013164418777569

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika, 46*, 337-341. https://doi.org/10.1007/BF02293740

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313-335. https://doi.org/10.1037/a0026802

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66,* 563-576. https://doi.org/10.1007/BF02296196Takeuchi, K., Yanai, H., & Mukherjee, B. N. (1982). *The Foundations of Multivariate Analysis.* New Delhi: Wiley Eastern.

Thurstone, L. L. (1935). *The Vectors of mind.* Chicago, IL: University of Chicago Press.

Weide, A. C., Scheuble, V., & Beauducel, A. (2021). Bayesian and maximum-likelihood modeling and higher-level scores of interpersonal problems with circumplex structure. *Frontiers in Psychology. Quantitative Psychology and Measurement*, *12,* 761378. https://doi.org/10.3389/fpsyg.2021.761378

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling, 26*, 646-661. https://doi.org/10.1080/10705511.2018.1545232

Appendix A

IBM SPSS-Syntax.

```
* Encoding: UTF-8.

* Syntax for the example presented in the manuscript.
* For use in other contexts, enter location and name of the data-file containing mean
  plausible values and adapt the variable number according to your model.

DATA LIST FILE="C:\Example_data_plausible_values.dat" fixed records=1
/1 x1 to x15 (15F6.3)   y1 to y10 (10F6.3)
Ksi1_meanPlausible (F6.3)  Ksi1_medianPlausible (F6.3) Ksi1_SD (F6.3) Ksi1_perc2p5 (F6.3) Ksi1_perc97p5 (F6.3)
Ksi2_meanPlausible (F6.3)  Ksi2_medianPlausible (F6.3) Ksi2_SD (F6.3) Ksi2_perc2p5 (F6.3) Ksi2_perc97p5 (F6.3)
Ksi3_meanPlausible (F6.3)  Ksi3_medianPlausible (F6.3) Ksi3_SD (F6.3) Ksi3_perc2p5 (F6.3) Ksi3_perc97p5 (F6.3)
Eta1_meanPlausible (F6.3)  Eta1_medianPlausible (F6.3) Eta1_SD (F6.3) Eta1_perc2p5 (F6.3) Eta1_perc97p5 (F6.3)
Eta2_meanPlausible (F6.3)  Eta2_medianPlausible (F6.3) Eta2_SD (F6.3) Eta2_perc2p5 (F6.3) Eta2_perc97p5 (F6.3).
Dataset name dataset1.
save outfile="C:\Example_data_plausible_values.sav".


MATRIX.

get P_Ksi/variables= Ksi1_meanPlausible Ksi2_meanPlausible  Ksi3_meanPlausible
/file='C:\Example_data_plausible_values.sav'.
get P_Eta/variables= Eta1_meanPlausible Eta2_meanPlausible
/file='C:\Example_data_plausible_values.sav'.

get x/variables=  x1 to x15
/file='C:\Example_data_plausible_values.sav'.
get y/variables=  y1 to y10
/file='C:\Example_data_plausible_values.sav'.


* In the following matrices are the values that are also given in Table 1.
* For use in other contexts, enter the corresponding values from your BSEM-OUTPUT:.

* Loadings of measured variables on Ksi.
compute Lx={
0.750, 0.066, 0.025;
0.845, 0.049, 0.002;
0.938, 0.031,-0.021;
0.845, 0.049, 0.002;
0.845, 0.049, 0.002;
```

```
0.031, 0.762, 0.023;
0.008, 0.858, 0.001;
-0.015, 0.953, -0.022;
0.008, 0.859, 0.000;
0.008, 0.858, 0.001;
0.064, 0.027, 0.749;
0.046, 0.002, 0.846;
0.029,-0.024, 0.942;
0.047, 0.002, 0.846;
0.047, 0.002, 0.846}.

* Intercorrelations of Ksi.
compute Phi={
 1.000, 0.275, 0.270;
 0.275, 1.000, 0.324;
 0.270, 0.324, 1.000
}.

* Loadings of measured variables on Eta.
compute Ly={
 0.160, 0.251;
 0.160, 0.251;
 0.999,-0.041;
 0.999,-0.041;
 0.999,-0.041;
-0.038, 0.534;
-0.038, 0.534;
-0.037, 0.533;
-0.038, 0.533;
-0.038, 0.534
}.

* Path coefficients from Ksi to Eta.
compute Gamma={
0.270,  0.000;
0.000,  0.037;
0.016,  0.447
}.

* Intercorrelations of Eta.
compute Ceta = {
1.000, 0.513;
0.513, 1.000
```

```
}.

* Computations.
compute P_Ksi=t(P_Ksi).
compute ncases=ncol(P_ksi).
* Mean-centering of P_Ksi.
compute mP=RSUM(P_Ksi)&/ncases.
compute ones=make(nrow(P_Ksi),ncol(P_Ksi),1).
compute mmP=Mdiag(mP)*ones.
compute P_Ksi=P_Ksi-mmP.

compute P_Eta=t(P_Eta).
* Mean-centering of P_Eta.
compute mP=RSUM(P_Eta)&/ncases.
compute ones=make(nrow(P_Eta),ncol(P_Eta),1).
compute mmP=Mdiag(mP)*ones.
compute P_Eta=P_Eta-mmP.

compute P={P_Ksi;P_Eta}.

compute x=t(x).
compute y=t(y).



compute C_P=INV(Mdiag(diag( P*t(P)&/(ncases-1) ))&**0.5) * P*t(P)&/(ncases-1)
* INV(Mdiag(diag( P*t(P)&/(ncases-1) ))&**0.5).
print C_P/format=F5.2/Title="Correlation of mean plausible values".

CALL Eigen(C_P, vec, eig).
compute C_P12=vec*Mdiag(eig)&**0.5*t(vec).

compute Gamma=t(Gamma).
compute Cetaksi=(Gamma)*Phi.
compute tcetaks=t(Cetaksi).

compute C={
Phi, tCetaks;
Cetaksi, Ceta }.

Print C/format=F5.2/Title="Correlation of factors according to the model parameters of BSEM".

CALL Eigen(C, vec, eig).
compute C12=vec*Mdiag(abs(eig))&**0.5*t(vec).
```

```
* Compute correlation-preserving plausible values according to Equation 10.
compute Pc=C12*INV(C_P12)*INV(Mdiag(diag( P*t(P)&/(ncases-1) ))&**0.5) *P.

print {INV(Mdiag(diag( Pc*t(Pc) )&/(ncases-1))&**0.5)*Pc*t(Pc)&/(ncases-1)
*INV(Mdiag(diag( Pc*t(Pc) )&/(ncases-1))&**0.5)}
/format=F5.2/Title="Check: Correlation of correlation-preserving mean plausible values. Should be equal to correlation
of factors".


* Determinacy.

compute Tdelta=Mdiag(diag( 1-Lx*Phi*t(Lx) )).
compute Sig_x=Lx*Phi*t(Lx) + Tdelta.

compute Tepsi=Mdiag(diag(1 - Ly*Ceta*t(Ly))).
compute Sig_y=Ly*Ceta*t(Ly) + Tepsi.


* Compute Determinacy of mean plausible values for Ksi.
compute D_Ksi=INV(Mdiag(diag( P_Ksi*t(P_Ksi)&/(ncases-1) )))&**0.5 * P_Ksi
* t(x)&/(ncases-1) * INV(Sig_x)*Lx*Phi.
print {t(diag(D_Ksi))} /format=F5.2/Title="Determinacy of mean plausible values for Ksi".

* Compute Determinacy of mean plausible values for Eta.
compute D_Eta=INV(Mdiag(diag( P_Eta*t(P_Eta)&/(ncases-1) ))) * P_Eta
*T(y)&/(ncases-1) * INV(Sig_y)*Ly*Ceta.
print {t(diag(D_Eta))} /format=F5.2/Title="Determinacy of mean plausible values for Eta".

* Compute Determinacy of correlation-preserving mean plausible values for Ksi (according Equation 14).
compute PcKsi = {Pc(1,:);Pc(2,:);Pc(3,:)}.
compute D_cKsi=INV(Mdiag(diag( PcKsi*t(PcKsi)&/(ncases-1) )))&**0.5 *PcKsi
*T(x)&/(ncases-1)*INV(Sig_x)*Lx*Phi.
print {t(diag(D_cKsi))} /format=F5.2/Title="Determinacy of correlation-preserving mean plausible values for Ksi
(according to Equation 12)".

* Compute Determinacy of correlation-preserving mean plausible values for Eta (according Equation 15).
compute PcEta = {Pc(4,:);Pc(5,:)}.
compute D_cEta=INV(Mdiag(diag( PcEta*t(PcEta)&/(ncases-1) ))) * PcEta
*T(y)&/(ncases-1)*INV(Sig_y)*Ly*Ceta.
print {t(diag(D_cEta))} /format=F5.2/Title="Determinacy of correlation-preserving mean plausible values for Eta
(according to Equation 13)".
```

```
save {t(Pc)}/outfile="C:\Example_data_correlation_preserving_plausible_values.sav"/variables
Pc_Ksi1 Pc_Ksi2 Pc_Ksi3 Pc_Eta1 Pc_Eta2.

END MATRIX.


* The following regression analyses are performed in order to check whether the
  correlation-preserving mean plausible values yield the same standardized
  coefficients as the BSEM model.

* Compare regression-coefficients of conventional mean plausible values...

Dataset activate Dataset1.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN   /DEPENDENT Eta1_meanPlausible
  /METHOD=ENTER Ksi1_meanPlausible Ksi2_meanPlausible  Ksi3_meanPlausible.

Dataset activate Dataset1.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN   /DEPENDENT Eta2_meanPlausible
  /METHOD=ENTER Ksi1_meanPlausible Ksi2_meanPlausible  Ksi3_meanPlausible.


* …with regression-coefficients of correlation-preserving mean plausible values:.

get file="C:\Example_data_correlation_preserving_plausible_values.sav".
Dataset name Dataset2.

Dataset activate Dataset2.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN   /DEPENDENT Pc_Eta1
  /METHOD=ENTER Pc_Ksi1 Pc_Ksi2 Pc_Ksi3.

Dataset activate Dataset2.
REGRESSION  /MISSING LISTWISE  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)  /NOORIGIN   /DEPENDENT Pc_Eta2
  /METHOD=ENTER Pc_Ksi1 Pc_Ksi2 Pc_Ksi3.
```

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Combining Correlated P-values From Primary Data Analyses

Jai Won Choi[1], Balgobin Nandram[2], & Boseung Choi[3]

[1] Statistical consultant, Meho Inc., U.S.A.

[2] Professor, Worcester Polytechnic Institute, U.S.A.

[3] Professor, Korea University, Sejong Campus, Korea

Correspondence: Jai Won Choi, 9504 Mary Knoll Drive, Rockville, MD, U.S.A.

Received: July 31, 2022 Accepted: October 3, 2022 Online Published: October 20, 2022

doi:10.5539/ijsp.v11n6p12 URL: https://doi.org/10.5539/ijsp.v11n6p12

Abstract

Research results on the same subject, extracted from scientific papers or clinical trials, are combined to determine a consensus. We are primarily concerned with combining p-values from experiments that may be correlated. We have two methods, a non-Bayesian method and a Bayesian method. We use a model to combine these results and assume the combined results follow a certain distribution, for example, chi-square or normal. The distribution requires independent and identically distributed (iid) random variables. When the data are correlated or non-iid, we cannot assume such distribution. In order to do so, the combined results from the model need to be adjusted, and the adjustment is done "indirectly" through two test statistics. Specifically, one test statistic (TS^{**}) is obtained for the non-iid data and the other is the test statistic (TS) is obtained for iid data. We use the ratio between the two test statistics to adjust the model test statistic (TS^{**}) for its non-iid violation. The adjusted TS^{**} is named as "effective test statistics" (ETS), which is then used for statistical inferences with the assumed distribution. As it is difficult to estimate the correlation, to provide a more coherent method for combining p-values, we also introduce a novel Bayesian method for both iid data and non-iid data. The examples are used to illustrate the non-Bayesian method and additional examples are given to illustrate the Bayesian method.

Keywords: assumed distribution, Correction ratio, Correlation, Model assumptions, P-values, Effective test statistic, Statistical inference

1. Introduction

Researchers use a model to combine the results, p-values or Z-scores, from sample surveys or clinical trials for the same subject or purpose. We consider these results are iid random variables and assume a certain distribution, for example normal, for statistical inference. Such a distribution requires iid-random variables.

However, these variables are more likely correlated as they are from the similar sample surveys or clinical trials for a specific topic or purpose. For example, poll results of presidential election or clinical trial results of one medication executed from different locations, or from the repeated trials at a same place (see Example 1). These results are often reported as p-values. We do not consider the previous procedures in obtaining p-values, and the k p-values are really the random variables. However, we are attacking a problem that is, indeed, very difficult because no aspect of the correlation is known, and moreover, there is a single sample of p-values, thereby making it impossible to find Pearson correlation.

The resulting p-values are non-iid random variables (see Example 1 and Appendix B). We present a method to show how an assumed distribution, which requires iid-random variables, can be applied to non-iid variables. To do so, non-iid variables need to be adjusted indirectly through its test statistics (TS^{**}). This adjustment is done by comparing two test statistics, one from the non-iid model and other from the iid model. The test statistic (TS^{**}) comes from a model with non-iid data, given null hypothesis, sample size and test level. Similarly, the other test statistic, (TS), comes from an assumed distribution with iid-random variables. We define correction factor as the ratio of TS^{**} to TS. Finally, we can get effective test statistic (ETS) of TS^{**} divided by the correction factor and this ETS is used to make statistical inference with the assumed distribution.

We use one of the two methods to combine the non-iid results or p^* values, Non-Bayesian or Bayesian. We show two methods for non-Bayesian in Section (3.1) show how to obtain ETS of correlated data (Choi and McHugh,1989), and in Section (3.2) show how to obtain ETS for TS^{**} of non-iid data, that involve not only correlation but also other non-iid-conditions, if any. Then, we use ETS with the assumed distribution.

The case of iid random variables to obtain TS

TS is based on a test statistic. It is the standard test statistic with which two other test statistics, TS^* or TS^{**}, are compared to measure the size of its deviation from TS, where TS^* is from a distribution of correlated variables and TS^{**} is from a distribution of non-iid variables. Below, we show how TS is obtained.

Suppose, p = $(p_1, \dots, p_n)$, $0 \leq p_i \leq 1$, i=1,…, n, are iid random variables with a known distribution function h(p|θ). One can make statistical inferences on p. Let the global null hypothesis H_o: $\theta_1 = \dots = \theta_n = \theta$ against alternative hypothesis $H_1: \theta_i \geq \theta$ for some i = 1, …, n. The hypothesis H_0 is reasonable as all the tests are done for a same purpose. We assume that h(p|θ) is a monotone function, and therefore it is optimal for combining p-values (Birnbaum, 1954).

We define test statistic (TS) as

$$TS_\alpha \text{or } t_\alpha = \text{T}(h(p|\theta), \alpha, n),$$

where the rejection test level α is obtained as

$$\alpha = 1 - \int_{-\infty}^{t_\alpha} \text{h(p}|\theta) \text{ dp.}$$

TS does not involve in hypothesis testing and it is based on the assumed distribution function $h(p|\theta)$ of iid p-values for given α and n. For example, h(p|θ) is Normal(μ, σ^2), or x_{2n}^2 chi-square 2n degrees of freedom. When we use $h(p|\theta)$ as base distribution of TS, we do not need actual p values, but the $h(p|\theta)$ implies p as iid random variables. For example, we only need sample size n and test level α to have table value of TS for x_{2n}^2, chi-square 2n degrees of freedom. The test level α is pre-selected by researcher. This TS is used only to compare to study test statistic, TS^* or TS^{**}, to measure its deviation from TS, and they involve in testing a null hypothesis at the same sample sized n and test level α of TS.

Above TS, based on $h(p|\theta)$ of iid-random variables p, is its own ETS. TS is compared to two study test statistics, TS^* based on correlated data and TS^{**} based on non-iid variables. We ignore the pre-procedures to obtain these data, and consider these data are the variables of our interest.

This paper has five more sections. In Section 2, we review pertinent literature. In Section 3, we present the non-Bayesian method. In Section 4, we show examples to illustrate the non-Bayesian method. In Section 5, we present Bayesian method to find the posterior mean of the combined p-value and some additional examples are presented. Section 6 includes a brief conclusion.

2. Pertinent Literature

Yoon et al (2021) used Meta analyses to increase statistical power by combining statistics (e.g., effect sizes, z- scores, or p-values) from multiple studies when they share the same null hypothesis under the assumption that all the data in each study have an association with a given phenotype. However, specific experimental conditions in each study can result in independent statistics that are derived from a null distribution. They showed the power of Meta analysis rapidly decreases as they were combined, Fisher's Method (Fisher, 1932), Weighted Fisher's method (wFisher), and Ordered p-values (ordMeta) increased power. The last two methods (i.e., wFisher and ordMeta), outperformed existing Meta-analysis when only a small number of studies n=2 is combined. The weighted Fisher's method (wFisher) assigned non-integer weights to each p-value, that are proportional to sample sizes. The wFisher and ordMeta are more robust than the test statistic of the Meta method.

Vovk and Wang (2020) got the average of k p-values $p_1,\dots,p_k$ to obtain one combined value without any parametric or distribution assumption. They reviewed previous results of arithmetic mean (AM $\bar{p}$) by multiplying 2 as $2\bar{p}$ and geometric mean (GM) replacing 2 by e (=2.718). They extended the recent risk aggregation technique to harmonic mean (HM) by multiplying log K for $K \geq 2$, scaling up by a factor of log k, where k is number of p-values. They also explore several other weighted averages of p-values. Note that the inequality of HM$\leq$ GM $\leq$AM, related to scaling factors, which is proved using Jensen's inequality (Casella and Berger, 2002).

Vovk and Wang (2020) showed several models to combine $p_1,\dots,p_k$ into a single p-value. assuming, $p_1,\dots,p_k$ are independent random variables. The simplest way to combine them is the Bonferroni method,

$$\text{F}(p_1,\dots,p_k) = K \min(p_1,\dots,p_k),$$

when $\text{F}(p_1,\dots,p_k)$ exceed 1, it can be replaced by 1. Other method, used to smooth out overestimation of above-mentioned method, is a general average:

$$M_{\emptyset,k}(p_1,\dots,p_k) = \varphi\left[\frac{\emptyset(p_1\}+,\dots,+\emptyset(p_K\}}{K}\right],$$

where $\emptyset(0,1) \to (-\infty, \infty)$ is a continuous strictly monotonic function and $\varphi[(0,1)] \to (0,1)$ is its inverse. For example, AM corresponds to the identity function $\emptyset(p)$= p, GM corresponds to $\emptyset(p)$= log p, and HM corresponds to $\emptyset(p)$= 1/p. They present more extensions of this basic idea.

Loughin (2004) compared several methods, when only p-values are available, in combining p-values from independent tests under combined hypothesis heuristically through simulation. They are minimum value (Tippett, 1931), Chi-square combining model (Fisher, 1932), scaled normal (Liptak, 1958), maximum value (Wilkinson, 1951), combinatoric uniform (Edington, 1972) and approximately scaled logistic (Rastogi, 1979).

Fisher's Model (FM) (1932) is $g(\boldsymbol{p}^*|\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log(p_i^*) = -2\log(p_1^* \dots p_n^*) = log\frac{1}{\{p_1^* \dots p_n^*\}^2}$. to combine $p_1^* \dots p_n^*$.

FM assumes the null hypothesis distribution follows x_{2n}^2, chi-square with 2n degrees of freedom for n independent random variables. This is not true when p^* are correlated. Other problem of FM arises when combining a large number of p^*-values. When n $\to \infty$, FM value$\to \infty$, i.e., combined value of even non-significant p-values becomes significant for a large n (Choi and Nandram, 2021).

Hess and Iyer (2007) used Fisher's Score combining p-values to detect differential genes array using Affymetrix expression arrays. Others (Tippett,1931, and Wilkinson,1951, George,1977, Stouffer,1949) suggest non-parametric methods to combine p-values.

Most methods, presented above, assumed independent p-values and did not address correlation or non-iid problems for statistical inference. Our research addresses a solution for this problem. However, this is a difficult problem because one cannot estimate the correlation in a straightforward manner, and this is an innovation in this paper as well. In a recent paper, Heard and Rubin-Delanchy (2018) showed how to choose between different methods to combine p-values. They also discussed the likelihood ratio for combining p-values and the weighted average of the logarithms of the p-values. However, there was no discussion about correlated p-values nor any discussion of the Bayesian approach, presumably there is none.

There is virtually no Bayesian attempt on the specific problem we are considering in this

paper. Specifically, we are combining a number of p-values, which may be dependent because the experiments are done under the same protocol, and similar procedures may be followed at the different experimental sites or laboratories. However, there is a sparse literature on the study of Bayesian p-values, not the combination of p-values. See Casella and Berger (1987) and the discussions that followed on reconciling Bayesian and frequentist evidence on the one-sided testing problem.

3. Non-Bayesian Method

Test statistics, TS^* for correlated variables and TS^{**} from non-iid variables, are compared by the standard rule, TS, for iid variables to see the size of their deviations from TS. We introduce these two test statistics, TS^* in (3.1) and TS^{**} in (3.2). We also present the correction factors, C^* and C^{**}, for TS^* and TS^{**} and its estimations. We also present Table 1 to illustrate practical application to clinical data.

In the introduction, we discussed the base test statistic TS for $h(p|\theta)$ with iid random variables p = $(p_1, \dots, p_n)$.as a standard rule to which TS^* or TS^{**} are measured.

In 3.1, the TS^* of g($p^*|\theta$) for correlated variables $p^* = (p_1^*, \dots, p_n^*)$ for given sample size and test level is compared to the base test statistic TS of $h(p|\theta)$ to find its difference, which is expressed as ratio, C^*= TS^* /TS. We call C^* correction factor (CF) as it corrects the impact of correlation on TS^*.

In 3.2, TS is now compared to TS^{**} for non-iid $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, which may carry not only correlation but also all other non-iid violations, if any. The difference between these two test statistics expressed as the ratio C^{**}= TS^{**} /TS. Here C^{**} corrects the impacts not only correlation but all other violations of iid condition.

In 3.3, we show how to estimate C^{**}. Three candidates are presented.

In 3.4, we illustrate TS, TS^{**}, and C^{**} in Table 1, using Fisher's Model F for TS^{**} and chi-square distribution C for TS. Table 1 is continuously used in the next Section 4. It shows for Fisher's Model users how to use the table values of C^{**} for possible violations of correlation or non-iid problem.

3.1 Correlated Random Variables, Model 1

Previously we introduced the base test statistic TS = T($h(p|\theta)$,α,n) for a known distribution h(p|θ) of iid random variables p = $(p_1, \dots, p_n)$, 0$\leq p_i \leq$1, i=1,…, n, for given test level α and sample size n.

Now we consider. We can obtain the test statistic (TS^*) for the combining model g($p^*|\theta$) of these correlated variables, $p^* = (p_1^*, \dots, p_n^*), 0 \le p_i^* \le 1$, for a given hypothesis H_O^*, test level α^*, correlation ρ and sample size n. We can assume g($p^*|\theta$) is its pseudo distribution and write

$$TS^* = \mathrm{T}(\mathrm{g}(p^*|\theta),\ H_O^*, \alpha^*, \rho, n).$$

Choi and McHugh (1989) discussed how to reduce the TS^* for the correlated variables in Chi-square testing. The g($p^*|\theta$) is erroneously assumed to follow h(p|θ), chi-square distribution x_{2n}^2. When the test statistic (TS) for distribution h(p|θ) is compaired to TS^* of the actual model g($p^*|\theta$), the test statistic, TS^* is largely inflated because of the correlation. Hence TS^* is reduced, dividing it by the correction factor C^*= [1 + ρ(n-1)] , ρ is the positive correlation among p^*-values, n is the sample size.

Choi and McHugh (1989) showed how to obtain the effective test statistic (ETS) of test statistic TS^* with this correction factor, C^*,

$$\mathrm{ETS}=\frac{TS^*}{C^*},$$

on $1 \le C^* < \infty$. It implies that the correlation of the variables $p^* = (p_1^*, \dots, p_n^*)$, is indirectly adjusted by the correction factor C^*. After such correction, we can now make statistical inference on the effective test statistic ETS with assumed distribution h(p|θ), for example chi-square distribution.

We can also achieve the same goal through effective sample size n_e of n, $n_e = \frac{n}{C}$ to obtain ETS (Choi,1980). For example, for binomial variables, x_i, i=1,…,n, that are correlated, its normal approximation of test statistic TS^* under the null hypothesis $H_0: p = 0$, is given as N(1,0)= $\frac{n\hat{p}}{\sqrt{p(1-p)}}$. We can use the reduced sample size $n_e = n/C$, to obtain effective test statistic, ETS =$\frac{n_e \hat{p}}{\sqrt{p(1-p)}}$.

3.2 Non-iid Random Variables, Model 2

In this section, we try to find the differences between the test statistic TS^{**} and basic test statistic TS, TS=T($h(p|\theta)$,α, n), and TS^{**}= T(g($p^{**}|\theta$), $H_O^{**}, \alpha^{**}, \rho, n^{**}$). Two types of differences can be considered: One is the correlation ρ in the variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$, and other includes all other known or unknown differences such as $h(p|\theta) \neq \mathrm{g}(p^{**}|\theta)$, $p^{**} \neq$ p, null hypothesis H_O^{**}, $\alpha \neq \alpha^{**}$, n $\neq n^{**}$.

The model $g(p^{**}|\theta)$ in TS^{**} is used to combine the non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$. The distribution $h(p|\theta)$ in TS is based on iid variables p= $(p_1, \dots, p_n)$. Users of the model $g(p^{**}|\theta)$ assume that $g(p^{**}|\theta)$ follows the distribution $h(p|\theta)$ as if p^{**} = p. It is a wrong assumption if $p^{**} \neq$ p. The aim of this section is to correct the wrong assumption indirectly by adjusting the test statistic, TS^{**}, while TS of assumed distribution h(p|θ) remains the same.

We have shown when TS is compared against TS^* for correlated variables $p^* = (p_1^*, \dots, p_n^*)$ in 3.1. Here in 3.2, we compare TS to TS^{**} for variables $p^{**} = (p_1^{**}, \dots, p_{n^{**}}^{**})$, which is not only correlated but also violated non-iid and other conditions, if any.

The total difference between the two test statistics, TS^{**} and TS, is defined as the ratio of these two test statistics:

$$C^{**}=\frac{TS^{**}}{\mathrm{TS}}, 0 < C^{**} < \infty.$$

Note that $TS^{**} \ge TS$ (Appendix A) when $1 < C^{**} < \infty$, and $TS^{**} < TS$ when $0 < C^{**} < 1$. The turning point greater than 1 or less than 1 depends on the size of p-values and the number n of the p-values as well as on the different changing speed, increasing or decreasing, of the TS^{**} and TS (see Table 1). We can ignore $TS^{**} < TS$ when $0 < C^* < 1$, since we assume only positive correlation of $p_1^{**}, \dots, p_{n^*}^{**}$ or consider only $TS^{**} \ge TS$ to correct positive correlation and other violation of TS^{**}.

To correct the impacts of non-iid and other violations, if any, we adjust TS^{**} by C^{**}as

$$ETS^{**} = \frac{TS^{**}}{C^{**}}, 0 < C^* < \infty.$$

Note $ETS^{**} > TS^{**} > TS$ on the interval $1 \leq C^{**} < \infty$ (Appendix A). The ETS** is the effective test statistic of the test statistic (TS**) on the interval,$1 < \boldsymbol{C}^{*} < \infty$. Here, the non-iid violation of the variables $p_1^{**}, \dots, p_{n^*}^{**}$, is indirectly corrected through C^{**}.

Lemma

The difference between the two test statistics, TS^{**} and TS can be expressed as the ratio , C^{**}=TS^{**}/ TS, $0 < C^{**} < \infty$, the correction factor, C^{**}, indirectly correct the correlation and other iid violations of TS^{**}. The effective test statistic is ETS^{**}=TS^{**} /C^{**}, on $1 < C^{**} < \infty$. Then, the effective test statistic ETS^{**} of test statistic, TS^{**}, is used for statistical inference with the originally assumed distribution $h(p|\theta)$.

Proof is outlined in Appendix A

3.3 Estimation of Correction Factor C^{**}

The correction factor C^{**} indirectly measures all violations including non-iid condition of p^{**}. In actual situation, it is difficult to obtain exact TS^{**} and hence C^{**}. To estimate $C^{**} = \frac{TS^{**}}{\text{TS}}, 1 < C^{*} < \infty$, we compare TS = T($h(p|\theta)$,$\alpha, n$) of assumed iid random variables p = $(p_1, \dots, p_n)$ to TS^{**}= T(g($p^{**}|\theta$), $H_0^{**}, \alpha^{**}, \rho, n^{**}$), of non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$. While the TS remains the same for given $h(p|\theta)$, α, n, the TS^{**} can be estimated by how we use $(p_1^{**}, \dots, p_{n^*}^{**})$ in the combing model g($p^{**}|\theta$). Below shows three ways of different use of these variables.

The three candidates are (1) is to use the minimum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, expressed as C_{Min}^{**}, (2) uses the maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, expressed as C_{Max}^{**}, (3) is the sum of individual values of TS^{**}, expressed as C_{Mix}^{**}, each term of $TS^{**}$is divided or individually weighted by all member weights (Example 1). All member weight is used because the weight of one member is one: when sample size is one (i.e., n=1), it is independent automatically regardless of the size of p-values, i.e., T($h(p|\theta), \alpha, n = 1$)= T(g($p^{**}|\theta$), $H_0^{**}, \alpha^{**}, \rho = 0, n^{**} = 1$), $h(p|\theta)$ = g($p^{**}|\theta$) for given $\alpha = \alpha^{**} = p = p^{**}$, ignoring the null hypothesis, H_0^{**}.as assumed distribution $h(p|\theta)$ is not involved in any null hypothesis. This is the only time the assumption is correct, or $h(p|\theta)$ = g($p^{**}|\theta$) (see First row, Table 1, Example 1).

Three possible correction factors are C_{Min}^{**}, C_{Max}^{**},and C_{Mix}^{**} (Appendix C). The choice depends on researcher's need. Thus, three different effective test statistics, ETS^{**}=TS^{**} /C^{**}, can be obtained when TS^{**} reduced by respective new correction factor:

$$ETS_{min}^{*} = \frac{TS^{**}}{C_{Min}^{*}},$$

$$ETS_{max}^{*} = \frac{TS^{**}}{C_{Max}^{*}},$$

and

$$ETS_{mix}^{*} = \frac{TS_1^{**}}{C_{Mix,1}^{*}} + \dots + \frac{TS_n^{**}}{C_{Mix,n}^{*}},$$

where $ETS_{max}^{**} \leq ETS_{mix}^{**} \leq ETS_{min}^{**}$, because $C_{Min}^{**} < C_{Mix,}^{**} < C_{Max,}^{**}$ (see Example 1). We may have the extreme cases of ETS_{max}^{**} and ETS_{min}^{**} when n^{*}-values of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ are widely spread out, and the minimum or maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ is comparatively very small or large, far away from the mean. In this situation, one may avoid the use of the two extreme cases and prefer to use middle value ETS_{mix}^{*} for the statistical inference in combining the value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$. Note the weights $C_{Mix,1}^{**}, \dots, C_{Mix,n}^{**}$ are each term weights for each TS_i^{**} of all member p_n^{**} (see Example 1, n=5 fifth row, for all 5 members, under each column of p-values).

3.4 Table 1, Numerical Example of Correction Factors C^{**}

The Table 1 below shows the numerical calculation to construct the test statistic TS (C), $TS^{**}(F)$, correction factors C^{**}, using chi-square value (C) for TS and Fisher's Model (F) for TS^{**} and the clinical trial data for $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$ (Example 1, Section 4).

One reason of presenting Table 1 here is to remind the users of Fisher's Model (FM) to be more careful if the data are correlated or non-iid variables. Often we find that, especially in medical journals, many people are still using FM without proper consideration of the problem as if data are iid random variables, Table 1 can be used to correct non-iid problems of their data when they use FM in combining p-values. Another reason to have Table 1 here is to help

understanding the text of next Section 4.

Table 1 shows the three numbers, FM (F), Chi-square model (C), and correction factor (C^{**}), by the p-vales on the columns, i.e., p= 0.01,0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, and the 15 numbers on the rows, i.e., n =1, 2, 3, 4, 5, 6, 7, 8, 9,10,11,12,13,14,15, each n-number means the same n p-values.(See Appendix A for the reason why we use the same p for n times). Recall that

$F = TS^{**}$ = T(g($p^{**}|\theta$), $H_0^{**}, \alpha^{**}, \rho, n^{**}$), test statistic for Fisher's Model g($p^{**}|\theta$), for given p^{**}, $H_0^{**}, \alpha^{**}, \rho, n^{**}$,

C = TS = T($h(p|\theta)$,α, n) of assumed base distribution $h(p|\theta)$ given α, n,

C^{**}= $\frac{F}{C} = \frac{TS^{**}}{\mathrm{TS}}, 1$

$0 < C^{**} < \infty$, in the Table 1, is the correction factor expressed as ratio of F

and C to compare them on the equal bases, (i.e., n=n**, and $\alpha=\alpha^{**} = p_i^{**}, \mathrm{i} = 1, \ldots, n^{**}$),

except correlation ρ and the forms of models g(.) and h(.), on the interval,$1 < C^{**} < \infty$,

this condition implies that C^{**} shows only impacts of correlation and model difference.

Note that here we use the five same values of p to induce the maximum correlation to F in C^{**}=$\frac{F}{c}$, while C remains the same, hence giving larger C^{**}, which, in turn, provides conservative or smaller ETS^{**}= $\frac{F}{C^{**}}$. Thus, users of C^{**}, in Table 1 will have conservative effective test statistic, ETS^{**}, when F is corrected by C^{**}.

To illustrate for the calculation of F, C, and C^{**}in Table 1, we take one cell for n=5, the fifth row and p=0.05 on the third column, Fisher's Model (FM), F = -2log 0.05 0.05 0.05 0.05 0.05)=29.96, using the same values five time for n=5 for the reason given above. The basic distribution, Chi-square value (C), C =18.31, for x_{2n}^2, 2n=10 degrees of freedom at $\alpha=\alpha^{**} = p_i^{**} = p^{**}$ = 0.05, from the table. The result is C^{**}= $\frac{F}{C}$= $\frac{29.96}{18.31}$.=1.64 as shown in the 5th row, n=5, and third column p=0.05 in Table 1. Other cells in Table1 follow the same steps to obtain F, C, and C^{**}.

Note we set the sample size n=n**=5, test level $\alpha=\alpha^{**} = p_i^{**} = 0.05$, to compare C and F on the equal bases except the correlation and the forms of two models, g(.) and h(.), i.e., g(.) ≠ h(.). Thus, the C^{**} shows the impacts of correlation and the wrong assumption of the model F in comparison to C.

We call C**=$\frac{F}{C} = \frac{TS^{**}}{TS}, 1 < C^{**} < \infty$, correction factor as they are indirectly used to correct or reduce TS^{**} for the violation of iid conditions and model assumption, for the data p_5^{**} = (05,.08, 0,09 0.10, 0.20), (see Appendix B). Effective Test Statistic (ETS^{**}=$\frac{TS^{**}}{C^{**}}$) is finally used for statistical inference. Note $ETS^{**} > TS, 1 < C^{**} < \infty$. (Appendix A).

Table 1. shows Fisher's Model F=TS^{**} and Chi-square Table value C= TS, and Correction Factors C**= F/C by the size of the nine p's, p=0.01, …, 0.9 on the columns, and the 15 numbers n=1,…,15 for the same n p-values on the rows

n of p	α =p→	0.01	0.02	0.05	0.1	0.2	0.3	0.5	0.7	0.9
n=1	F	9.21	7.82	5.99	4.61	3.22	2.41	1.39	0.71	0.21
	C	9.21	7.82	5.99	4.61	3.22	2.41	1.39	0.71	0.21
	C**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
n=2	F	18.42	15.65	11.98	9.21	6.44	4.82	2.77	1.43	0.42
	C	13.28	11.67	9.49	7.78	5.99	4.88	3.36	2.19	1.06
	C**	**1.39**	**1.34**	**1.26**	**1.18**	**1.07**	**0.99**	**0.83**	**0.65**	**0.40**
n=3	F	27.63	23.47	17.97	13.82	9.66	7.22	4.16	2.14	0.63
	C	16.81	15.03	12.59	10.64	8.56	7.23	5.35	3.83	2.20
	C**	**1.64**	**1.56**	**1.43**	**1.30**	**1.13**	**1.00**	**0.78**	**0.56**	**0.29**
n=4	F	36.84	31.30	23.97	18.42	12.88	9.63	5.55	0.56	0.84
	C	20.09	18.17	15.51	13.36	11.03	9.52	7.34	5.53	3.49
	C**	**1.83**	**1.72**	**1.55**	**1.38**	**1.17**	**1.01**	**0.76**	**5.53**	**0.24**
n=5	F	46.05	39.12	29.96	23.03	16.09	12.04	6.93	3.57	1.05
	C	23.21	21.16	18.31	15.99	13.44	11.78	9.34	7.27	4.87
	C**	**1.98**	**1.85**	**1.64**	**1.44**	**1.20**	**1.02**	**0.74**	**0.49**	**0.22**
n=6	F	55.26	46.94	35.95	27.63	19.31	14.45	8.32	4.28	1.26
	C	26.22	24.05	21.03	18.55	15.81	14.01	11.34	9.03	6.30
	C**	**2.11**	**1.95**	**1.71**	**1.49**	**1.22**	**1.03**	**0.73**	**0.47**	**0.20**
n=7	F	64.47	54.77	41.94	32.24	22.53	16.86	9.70	4.99	1.48
	C	29.14	26.87	23.68	21.06	18.15	16.22	13.34	10.82	7.79
	C**	**2.21**	**2.04**	**1.77**	**1.53**	**1.24**	**1.04**	**0.73**	**0.46**	**0.19**
n=8	F	73.68	62.59	47.93	36.84	25.75	19.26	11.09	5.71	1.69
	C	32 00	29.63	26.3	23.54	20.47	18.42	15.34	12.62	9.31
	C**	**2.30**	**2.11**	**1.82**	**1.56**	**1.26**	**1.05**	**0.72**	**0.45**	**0.18**
n=9	F	82.89	70.42	53.92	41.45	28.97	21.67	12.48	6.42	1.90
	C	34.81	32.35	28.87	25.99	22.76	20.60	17.34	14.44	10.86
	C**	**2.38**	**2.18**	**1.87**	**1.59**	**1.27**	**1.05**	**0.72**	**0.44**	**0.17**
n=10	F	92.10	78.24	59.91	46.05	32.19	24.08	13.86	7.13	2.11
	C	37.57	35.02	31.41	28.41	25.04	22.77	19.34	16.27	12.44
	C**	**2.45**	**2.23**	**1.91**	**1.62**	**1.29**	**1.06**	**0.72**	**0.44**	**0.17**
n=11	F	101.3	86.06	65.91	50.66	35.41	26.49	15.25	0.44	2.32
	C	40.29	37.66	33.92	30.81	27.3	24.94	21.34	18.1	14.04
	C**	**2.51**	**2.29**	**1.94**	**1.64**	**1.30**	**1.06**	**0.71**	**0.43**	**0.17**
n=12	F	110.5	93.89	71.9	55.26	38.63	28.90	16.64	8.56	2.53
	C	42.98	40.27	36.42	33.20	29.55	27.10	23.34	19.94	15.66
	C**	**2.57**	**2.33**	**1.97**	**1.66**	**1.31**	**1.07**	**0.71**	**0.43**	**0.16**
n=13	F	119.7	101.7	77.89	59.87	41.85	31.30	18.02	9.27	2.74
	C	45.64	42.86	38.89	35.56	31.79	29.25	25.34	21.79	17.29
	C**	**2.62**	**2.37**	**2.00**	**1.68**	**1.32**	**1.07**	**0.71**	**0.43**	**0.16**
n=14	F	128.9	109.5	83.88	64.47	45.06	33.71	19.41	9.99	2.95
	C	48.28	45.42	41.34	37.92	34.03	31.39	27.34	23.65	18.94
	C**	**2.67**	**2.41**	**2.03**	**1.70**	**1.32**	**1.07**	**0.71**	**0.42**	**0.16**
n=15	F	138.2	117.4	89.87	69.08	48.28	36.12	20.79	10.7	3.16
	C	50.89	47.96	43.77	40.26	36.25	33.53	29.34	25.51	20.6
	C**	**2.71**	**2.45**	**2.05**	**1.72**	**1.33**	**1.08**	**0.71**	**0.42**	**0.15**

Note in Table 1, C** = F/C is increasing from1.39 to 2.71 when n=2 increases to n=15 on the first column of p=0.01. It means that F is increasing faster than C as the number n of same p-values is increasing. This trend is reversed in the seventh column of p=0.5, C** is decreasing from 0.83 to 0.71 when n=2 increases to n=15. i.e., F decreasing faster than

C.

Similar trend exists on the rows, for the second-row n=2, C** is decreasing from 1.39 to 0.40 when p=0.01 increases to p=0.9. The change point C** greater than 1 to less than 1 is p=0.5, it is true for all the 15 rows.

Note that we ignore when $C^{**} = \frac{TS^{**}}{TS}, 0 < C^{**} < 1$, it happens data are negatively correlated. or

$TS^{**} < TS$ which happens when C^{**} does not reduce the impacts of non-iid inflation on TS^{**}.

4. Examples

Two examples are presented. (1) Effective Test Statistics ETS* of the Fisher's Model (FM) to combine p^*-values from clinical trial data at Minneapolis Veterans Administration (VA) Hospital. (2) Random group method for a large sample of n variables (Choi and Nandram, 2021). Using random grouping, we divide a large sample into k manageable random groups and obtain one p value from each group. Then the k p-values are combined, using FM.

4.1 Example 1. Fisher's Model (Fisher, 1932) to Combine Clinical Trial Results

All Parkinson patients, visiting the Neurology Department of Minneapolis VA hospital, are the population during the study period in 1970 (Choi, 1970). In our example, a sample of 36 patients is randomly selected from all the visitors. The 36 patients randomly ordered and took either Symmetrel, a candidate for Parkinson medication, or placebo, for 20 weeks crossover design, starting by coin toss, one week medication and one week placebo double blindly.

After each week, they took 5 tests: walking, tremor, stiffness, arm movement, and eye movement, to measure the impacts of medication or placebo. These tests are equally weighted assuming no residual effects, and calibrated from one to ten, one for no effect and 10 for the best result. The differences of on and off weeks are measured. Each patient provides 10 differences during 20 trial weeks and obtain one mean difference for each patient.

Again, find one mean differences from 36 patients for each of 5 tests, providing one mean difference from each of 5 tests. Using student-t test for the mean differences under the null hypothesis of no difference, we have 5 p-values from 5 tests, n=5, combined with Fisher's Model (FM), assuming they are iid random variables and follow Chi-square 10 degrees of freedom, x^2_{10}.

We have five p values of t-test under the null hypothesis of no mean differences. Once we have p-values, we ignore the previous procedures to obtain them and they are the random variables of our interest and may have their own distribution. The five p values are p^*_5 = (05,.08, 0,09 0.10, 0.20).

Fisher's model (FM) combines these 5 p-values.

FM = -2 log (0.05 x 0.08 x 0,09 x 0.10 x 0.2)

= $-$ 2(log 0.05 + log 0.08 + log 0.09 + log 0.10 + log 0.20)

=$-2(-2.9957$ - 2.5257 - 2.4080 - 2.3026 - 1.6094)

= 23.6828.

When we compare FM=23.6828 to the assumed Chi-square 10 degrees of freedom at α = 0.01 = 23.209, FM is significant as 23.6828 $>$ 23.209 at α = 0.01 of x^2_{10}.

However, the clinical trial data p^*_5 = (05,.08, 0,09 0.10, 0.20) are correlated (see Appendix B) or non-iid random variables, and thus, we cannot assume FM is distributed as chi-square 10 degrees of freedom. Therefore, FM = 23.6828 should be reduced for the violations of iid condition of p^*_5.

Most data are correlated in the real world as there is hardly any independent data.

But statisticians, in general, blindly assume their data are iid random variables. Thus, it is necessary to check out the independence and other characteristics of their data beforehand.

The three candidates, C^{**}_{Min}, C^{**}_{Max}, and $* C^{*}_{Mix}$ of Correction Factor are introduced in 3.3. They are used to reduce FM for iid violations.

(1) $C^{**}_{\min} = \frac{\boldsymbol{F(min)}}{\boldsymbol{C(min)}} = \frac{-2\log(0.05\ 0.05\ 0.05\ 0.05\ 0.05),\alpha=0.05}{x^2_{10,}(p=\alpha=0.05,n=5)} = 1.64$, using minimum($p^*_5$)=0.05.

(2) $C^{**}_{max} = \frac{F(max)}{C(max)} = \frac{-2\log(0.2\ 0.2\ 0.2\ 0.2\ 0.2),,\alpha=0.2}{x^2_{10,}(p=\alpha=0.2,n=5)} = 1.29$, using maximum($p^*_5$)=0.2.

Since individual weights are C^{**}=1 for n=1 (see first row, Table 1), we use an alternative weight.

(3) $C^{**}_{Mix,5} = \frac{F_i}{C_i}$ =1.64, 1.52, 1.46, 1.44. 1.20 (see, C^{**} in Table 1**,** row 5 for n=5 and corresponding columns of p= 0.05, 0.08, 0,09, 0.10, 0.20.

TS^{**} for Fisher's Model result (F) is adjusted by this correction factor (C^{**}) to obtain the effective test statistics (ETS^{**}) as shown below.

First, we find the minimum value of p^{**}_5 = (05,.08, 0,09 0.10, 0.20, which is 0.05, and use 0.05 five times to find FM (F) as explained the reason why we use the same number 0.05 five times. Then adjust FM by C^{**}_{min}=1.64 (Table 1, row n=5 and column p=0.05). We have

FM(min=0.05) = -2 log(0.05 0.05 0.05 0.05 0.05) = 2(5 x 2.99573) =29.9573,

$$\text{ETS**(min=0.05)} = \frac{TS^{*}(0.05)}{C^{**}_{min=0.05}} = \frac{29.9573}{1.64} = \mathbf{18.2667.}$$

Second, similarly, the maximum value 0.2 is used five times in FM, and FM(max=0.2)

is adjusted by $C^{**}_{max=0.2}$=1.29 (Table 1, row n=5 and column p=0.2). We have

FM(max=0.2)= -2 log(0.2 0.2 0.2 0.2 0.2)= 2(5x 1.60944)=16.0944.

$$\text{ETS**(max=0.2)} = \frac{TS^{*}(0.2)}{C^{**}_{max=0.2}} = \frac{16.0944}{1.29} = \mathbf{12.4763.}$$

Third, we obtain FM(mix) of individual value adjusted by individual combined weights C^{**}_i = 1.4, 1.52, 1.46,1.44,1.20 (Table 1, row 5, n =5 and columns corresponding to 0.05, 0.08, 0.09, 0.1, 0.2). The main reason why we use individual combined weights is, when n=1, individual weights C^{**}=1 regardless of p-vales. One sample is always independent so both FM and assumed chi-square distribution remain the same for given test level when sample size is one (see Table 1, row 1, C^{**}=1 for all p=values). We have

FM(mix) = -2 {log 0.05 + log 0.08 + log 0.09 + log 0.1 +log 0.2}, and each term is divided by the corresponding individual combined weight for the given reason. Hence, we have been

$$\text{ETS**(mix)} = \frac{-2\{\log 0.05\}}{1.64} + \frac{-2\{\log 0.08\}}{1.52} + \frac{-2\{\log 0.09\}}{1.46} + \frac{-2\{\log 0.1\}}{1.44} + \frac{-2\{\log 0.2\}}{1.20}$$

$$= \frac{2x2.99573}{1.64} + \frac{2x2.5257}{1.52} + \frac{2x2.408}{1.46} + \frac{2x2.3026}{1.44} + \frac{2x1.6094}{1.20}$$

$$= \frac{5.99146}{1.64} + \frac{5.0514}{1.52} + \frac{4.816}{1.46} + \frac{4.6052}{1.44} + \frac{3.2188}{1.20}$$

=3.4521+3.3233 +3.2986 +3.1981 +2.6823 =**15.9544.**

Results show that

ETS**(max=0.2)= **12.4763** < ETS**(mix) =**15.9544** < ETS** (min=0.05)= **18.2667.**

ETS**(min=0.05)=18.2667 is significant at α =0.05 of $x^2_{10,}$ (=18.307), but other two, ETS**(max=0.2)= **12.4763** and ETS**(mix) =**15.9544** are not significant.

In the beginning of this example, Fisher's Model gives FM=23.6826, without correction, which is significant at α = 0.01 of x^2_{10} (23.209). This FM is very much inflated when compared to above corrected results. Only one not-corrected value 29.9537 of FM(min=0.05) is bigger than the not-corrected FT=23.6826.

When the Maximum, here 0.2 or Minimum, 0.05, of p-values are too far away from the mean or relatively too small or too big, one may prefer the mixed value, ETS**(mix)=15.9544, for statistical inference, which is not significant at α = 0.01 of x^2_{10} (23.209), even at α =0.05 of $x^2_{10,}$ (=18.307).

4.2 Example 2. P-values from Random Groups of a Large Sample

When the existing methods, for example normal test or student t-test, are used for statistical inference, we encounter the large sample problems (Choi and Nandram, 2021). The reason is such test is the function of its variance, which in turn, function of sample size. The variance becomes too small when the sample size is large or too large when sample size is

too small. We consider the case of too large sample size, and test statistic becomes significant for the sample size over certain level (Choi and Nandram, 2021).

4.2.1 The Large Sample Problem

We indicate the large sample problem and show a solution using Random Group Method (Choi and Nandram, 2021). A concrete example is as follows. Let $x_1, x_2, \ldots, x_n$ be the realization of iid random variables $X_1, X_2, \ldots, X_n$, distributed as N(μ, σ^2), where σ^2 is known and inference is required about $\boldsymbol{\mu}$. We test the null hypothesis H_0: $\mu = \mu_0$ against alternative H_1: $\mu < \mu_0$. Let $\bar{\boldsymbol{x}}_{\boldsymbol{0}}$ be observed value of the sample mean, $\bar{x}$. Then the p-value of the test is

$$\mathrm{P}(\bar{\boldsymbol{x}} \leq \bar{\boldsymbol{x}}_{\boldsymbol{0}} \mid \boldsymbol{H_0})$$

$$= \mathrm{P}\left(\frac{\bar{x} - \boldsymbol{\mu_0}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \leq \frac{\bar{\boldsymbol{x}}_{\boldsymbol{0}} - \boldsymbol{\mu_0}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \mid \boldsymbol{H_0}\right)$$

$$= \boldsymbol{\Phi}\{\sqrt{\boldsymbol{n}}\ (\bar{\boldsymbol{x}}_{\boldsymbol{0}} - \boldsymbol{\mu_0})\ \}$$

Here $\boldsymbol{\Phi}(.)$ is the cdf of standard normal random variable. Therefore, if n is very large and $\bar{\boldsymbol{x}}_{\boldsymbol{0}} \leq \boldsymbol{\mu_0}$, p-value $\approx \boldsymbol{0}$ which shows large sample problem (Choi and Nandram, 2021). We use the following steps to solve this problem.

Step one

We divide a large sample of size n into a number of random groups so that each can be tested by the usual method. Let $\boldsymbol{x_n} = x_1, x_2, \ldots, x_n$ be a large sample of size n from N(μ, σ^2). When n is a large number, we cannot do the usual test. We want to divide the sample into h smaller samples of size m, 1< m < n, using Random Group Method. The smaller samples enable us to perform a traditional test (e.g., Normal test, t-test) for testing a hypothesis, H_0: $\mu = \mu_0$. Choi and Nandram (2021) showed how to divide the large sample into h smaller samples. Each sample provides one test statistic

$$t_i = \mathrm{T}(f(\mathrm{p} \mid \mu, \sigma^2), m_i, H_0, \alpha), \ m_i = \mathrm{m}, \quad \mathrm{i} = 1, \ldots, \mathrm{h}.$$

and the h test statistics provide h test scores $p_1, \ldots, p_h$ at the test level $\alpha_i = \alpha$, i=1, …,h.

Step two

When h p-values are iid variables, we can use Fisher's Model is assumed to be chi-square 2h degrees of freedom. We assume random groups are independent, we may assume h p-values are also independent, p = $p_1, \ldots, p_h$ are distributed as chi-square distribution, f(p|θ). We can make statistical inference with chi-square test result. However, If the p-values are correlated, we can use the correction factor in Table 1, to correct such impacts on Fisher's Model value.

Numerical example

A student presented data analysis of three sets of data; each includes 1500 persons' dental records. All the three t-tests of hypothesis H_0: $\mu = \mu_0$ were significant due to large sample size. Suggestion was to randomly divide 1.500 into 50 groups of 30 persons. If out of 50 t-tests, 45 tests (90%) of the 50 tests were significant at p=0.05, then it is also 90% significant for the 1500 persons' data at the same level at p=0.05 (Choi and Nandram, 2021). Similarly, it can be done for the remaining two groups.

5. Bayesian Model for Combining P-values

The Bayesian paradigm has the advantage of coherence, but the calculation of p-values is incoherent within the Bayesian paradigm because the computation of a tail area of a posterior distribution is not coherent. This is why Bayesians have hardly worked on this problem; see Casella and Berger (1987) and the discussions that followed. The combined p-value is an appropriate posterior mean, μ, say. However, note that μ is a parameter in the Bayesian paradigm, and it is a random variable.

It is not simple to include a correlation among the p-values since the sample of p-values is small. For the non-Bayesian method, we have constructed a correlation based on a distance measure (see Appendix B); otherwise, it is impossible to estimate this correlation. Here we will separate the data into groups to get an intra-cluster correlation.

The problem of combining a number of p-values, from the studies on the same subject, is one of data integration, which is currently a hot topic, see, for example, Nandram et al (2021) for model-based methods using both non-Bayesian and Bayesian approaches.

5.1 The Case of Independence

Suppose that we have the results of p-values $\hat{p}_1, \ldots, \hat{p}_n$ from n data sets, and these values are independent. We can also use an appropriate prior to reflect previous procedures to obtain p-values.

Let iid $\hat{p}_1,\ldots, \hat{p}_n \sim \text{Beta}\{\mu\frac{1-z}{z}, (1-\mu)\frac{1-z}{z})\}$ and $E(\hat{p}_i) = \mu$, $0 \leq \mu, z \leq 1$.

This is a useful reparameterization of the parameters of the Beta distribution in which both (μ, z) lie in (0,1), which leads to easy computation. See Nandram (2016) where this reparameterization was first introduced. A priori, we assume that

$$\mu, z \sim U(0,1),$$

essentially a non-informative prior.

We want to make inference about μ, combined p values. Letting $\hat{p}_a = \prod_{i=1}^{n} \hat{p}_i$, and $\hat{p}_b = \prod_{i=1}^{n}(1-\hat{p}_i)$, the posterior density of (μ, z) is

$$\pi(\mu,z|\hat{p}) \sim \left[\frac{\Gamma(\frac{1-z}{z})}{\Gamma\mu(\frac{1-z}{z})\Gamma(1-\mu)(\frac{1-z}{z})}\right]^n \hat{p}_a^{\mu\frac{1-z}{z}-1} \hat{p}_b^{(1-\mu)\frac{1-z}{z}-1}, \; 0 \leq \mu, z \leq 1.$$

For the samples from the posterior density, one can also use the Gibbs sampler (Casella and George,1992) to obtain μ and z for given p-values; but we use a random sampler that does not need any convergence monitoring.

The posterior summaries we use are the posterior mean (PM), posterior standard distribution (PSD), posterior coefficient of variation (PCV) and 95% highest density interval (HPDI).

Consider Example 1 on combining the five p-values, .05, .08, ,09, .10, .20. Applying our method based on the Beta model to these p-values, we computed the combined p-value, μ, and the posterior summaries are PM=.121, PSD=.032, PCV=.266, HPDI=(.069, .191). Therefore, the null hypothesis is not significant at the 5% significant level and perhaps not even at 10% significant level.

Table 2 has results of a small simulation study, which is used to provide many different examples. We generated n p-values, n=10,…,100, and we compare the combined p-value, the posterior mean of μ; we also look at z. Again, we show posterior summaries in Table 2 of the two variables, μ *and* Z, by sample size on the columns, and posterior mean (PM), posterior standard deviations (PSD), coefficient of variations (PCV) and 95% HPDIs of μ and z on the rows. Again, not that μ-values represent the posterior mean of the p-values, which range $0.05529 < \mu < 0.09157$. Note that the PSDs are decreasing as the sample size n increases. This also gives smaller PCVs and narrower 95% HPDIs e.g., at n=2 the 95% HPDI for μ is (.02945, .16355).

Table 2. Posterior summaries of μ *and* z including intervals

Sample size n	PM	PSD	PCV	95% Lower bound	95% Upper bound
n=10 μ	**0.09157**	**0.03414**	**0.37282**	**0.03945**	**0.16355**
z	0.09908	0.05641	0.56934	0.02105	0.20441
n=20 μ	**0.06136**	**0.01395**	**0.22729**	**0.04007**	**0.09056**
z	0.05462	0.02196	0.40201	0.02156	0.09700
n=30 μ	**0.05716**	**0.01028**	**0.17992**	**0.04096**	**0.07916**
z	0.04721	0.01501	0.31800	0.02101	0.07358
n=40 μ	**0.05810**	**0.00821**	**0.14122**	**0.04099**	**0.07149**
z	0.03979	0.01092	0.27439	0.02117	0.06064
n=50 μ	**0.05596**	**0.00675**	**0.12061**	**0.04206**	**0.06934**
z	0.03771	0.00901	0.23902	0.02110	0.05349
n=60 μ	**0.05545**	**0.00640**	**0.11540**	**0.04149**	**0.06795**
z	0.03787	0.00818	0.21608	0.02107	0.05085
n=70 μ	**0.05975**	**0.00616**	**0.10310**	**0.05117**	**0.07057**
z	0.04001	0.00760	0.19006	0.03101	0.06021
n=80 μ	**0.05529**	**0.00616**	**0.11149**	**0.04092**	**0.06617**
z	0.04357	0.00808	0.18538	0.03098	0.05878
n=90 μ	**0.05751**	**0.00571**	**0.09927**	**0.04879**	**0.07038**
z	0.04436	0.00798	0.17976	0.03099	0.05867
n=100 μ	**0.05859**	**0.00573**	**0.09778**	**0.05099**	**0.07015**
Z	0.04580	0.00778	0.16985	0.03101	0.05922

We may be able to include all information of first stage as prior replacing $\mu, z \sim U(0,1)$. This

Will be done in a future study. We can use independent Beta distributions with specified parameters, and this will depend on the amount of information available.

To motivate the case, where we include an intra-class correlation, we provide another Bayesian analogue of Fisher's model of combining p-values. Let $p_{i,}$ i= 1,…,n, denote the n p-values, and let $q_i = \log\{p_i/(1-p_i)\}$, independent, then a simple model is

$$q_i \mid \mu, \sigma^2 \sim \text{Normal}(\theta, \sigma^2)$$

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This is a standard non-informative prior (a version of Jeffrey's objective prior), but as always leading to proper posterior distribution for (θ, σ^2).

Here the combined p-value is $\emptyset = e^{\theta}/(1+e^{\theta})$. The posterior density of θ is a Student's t density, and inference about $\emptyset$ is obtained by sampling the Student's t density and computing $\emptyset$. For the example on the five p-values, for inference about $\emptyset$, we have posterior summaries, which are PM=0.099, PSD=0.033, PCV=0.334, HPDI=(0.044, 0.162). Again, the test is not significant at the 1 % significant level.

5.2 Including Correlation

We add an intra-cluster correlation as follows**.** We find all l = n(n-1)/2 distinct pairs of $q_i, \ldots, q_n$, and we form a Bayesian one-way random effect model, each cluster having just two values. Let y_{i1}, y_{i2}, i= 1,…,l, denote the distinct pairs which form the clusters. Then we assume the model,

$$y_{i1}, y_{i2} \mid \mu_i, \sigma^2 \widetilde{ind} N(\{\mu_i, (1- \rho)\sigma^2\}$$
$$\mu_i \mid \theta, \sigma^2, \rho \widetilde{ind} \text{N}(\theta, \rho\sigma^2), \quad \text{i= 1,…,}l,$$

$$\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}.$$

It is important to note that $\text{cor}(y_{i1}, y_{i2} \mid \vartheta, \sigma^2, \rho) = \rho$ in (0,1). We have actually used the traditional non-informative prior for $\pi(\theta, \sigma^2, \rho)$; this prior causes no impropriety issues (see Nandram, Toto and Choi, 2011) for proofs.

Also, note that we are actually assuming a composite likelihood because the pairs are not independent (i.e., each pair has one common unit), for example, see Varin, Reid and Firth (2011) for a discussion of composite likelihood. Again, the combined p-value is $\emptyset = e^{\theta}/(1+e^{\theta})$. This is the same as for the case when no correlation is assumed.

Using Bayes' Theorem, the joint posterior density is

$$\pi(\boldsymbol{\mu}, \theta, \sigma^2, \rho \mid \mathbf{q}) =$$
$$\pi_1(\boldsymbol{\mu} \mid \theta, \sigma^2, \rho \mid \mathbf{q})\ \pi_2(\theta \mid \sigma^2, \rho \mid \mathbf{q})\ \pi_3(\sigma^2 \mid \rho \mid \mathbf{q})\ \pi_3(\sigma^2 \mid \rho \mid \mathbf{q})\ \pi_4(\rho \mid \mathbf{q}).$$

Here, $\pi_1(\boldsymbol{\mu} \mid \theta, \sigma^2, \rho \mid \mathbf{q})$, $\pi_2(\theta \mid \sigma^2, \rho \mid \mathbf{q})$, and $\pi_3(\sigma^2 \mid \rho \mid \mathbf{q})$, have simple forms, and $\pi_4(\rho \mid \mathbf{q})$ has nonstandard form but it can be sampled using a grid method (e.g., Nandram, Toto and Choi, 2011). It is also true that the joint posterior density is proper, provided $l \geq 2$, see Nandram, Toto, and Choi (2011). Therefore, it is easy to sample the posterior density of θ and so $\emptyset$. To make inference about $\emptyset$, we draw 10,000 samples of the posterior density of $\emptyset$. No monitoring is required because a Markov chain Monte Carlo sampler is not used.

As summaries of the posterior density of $\emptyset$, we have PM=0.078, PSD=0.017, PCV=0.217, and the 95% HPDI= (0.048, 0.112). Therefore, the combined test is not significant at 5% significant level. Note that when we assume no correlation, PM=0.099 a bit larger, and the HPDI= (0.044, 0.162) a bit wider. The posterior summaries of ρ are PM=0.147, PSD=0.125, PCV=0.851, 95% HPDI=(0.001, 0.603); so, there is a small correlation.

As another example, when we increased the number of p-values to 10 (i.e., duplicate the five p-values to get 05, .08, ,09, .10, .20, 05, .08, ,09, .10, .20); there is an increase in precision but the results remain essentially the same. The posterior summaries of ρ are PM= 0.147, PSD= 0.125, PCV= 0.851, 95% HPDI= (0.001, 0.393); so that there is a small correlation, not much of a difference

6. Conclusion

We have used a model combine test scores on the same topic. Here, we assume a distribution for the data model. We compare the two test statistics, one from assumed distribution h(.) of iid-data and other from pseudo-distribution g(.) of

non-iid data. We define the differences between them as the ratio of the two. As the actual data may include impacts of not only correlation but also other difference of iid and non-iid conditions. We describe how to reduce the test statistics of non-iid data to make statistical inferences with the assumed distribution of iid variables.

We have considered two-stage procedure. The first stage is sampling and pre-processing to obtain the p-values. The second stage is the analysis of the first stage results.

Suppose that h independent samples. $y_1, \ldots, y_{n_i}$, i=1,…,h, are randomly taken from the population for an investigation on a same subject and suppose the sample follows true distribution f(y|θ). Each sample provides one test result from significant testing at a critical level α under a null hypothesis, providing test statistics.

$$t_i = \mathrm{T}(f(y_i|\theta), H_0, \alpha, n_i), \ \alpha_i=\alpha, \ \mathrm{i}=1,\ldots, \mathrm{h},$$

These test statistics provide h p*-values,

$$\alpha = 1 - \int_{-\infty}^{t_i} f(y_i|\theta) \ \mathrm{d}y_i, \ \mathrm{i} = 1,\ldots,\mathrm{h}.$$

Some assume the two stages are connected and the second stage is a continuation of the first. If the information such as sample design, sample, $f(y_i|\theta), H_0, \alpha,$ and sample size n_i are available, we can use this information in the second stage to combine the p_i^*-values to increase efficiency. Yoon et al.(2021) incorporate sample size n_i to combine p*-values. If one wants to include other information in Bayesian modeling, it is possible to use them as prior information.

The validity check of these estimations can be added in the future extension using the variance or coefficient of variation, and 95% confidence interval of each estimation through simulation.

It will be useful to carry out further study of the combination of correlated p-values in the Bayesian paradigm. For one thing, it will allow us to incorporate further information that can improve posterior inference. When available, information such as sample size and site covariates can be included in the combination of correlated p-values.

References

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association, 82*(397), 106-111.

Casella, G., & Berger, R. L. (2002). Statistical Inference, Second Edition. Duxbury, Pacific Grove, Ca.

Casella, G., & Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

Choi, J. (1970). Effectiveness of Symmetrel for Parkinson patients. Technical Report, Parkinson Laboratory, Neurology Department, VA Hospital, Minneapolis.

Choi, J. W. (1980). Ph. D. Thesis, University of Minnesota.

Choi, J. W., & McHugh, R. (1989). An adjustment factor for goodness and independent test for correlated and weighted observations. *Biometrics, 43,* 976-996.

Choi, J., & Nandram, B. (2021). Large sample problems. *International Journal of Statistics and Probability, 10*(2), 81-89.

Fisher, R. A. (1932). Statistical Methods for Research Workers (4th Edition), London: Oliver and Boyd.

George, E. O. (1977). *COMBINING INDEPENDENT ONE-SIDED AND TWO-SIDED STATISTICAL TESTS--SOME THEORY AND APPLICATIONS*. University of Rochester.

Hartung, J., & Knapp, G. (2005). Models for combining results of different experiments: retrospective and prospective. *American Journal of Mathematics and Management Sciences, 25,* 149-188.

Hartung, J., Bockenhoff, A., & Knapp, G. (2003). Generalized Cochran-Wald statistics in combination of experiments. *Journal of Statistical Planning and Inference, 113,* 215-237**.**

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika*, *105*(1), 239-246.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Boston Academic Press.

Held, L., Pawel, S., & Schwab, S. (2020). Replication power and regression to the mean. *Significance*, *17*(6), 10-11.

Hess, A., & Iyer, H. (2007). Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *Bmc Genomics*, *8*(1), 1-13.

Higgins, J. P., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in medicine*, *23*(11), 1663-1682.

Iyer, H. K., Wang, C. J., & Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, *99*(468), 1060-1071.

Li, Z., & Begg, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association, 89*, 1523-1527.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, *3*, 171-197.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis*, *47*(3), 467-485.

Matthews, R. (2021). The p-value statement, five years on. *Significance*, *18*(2), 16-19.

Nandram, B. (2016). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics, 32*(1), 187-208.

Nandram, B., Choi, J. W., & Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability, 10*(6), 4-17**.**

Nandram, B., Toto, M. C., & Choi, J. W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation, 81,* 1593-1608.

Rustagi (Ed.) (1979). Symposium on Optimizing Methods in Statistics, Academic Press, New York, 1979. 345-366.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. A., Jr. (1949). The American Soldier, Volume I. Adjustment during Army Life. Princeton, N.J.: Princeton University Press.

Tippett, L. H. C. (1931). The Methods of Statistics. London: Williams and Norgate Ltd.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica, 21*(1), 5-12.

Vovk, V., & Wang, R. (2020). Combining p-values via averaging. *Biometrika*, *107*(4), 791-808.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin, 48,* 156-158.

Yoon, S., Baik, B., Park, T., & Nam, D. (2021). Powerful p-value combination methods to detect incomplete association. *Scientific reports*, *11*(1), 1-11.

Appendix A, outline for the proof of Lemma

Correlation, Model 1

Consider the correlated random variables $p^* = (p_1^*, \dots, p_n^*)$. Choi and McHugh (1989) show how to adjust the TS_α^* based on correlated variables in Chi-square testing. Test Statistic (TS_α^*) for correlated data p^* is largely inflated and corrected by the correction factor C= [1 + ρ(n-1)] , ρ is the correlation among n p^*-values. $1 < C < \infty$.

$ETS_\alpha^* = \frac{TS_\alpha^*}{C}$. ETS_α^* can also be obtained by effective sample n_e of n, $n_e = \frac{n}{C}$. (Choi, 1980).

Non-iid case, correlation and other non-iid violations, Model 2

Here, we try to find the non-iid problem of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, indirectly through its test statistics TS^{**}, which is compared to test statistic TS of iid variables. The total difference between the two test statistics, TS^{**} and TS, can be expressed as the ratio of these two, $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, is used to get effective test statistics (ETS), which is used for statistical inference with $h(p|\theta)$.

$C^{**} = \frac{TS^{**}}{\mathrm{TS}} = \frac{\mathrm{T}*(\mathrm{g}(p^{**}|\theta), H_O^{**}, \alpha^{**}, \rho, n^{**}),}{\mathrm{T}(h(p|\theta), \alpha, n)}$.

The ratio, $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, $0 < C^{**} < \infty$ We consider C^{**} only on $1 \le C^{**} < \infty$, for positive correlation or TS^{**}> TS.

We do not consider or ignore TS^{**} < TS on $0 < C^{**} < 1$, for it does not reduce inflated TS^{**} for the impacts of non-iid violation (see Proof below). It happens also for negative correlation in C= [1 + ρ(n-1)] (see Method 1).
To prove TS^{**}> TS, consider two disjoint intervals, $(0 < C^{**} < \infty) = \{(0 < C^{**} < 1)\ U\ (1 \le C^{**} < \infty)\}$.

Let the effective test statistic be $ETS^{**} = \frac{TS^{**}}{C^{**}}$, and correction factor be $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$.

It is easy to see that $ETS^{**} < TS^{**}$ from $ETS^{**} = \frac{TS^{**}}{C^{**}}$, TS^{**} < TS from $C^{**} = \frac{TS^{**}}{\mathrm{TS}}$, on the interval $(0 < C^{**} < 1)$.

Similarly, $ETS^{**} \ge TS^{**}$ and TS^{**} > TS, on the other interval $(1 \le C^{**} < \infty)$.

The difference between TS^{**} and TS, $C^{**} = \frac{TS^{**}}{\mathrm{TS}} = \frac{\mathrm{T}(\mathrm{g}(p^{**}|\theta), H_O^{**}, \alpha^{**}, \rho, n^{**}),}{\mathrm{T}(h(p|\theta), \alpha, n)}$, C^{**} is less than 1 or greater than 1 depending also on n, p = α, and the increasing or decreasing speed of TS^{**} and TS (see Table 1).
If all the above conditions of TS^{**} and TS are same except ρ of $p_i^{**}s$, ignoring H_0^*, and $\alpha = \alpha^{**} = p^{**}$, and n = n^{**}, the proof depends only on correlation ρ : $0 \le \rho(p_i^{**}, p_{i'}^{**}) \le 1$, i $\ne i'$, for i, i' = 1, …, n. Model 1 can be used in this case.

(1) If $\rho = 0$, $C^{**} = \frac{TS_{\alpha^{**}}^{**}}{TS_\alpha} = \frac{\mathrm{T}(\mathrm{g}(p^{**}|\theta), \alpha^{**}, \rho, n^{**})}{\mathrm{T}(h(p|\theta), \alpha, n)} = 1$,. It is also true $C^* = 1$ when n=1. The sample size one is always independent, $\rho = 0$ and $\mathrm{T}(h(p|\theta),, \alpha, n = 1) = (\mathrm{g}(p^{**}|\theta),, \alpha^{**}, \rho = 0,\ n^{**} = 1)$ for g(.) =h(.) and $\alpha = \alpha^{**} = p^{**}$=p. This is the only time that FM for g(.) assumed correctly to be distributed as c hi-square C for h(.)

(2) If $0 < \rho \le 1$ and $2 \le n$, the correction factor $C^* = 1 + \rho(n-1), 1 < C^* < \infty$ (Choi and McHugh 19 89) and, if $\alpha = \alpha^{**} =$ p = $p_i^{**}, i = 1 \ldots, n^{**}, and\ n^{**} = n$, the effective test statistic $ETS_{\alpha^{**}}^* = \frac{TS_{\alpha^{**}}^{**}}{C^*} = \frac{TS_{\alpha^{**}}^{**}}{1+\rho(n-1)}$, C^* reduces the correlation impact of $TS_{\alpha^{**}}^{**}$.

For example: If the correlation among the 5 p-values of data 0.05, 0.08, 0.09, 0.10, 0.20, is ρ=0.42 (Appendix B). The correction factor $C = 1 + \rho(n-1) = 1 + 0.42(5-1) = 2.68$ and the Fisher's Model Test Statistic FM= $TS_{\alpha^{**}}^{**}$=23.68 is reduced as, $ETS_{\alpha^{**}}^{**} = \frac{23.68}{2.68}$=8.8361, this effective Test Statistic not significant at α =0.01 of $x_{10,}^2$ (=23.209).

If $\rho = 1$, for n= 5, C= $[1 + \rho(n-1)] = 1 + 1.0(5-1) = 5.00$, which is the largest correction value for any given n, and it ,in turn, gives the smallest $ETS_{\alpha^{**}}^* = \frac{23.68}{5} = 4{,}74$.

(3) We can also use the effective sample size n_e^*, $n_e^* = \frac{n^*}{C^*}, 1 \le C^* < \infty$ to obtain ETS^* (Choi, 1980) .

(4) The turning point also depends on the increasing or decreasing speed of $TS_{\alpha^{**}}^{**}$ and TS_α, $TS_{\alpha^{**}}^{**} < TS_\alpha$ when $0 < C^{**} < 1$ and $TS_{\alpha^{**}}^{**} > TS_\alpha$ when $1 < C^{**} < \infty$. We can ignore the case $TS_{\alpha^{**}}^{**} < TS_\alpha\ on\ 0 < C^{**} < 1$, as it happens for negative correlation of p^{**} variables. The change point from less than 1 to more than 1 also depends on the sample size n^{**} and size of p^{**}, for example, Table 1 shows the turning point is $at\ p^{**}$= 0.5 in the column and for all n on the rows,

Appendix B, the correlation of one sample

For one group of data including n variables $p_1, \dots, p_n$, currently there is no formula available to calculate ρ between the variables. We define $\rho_{(p_i p_j)} = \frac{\mathbf{1}}{|\boldsymbol{p_n - p_1}|} \frac{\sum_{i>j}^{n} |\boldsymbol{p_i - p_j}|}{\boldsymbol{n(n-1)/2}}$ for the continuous variables, $p_1, \dots, p_n$.

For example, p = (05,.08, 0,09, 0.10, 0.20),

$$\rho_{(p_i p_j)} = \frac{(0.03+ 9.04+ 0.05+ 0.15)+(0.01+ 0.02+ 0.12)+(0.01+ 0.11)+0.1}{|0.2-0.05)/|(5x4)/2} = \frac{0.27+0.15+0.12+0.1}{0.15x\ 10} = \frac{0.64}{1.5} = 0.4207$$

Appendix C, the three candidates of correction factor

TS = $T(h(p|\theta),\alpha,n)$ of iid random variables p = $(p_1, \dots, p_n)$ remain the same for given test level α and sample size n, while TS**= T(g($p^{**}|\theta$), $H_O^{**}, \alpha^{**}, \rho, n^{**}$) on the non-iid variables $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$

(1) C_{Min}^{**} uses the minimum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, all n** valuers are the same $p_{min.}^{**}$= $Min(p^{**})$ $= p_{min,i}^{**}$, i=1,…, n^{**}. to obtain the test statistic (TS**). The same minimum values are used to induc e the maximum correlation and in turn conservative TS**. (see Example 1 and Table 1)

$$C_{Min}^{**} = \frac{TS_{\boldsymbol{\alpha^*,min}}^{**}}{TS_\alpha} = \frac{T(g(p_{min,}^{**}, \dots, p_{min,}^{**}|\theta),\ H_O^{**}, \alpha^{**}, \rho, n^{**})}{T(h(p|\theta), \alpha, n)}.$$

(2) C_{Max}^{**} uses the maximum value of $p^{**} = (p_1^{**}, \dots, p_{n^*}^{**})$, similarly all n** valuers are $p_{max.}^{**}$= $Max(p^{**}) = p_{max,i}^{**}$, i=1,…, n^{**}.

$$C_{Max}^{**} = \frac{TS_{\alpha,max}^{**}}{TS_\alpha}, = \frac{T(g(p_{max.}^{**}, \dots, p_{max.}^{**}|\theta),\ H_O^{**}, \alpha^{**}, \rho, n^{**})}{T(h(p|\theta), \alpha, n\ )}.$$

(3) $C_{Mix}^{**} = C_{Mix,1}^{**} +, \dots, + C_{Mix,n^{**}}^{**}$,

where $C_{Mix,i}^{**} = \frac{TS_{.i}^{**}}{TS_n} = \frac{(T(g(p_i^{**}|\theta), H_O^*, \alpha^{**}, \rho, i))}{T_i(h(p_i|\theta), \alpha, n^{**})}$. $i = 1, \dots, n^{**}$.

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Review of Copula for Bivariate Distributions of Zero-Inflated Count Time Series Data

Dimuthu Fernando[1], Mohammed Alqawba[2], Manar Samad[3], Norou Diawara[1]

[1.] Department of Mathematics & Statistics, College of Sciences, Old Dominion University, Norfolk VA, USA

[2.] Department of Mathematics, College of Science and Arts, Qassim University, Ar Rass 51452, Saudi Arabia

[3.]Dept. of Computer Science, Tennessee State University, Nashville, TN, USA

Correspondence: Norou Diawara, Department of Mathematics & Statistics, College of Sciences, Old Dominion University, Norfolk VA, USA

Received: August 29, 2022 Accepted: October 11, 2022 Online Published: October 26, 2022

doi:10.5539/ijsp.v11n6p28 URL: https://doi.org/10.5539/ijsp.v11n6p28

Abstract

The class of bivariate integer-valued time series models, described via copula theory, is gaining popularity in the literature because of applications in health sciences, engineering, financial management and more. Each time series follows a Markov chain with the serial dependence captured using copula-based distribution functions from the Poisson and the zero-inflated Poisson margins. The copula theory is again used to capture the dependence between the two series.

However, the efficiency and adaptability of the copula are being challenged because of the discrete nature of data and also in the case of zero-inflation of count time series. Likelihood-based inference is used to estimate the model parameters for simulated and real data with the bivariate integral of copula functions. While such copula functions offer great flexibility in capturing dependence, there remain challenges related to identifying the best copula type for a given application. This paper presents a survey of the literature on bivariate copula for discrete data with an emphasis on the zero-inflated nature of the modelling. We demonstrate additional experiments on to confirm that the copula has potential as greater research area.

Keywords: count time series, copula, Zero-Inflated, count data, Poisson distribution

Subject Classification: 62H05, 62H10

1. Introduction

In the study of multivariate distributions, copula functions are gaining popularity in recent years. They are attractive as they can handle internal and mutual dependences among variables. The copula was first introduced in the Sklar (1954) paper, a paper that Frechet helped publish. Hoffding (1940) is also credited for almost innovating the concept of copula . Many problems in practical situations are modeled under related distributions using copula functions, in contrast to classical multivariate (Gaussian) distributions for count data. As such, the literature shows a growing interest in the investigation of dependence for sequences of counts in time series cases. The simplest of such sequences are bivariate count time series data. Copula functions have gained popularity in building such bivariate and multivariate distributions as the desire to understand the structure in massive time series count data is becoming more common. For diseases and rare events, observed counts over time appear in a high frequency of zeros (zero inflation), which is discussed in MÖller et al. (2020) and Young et al. (2020).

Sklar (1959) introduced a method to build in the bivariate and multivariate distributions for two random variables. The idea of joint distribution, especially in the bivariate case can be traced back to Frechet (1951, 1956, 1958). Morgensetrn (1956), Plackett (1965), Farlie (1969) and many other authors could be included in this systematic approach of constructing bivariate distributions with specific marginals and different dependence measures. See examples such as Gumbel (1958) or Johnson and Tenenbein (1981). In that same line of thought, Cook and Johnson (1981) asked two questions that are still of relevance. The questions are: 1) “Is there a distribution that appears to be the most promising candidate for non-normal types of data?” 2) “Is the resulting distribution or model fit significantly better than that obtained from the multivariate normal distribution?”

Finding a unique copula for a joint distribution requires one to know the form of the joint distribution. When using copula, one can separately model the marginal distributions and the dependence structure, which makes the copula

approach unique. Choosing the appropriate copula for a particular scenario means finding the one that best captures the dependence in data. Many variants of copulas have been proposed in the literature where each of these is suitable for different dependence structures. For example, Gaussian copula is flexible, and it allows for equally positive and negative dependence. The Clayton copula cannot account for negative dependence, and it exhibits strong left tail dependence. Similar to Gaussian copula, Frank copula allows for both positive and negative dependence between the marginals.

Copulas offer a flexible framework to combine distributions. It is unique if marginal densities are continuous. However, if some of the marginal distributions are discrete, the unicity cannot be obtained automatically.

Many copula functions have been identified, from the extreme of independent variables (the so-called independent copula or the product) to the max or min copula. The dependence is then captured by a selection of parameters and criteria associated with the range and properties of model parameters.

Moreover, high dimensional copulas have been introduced via bivariate copulas, under different decompositions and structures. These structures are known as the canonical vine (C-vine) or drawable vine (D-vine). References to C and D vines can be found in Bedford and Cooke (2002), Joe et al. (2010), and Aas et al. (2009). Gräler (2014) proposed the convex combination of bivariate copula densities incorporating the distance [between what?] as a parameter in the spatial setting. The application of copula functions can be found in finances (Czado et al, 2012), hydrology (Yu et al., 2020), transportation (Irannezhad et al., 2017), health care (Shi and Zhang, 2015), and more. The Farlie-Gumbel-Morgenstern (FGM) family of copula can be used to establish relationship between predictors (Durante and Sempi, 2016)).

Within the count time series, if we look at the binary data, there is a growing interest in the description of multivariate distributions under pair copulas (Lin and Chaganty, 2021). Panagiotelis et al. (2012) presented pair copula constructions for discrete multivariate data. Their algorithm is explained as a product of bivariate pair copula, demonstrating the great potential of vine copula approaches. They stated that the model selection for C or D vine remains an important open problem, with a particular emphasis on the conditional independence identification (Czado, 2019, Deng and Chaganty, 2021,). From there, the idea of using the D vine for modeling counts with excess zeros and temporal dependence is presented in Sefidi et al. (2020). Perrone and Durante (2021) highlighted the link between the extreme discrete copula and mathematical concept of convex polytope, which is an idea spinning from the class of bivariate distributions (Rao and Subramanyam (1990).

There are numerous problems and interesting challenges related to time series of counts. Davis et al. (2016, 2021) presented extensive literature and many examples of count time series. Fokianos (2021) and Armillota and Fokianos (2021) presented a Poison network autoregression for counts. In the statistical process control, Fatahi et al. (2012) proposed the monitoring of rare events under the copula based bivariate zero-inflated Poisson. van Den Heuvel et al. (2020) proposed corrections to such results adding the negative correlation option.

With these studies and observations in mind, this paper presents reviews and updates related to the copula for bivariate distributions of zero-inflated count time series and highlights research directions. Motivated by multivariate datasets acquired using correlation structures, our goal is to review the bivariate count and zero-inflated count time series for inference and application purposes under copula modeling. We give some insights into the bivariate count copula and its recent developments. We organize our discussion as follows. In Section 2, copulas for discrete count and zero-inflation of discrete count time series data are described. The use of univariate and bivariate copula for discrete data is discussed in Section 3. Extensions of discrete bivariate copulas are described in Section 4. We conclude this paper with an extended discussion on future work.

2. Copula for Zero-inflated of Discrete and Count Time Series Data

This section introduces the general form for multivariate copula, and its Gaussian representation. We also give an explicit definition of the zero inflated counts time series data.

2.1 Simple Gaussian Copula Example

Masarotto and Varin (2012) introduced a Gaussian copula model which can be used to model time series data in the presence of covariates. The corresponding regression model can be written as follows.

$$\mathrm{Y_t} = \mathrm{g}(X_t, \epsilon_\mathrm{t}\, \theta), \text{for t} = 1, .., \mathrm{n},$$

where $g(.)$ is a function of the covariates X_t and ϵ_t, which capture the serial dependence. The parameter $\boldsymbol{\theta}$ is a vector of marginal regression coefficients. The joint distribution function of the time series $\{Y_t\}\, for\, t = 1, \dots, n$ can be constructed using the Gaussian copula as follows.

$$F(y_1, y_2, \dots, y_n) = P(Y_1 \le y_1, Y_2 \le y_2, \dots, Y_n \le y_n) = \Phi_{R(\rho)}(\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2)) \dots, \Phi^{-1}(F_n(y_n))) \quad (1)$$

Here, Φ^{-1} is the inverse CDF of standard normal distribution, and $\Phi_{R(\rho)}$ is the joint CDF of a multivariate normal distribution with a mean vector of zeros and covariance matrix *R*.

2.2 Review of Copula for Discrete Data

Copula distributions are becoming increasingly popular in many areas of statistical data sciences. For example, in engineering, copula distributions are used to model the shear fore for cantilever beams and for beams with multiple point loads (Zhang and Lam, 2016). In pharmaceutical quality control, two correlated characteristics sample data are presented in Fatahi et al. (2012). The authors describe the bivariate Poisson distribution with the evidence of zero-inflation. Sukparungsee et al. (2021) developed a bivariate copula for control chart effectiveness. They show the bivariate copula distribution on Hotelling's T^2 over the multivariate cumulative sum for positive, negative, weak, moderate, and strong correlations when the assumption of multivariate normality is violated. Van den Heuvel et al. (2020) extended the idea from Fatiha et al. (2012) and included negative correlation case, and an upper control limit on the sum of bivariate random variables. Copulas are elegantly captured in the Genest and MacKay (1986), Genest (1987) and also in Han and De Oliveira (2016 and 2020), among others. In the financial sector, a recent work by Nikoloulopoulos and Moffatt (2019) reminds us of the need to study dependence structures. There are also more general ambitions for the bivariate copula from a bigger perspective than we expect to show the aggregated effects in many other areas.

The list of copula functions is very large. The work of Größer and Okhrin (2021) presents a summary of bivariate copula followed by the construction of multivariate copula using pair copula decompositions. They provide examples for each copula family and provide an overview of how copula theory can be used in various fields of data science.

Yang et al. (2014) proposed the Ali-Mikhael-Haq (AMH) copula-based function to investigate the joint risk probabilities of rainstorms, wind speeds, and storm surges. The proposed model was developed to assess the impact based on marginal distributions of maximum daily rainfall and extreme gust velocity. Alqawba et al. (2021) constructed a class of bivariate integer-valued time series models using copula theory. Applying either the bivariate Gaussian copula or the bivariate t copula functions, they jointly modeled two copula-based Markov time series models. They applied their method on bivariate count time series data, where the marginals follow either a Poisson or zero-inflated Poisson distribution.

Safari et al. (2020) proposed a bivariate copula regression model to analyze cervical cancer data. They applied a bivariate copula to model and estimate joint distribution parameters. Nikoloulopoulos and Moffatt (2019) used bivariate copulas to jointly model bivariate ordinal time-series responses with covariates for risks assessment of married couples. They proposed a copula-based Markov modelling of ordinal time-series responses and used another copula to couple their conditional (on the past) distributions at each time point. Copula families such as the Bivariate normal (BVN), Frank, Gumbel and bivariate t-copula were used to model the univariate time series as well as to couple them together.

The work of Nikoloulopoulos & Karlis (2010) presents a regression copula-based model where covariates are used not only for the marginal but also for the copula parameters. They measured the effect of covariates on dependence structure by building a fully parametric copula-based model while considering six one-parameter copula families, namely Frank, Galambos, Gumbel, Mardia–Takahasi (M-T), and normal to build the dependence structure.

Karlis & Pedeli (2013) presented a bivariate integer-valued autoregressive process (BINAR(1)) in which the cross-correlation was modeled using a copula to accommodate both positive and negative correlation. They presented an application of the Frank and Gaussian copula to model dependence, and marginal time series were modeled using Poisson and negative binomial INAR(1) distributions.

Ma et al. (2020) proposed a copula approach utilizing a Gaussian copula with random effects to model correlated bivariate count data regression.

2.3 The Zero-Inflated Discrete Data

Zero inflation models can be found in many studies from Lambert (1992) to Hall (2000) and recently in Rigby et al. (2019). The zero-inflated count regression models are described as follows.

- Zero-Inflated Poisson (ZIP) Distribution (Lambert, 1992):

$$F_{Y_t}(m) = \omega_t + (1-\omega_t)e^{-\lambda_t}\sum_{y_t=0}^{m}\frac{\lambda_t^{y_t}}{y_t!}. \quad (2)$$

- Zero-Inflated Negative Binomial (ZINB) Distribution (Ridout et al, 2001):

$$F_{Y_t}(m) = \omega_t + \frac{(1-\omega_t)}{\Gamma(\kappa_t)}\left(\frac{\kappa_t}{\kappa_t+\lambda_t}\right)^{\kappa_t} \sum_{y_t=0}^{m} \frac{\Gamma(\kappa_t+y_t)}{y_t!} \left(\frac{\lambda_t}{\kappa_t+\lambda_t}\right)^{y_t}.$$

- Zero-Inflated Conway-Maxwell-Poisson (ZICMP) Distribution (Sellers and Raim, 2016):

$$F_{Y_t}(m) = \omega_t + \frac{(1-\omega_t)}{Z(\lambda_t+\kappa_t)} \sum_{y_t=0}^{m} \frac{\lambda_t^{y_t}}{(y_t!)^{\kappa_t}},$$

where $\lambda_t = \exp(\boldsymbol{X}_t'\beta), \omega_t = \frac{\exp(\boldsymbol{z}_t'\gamma)}{1+\exp(\boldsymbol{z}_t'\gamma)}$, and $\kappa_t = \exp(\boldsymbol{W}_t'\alpha)$

are the associated covariate vectors affecting the intensity parameter λ_t, the zero-inflation parameter ω_t and the dispersion parameter κ_t, respectively.

The term $\sum_{y_t=0}^{m} \frac{\lambda_t^{y_t}}{(y_t!)^{\kappa_t}}$ is the normalizing function of the CMP.

Different variants of similar regression models have been proposed in the literature. A noteworthy use of copula for zero-inflated data is studied in Shamma et al. (2020), where the inflation is built from a geometric count time series in an integer-valued autoregressive (INAR) process.

3. Univariate and Bivariate Copula Models for Count Time Series Data

3.1 Univariate Copula-Based Model for Count Time Series Data

First order Markov model

Alqawba, & Diawara (2021) introduced a class of Markov zero inflated count time series model where the joint distribution function of the consecutive observations is constructed through copula functions. Suppose $\{Y_t\}$ zero-inflated count time series first order Markov chains the multivariate joint density distribution of $Y_1, Y_2, \ldots, Y_n$ can be constructed as below.

$$Pr(Y_1 = y_1, \ldots, Y_n = y_n) = Pr(Y_1 = y_1) \prod_{t=2}^{n} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1})$$

Using the copula theory, the joint distribution function of Y_t, Y_{t-1} can be written as below.

$F_{12}(y_t, y_{t-1}) = C(F_t(yt), F_{t-1}(y_{t-1}); \boldsymbol{\delta})$ *where* $\boldsymbol{\delta}$ *is bivariate copula parameter vector.*

Hence, we can calculate the transition probability as below.

$$\Pr(Y_t = y_t | Y_{t-1} = y_{t-1}) = \frac{\Pr(Y_t = y_t, Y_{t-1} = y_{t-1})}{f_{t-1}(y_{t-1})}$$

Where

$$Pr(Y_1 = y_1, Y_{t-1} = y_{t-1}) = F_{12}(y_t, y_{t-1}) - F_{12}(y_t - 1, y_{t-1}) - F_{12}(y_t, y_{t-1} - 1) + F_{12}(y_t - 1, y_{t-1} - 1)$$

Likelihood and parameter estimation under first order Markov model

The likelihood function of the first order Markov model is given by

$$L(\vartheta, y) = \Pr(Y_1 = y_1; \theta) \prod_{t=2}^{n} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}; \vartheta) \quad (3)$$

The log likelihood function ($l(\vartheta; y)$) is given by

$$l(\vartheta; y) = \log \Pr(Y_1 = y_1; \theta) + \sum_{t=2}^{n} log \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}; \vartheta)$$

Where θ *and* δ are the parameter vectors of the marginals and the dependence structure, respectively. For the Gaussian copula family, the likelihood function involves a bivariate integral of the normal probability in $C(.;\delta)$ which means that the function is not in a closed form and we need approximations for the rectangle probabilities.

The simulation study was conducted using the **R software** by the **'optim'** function in the **"stats" package.** We simulate first order stationary Markov processes with joint distribution of consecutive observations following the bivariate Gaussian copula. The marginal distributions are chosen to be the Poisson and ZIP distributions. We present the simulation results for a first order Markov model with Poisson marginals. The parameter λ represents the mean of a marginal Poisson, ω is the measure of zero inflation, and δ is the serial dependence associated with time series data.

We found that the estimate of these parameters is fairly stable where the precision increases with increasing sample size. Table 1 and Table 2 show the estimates of copula parameters for positive and negative autocorrelations, respectively. The estimates are described by standard measures of variation, including standard deviation, mean square error and mean absolute error.

Univariate ZI count time series models

For positive serial dependence with λ=3, ω=0.3, δ =0.6

Table 1. Parameter estimates for the univariate ZI Poisson model with positive autocorrelation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda(3)$	2.990	0.347	0.1200	0.282
	$\omega(0.3)$	0.288	0.083	0.0070	0.006
	$\delta(0.6)$	0.577	0.091	0.0080	0.073
300	$\lambda(3)$	3.013	0.192	0.037	0.152
	$\omega(0.3)$	0.293	0.046	0.002	0.037
	$\delta(0.6)$	0.596	0.046	1.433	1.196
500	$\lambda(3)$	3.006	0.154	0.024	0.120
	$\omega(0.3)$	0.295	0.035	0.001	0.028
	$\delta(0.6)$	0.596	0.037	0.001	0.028

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

For negative serial dependence with λ=3, ω=0.3, δ =-0.6

Table 2. Parameter estimates for the univariate ZI Poisson model with negative autocorrelation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda(3)$	3.045	0.280	0.080	0.234
	$\omega(0.3)$	0.299	0.046	0.002	0.036
	$\delta(-0.6)$	-0.618	0.087	0.0070	0.072
300	$\lambda(3)$	3.019	0.152	0.023	0.119
	$\omega(0.3)$	0.298	0.030	0.0007	0.002
	$\delta(-0.6)$	-0.605	0.050	0.003	0.040
500	$\lambda(3)$	3.014	0.112	0.0127	0.009
	$\omega(0.3)$	0.299	0.019	0.0004	0.015
	$\delta(-0.6)$	-0.603	0.040	0.002	0.031

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

Applications

Alqawba & Diawara (2021) applied the proposed model to analyze monthly count of strong sandstorms recorded by the AQI airport station in Eastern Province, Saudi Arabia. The data set consists of 348 monthly counts of strong sandstorms, starting from January 1978 to December 2013. The bar plots suggest that both counts follow Zero inflated Poisson distribution, whereas the ACFs indicate that the counts are serially dependent. Finally, to illustrate the superiority of the proposed method they compare the method with zero-inflated integer-valuedautoregressive (ZIINAR) models.

3.2 Bivariate Copula-Based Model for Count Time Series Data

Copula based bivariate model

Suppose we have $\{Y_{1t}\}$ and $\{Y_{2t}\}$ jointly observed at timepoints t=1, 2, …, n, with the assumption that each series $\{Y_{1t}\}$ and $\{Y_{2t}\}$ follows a copula-based Markov process described on section 3.1. Let's mean vector, correlation matrix of the bivariate series as μ_t

and $\tau(t, t-1)$ which are described as below.

$$\mu_t = E(Y_t) = \begin{bmatrix} E(Y_{1t}) \\ E(Y_{2t}) \end{bmatrix}$$

$$\tau(t, t-1) = COV(Y_t, Y_{t-1}) \begin{bmatrix} COV(Y_{1t}, Y_{1,t-1}) & COV(Y_{1t}, Y_{2,t-1}) \\ COV(Y_{2t}, Y_{1,t-1}) & COV(Y_{2t}, Y_{2,t-1}) \end{bmatrix}$$

Here the diagonal elements of the matrix represent the serial dependence between two series, while the off-diagonal elements describe the cross-correlation between two time series.

The joint distribution of Y_{1t} and Y_{2t} given Y_{1t-1}, Y_{2t-1} for t=1, 2,…,n is given by

$$\mathrm{f}(y_{1t}, y_{2t} | y_{1t-1}, y_{2t-1}) = \int_{V^{-1}(F_{1,t}^-)}^{V^{-1}(F_{1,t}^+)} \int_{V^{-1}(F_{2,t}^-)}^{V^{-1}(F_{2,t}^+)} V_2(z_1, z_2, R) d z_2 \, dz_1$$

where V^{-1} is either the inverse cdf (Cumulative distribution function) of the normal distribution or the t-distribution with $V_2(., R)$ being the bivariate normal or t-distribution, respectively. R is correlation matrix capturing the cross correlation between two time series which is described below.

$$\mathrm{R}=\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The limits of the bivariate integral can be calculated as below.

$F_{i,t}^+ = F(y_{i,t} | y_{i,t-1})$ *and* $F_{i,t}^- = F(y_{i,t} - 1 | y_{i,t-1})$, for i= 1,2 where,

$$F(y_{i,t} | y_{i,t-1}) = \frac{F_{12}(y_{i,t}, y_{i,t-1}) - F_{12}(y_{i,t}, y_{i,t-1} - 1)}{f_{t-1}(y_{i,t-1}; \boldsymbol{\theta})}$$

and

$$F_{12}(y_{i,t}, y_{i,t-1}) = C(F_t(y_{i,t}), F_{t-1}(y_{i,t-1} - 1); \delta)$$

$C(.; \delta)$ represents the bivariate copula function with dependence parameter δ, describing the serial dependence in a single series, and $\boldsymbol{\theta}$ is a vector of the marginal parameters.

Likelihood and parameter estimation for the bivariate model

Likelihood based inference were conducted with maximizing the log-likelihood function of the bivariate distribution. The corresponding likelihood function for the joint distribution is given by,

$$L(\vartheta, y) = f(Y_{11}, Y_{21}). \prod_{t=2}^{n} f(Y_{1t}, Y_{2t} | Y_{1t-1}, Y_{2t-1}) \qquad (4)$$

Where $\vartheta = (\theta', \delta_1, \delta_2, \rho)'$,where θ is the marginal parameter vector and δ_1, δ_2 are parameters associated with the serial dependence in each time series respectively. The cross correlation between the two-time series is captured by ρ.

We can construct the log -likelihood function $l(\vartheta, y)$ as below.

$$l(\vartheta, y) = \log(f(Y_{1t}, Y_{2t})) + \sum_{t=2}^{n} \log f(Y_{1t}, Y_{2t} | Y_{1t-1}, Y_{2t-1}).$$

The likelihood function ($l(\vartheta, y)$) contains either a bivariate normal or t-integral function which unable us to use the standard maximization procedures to get the ML estimates. Due to this reason, we evaluated the bivariate integral function using the standard randomized importance sampling method.

We present simulation results for the proposed bivariate model in Section 3.1 after expanding from univariate to bivariate model. For each univariate time series, we considered a copula-based Markov model, where a copula family was used for the joint distribution of subsequent observations, and then, coupled these two-time series using another copula at each time point.

The parameters of the marginal Poisson distribution are shown in Table 3 and Table 4 for positive and negative cross correlations, respectively. Here $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ denote the means, $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ denote zero inflation parameters, $\boldsymbol{\delta_1}$ and $\boldsymbol{\delta_2}$ denote the serial dependence of marginal distributions. ρ is measure of the cross correlation between the two time series distributions.

The Gaussian copula was used to construct marginal distributions for 300 replicates with sample sizes of 100,300 ,500 and the true parameter values are presented in brackets. The count time series with positive cross correlation is presented in Figure 1, and the joint density is shown in Figure 2. When observing the parameter estimates displayed in Table 3, we can state that the estimated values are more precise and converges to the true parameter values as the

sample size increases.

Bivariate ZI count time series models

Table 2. Parameter estimates for the bivariate ZI Poisson model with positive cross correlation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda_1(3)$	3.4021	0.3887	0.3123	0.4599
	$\omega_1(0.3)$	0.3333	0.0835	0.0081	0.0701
	$\lambda_2(5)$	5.1993	0.3832	0.1860	0.3337
	$\omega_2(0.4)$	0.4026	0.0686	0.0047	0.0537
	$\delta_1(0.6)$	0.5425	0.0837	0.0103	0.0788
	$\delta_2(0.4)$	0.3628	0.0963	0.0106	0.0806
	$\rho(0.5)$	0.4822	0.0911	0.0086	0.0748
300	$\lambda_1(3)$	3.4051	0.1974	0.2030	0.4082
	$\omega_1(0.3)$	0.3380	0.0447	0.0034	0.0471
	$\lambda_2(5)$	5.1816	0.2097	0.0768	0.2226
	$\omega_2(0.4)$	0.4065	0.0386	0.0015	0.0309
	$\delta_1(0.6)$	0.5540	0.0433	0.0040	0.0524
	$\delta_2(0.4)$	0.3669	0.0544	0.0040	0.0492
	$\rho(0.5)$	0.4711	0.0493	0.0033	0.0441
500	$\lambda_1(3)$	3.4105	0.1721	0.1980	0.4108
	$\omega_1(0.3)$	0.3408	0.0365	0.0030	0.0456
	$\lambda_2(5)$	5.1843	0.1622	0.0602	0.2028
	$\omega_2(0.4)$	0.4084	0.0293	0.0009	0.0246
	$\delta_1(0.6)$	0.5558	0.0320	0.0030	0.0465
	$\delta_2(0.4)$	0.3700	0.0430	0.0027	0.0413
	$\rho(0.5)$	0.4720	0.0392	0.0023	0.0379

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

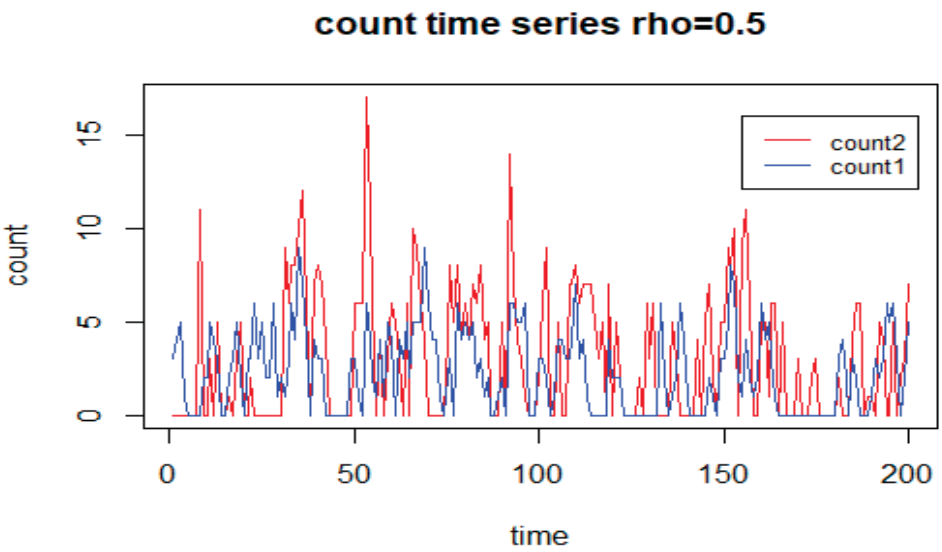

Figure 1. Plot of individual ZI count time series with positive cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

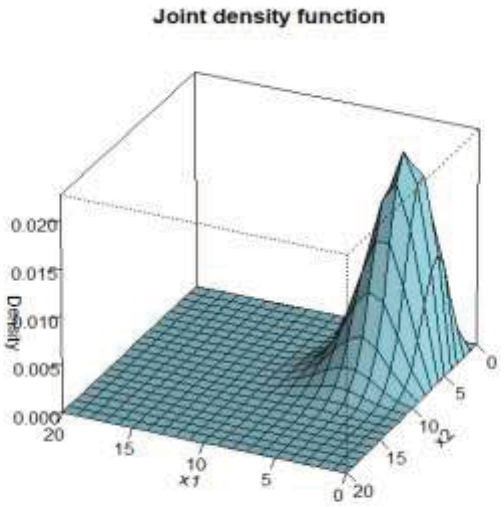

Figure 2. Joint probability density function for the bivariate ZI model with positive cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

There are times when the correlation is negative and table 4 shows the parameter estimates for such scenarios. The Gaussian copula was again used in constructing marginal distributions for 300 replicates with sample sizes of 100, 300, 500 and the true parameter values are presented in brackets. The count time series with negative cross correlation is illustrated in Figure 3, and the joint density is shown in Figure 4. The estimated parameters in Table 4 are more precise and converge to the true parameter values with increasing sample size as observed before.

These results are new because a large body of the literature focuses on positive correlations. Therefore, our proposed algorithm can handle less restrictive cases of ZI count time series data.

Table 3. Parameter estimates for the bivariate ZI Poisson model with negative cross correlation

Sample Size	Parameters	Estimate	St_Dev	MSE	MAE
100	$\lambda_1(3)$	3.417	0.388	0.324	0.474
	$\omega_1(0.3)$	0.341	0.084	0.074	0.074
	$\lambda_2(5)$	5.225	0.382	0.196	0.354
	$\omega_2(0.4)$	0.408	0.070	0.056	0.056
	$\delta_1(0.6)$	0.549	0.085	0.010	0.077
	$\delta_2(0.4)$	0.368	0.103	0.012	0.086
	$\rho(-0.4)$	-0.391	0.104	0.011	0.081
300	$\lambda_1(3)$	3.4072	0.2016	0.2063	0.4110
	$\omega_1(0.3)$	0.3378	0.0455	0.0035	0.0477
	$\lambda_2(5)$	5.2100	0.1965	0.0826	0.2331
	$\omega_2(0.4)$	0.4077	0.0379	0.0015	0.0313
	$\delta_1(0.6)$	0.5529	0.0458	0.0043	0.0534
	$\delta_2(0.4)$	0.3683	0.0537	0.0039	0.0499
	$\rho(-0.4)$	-0.3815	0.0559	0.0035	0.0465
500	$\lambda_1(3)$	3.4181	0.1727	0.2045	0.4182
	$\omega_1(0.3)$	0.3364	0.0348	0.0025	0.0412
	$\lambda_2(5)$	5.1984	0.1575	0.0641	0.2138
	$\omega_2(0.4)$	0.4094	0.0304	0.0010	0.0254
	$\delta_1(0.6)$	0.5524	0.0321	0.0033	0.0493
	$\delta_2(0.4)$	0.3731	0.0417	0.0025	0.0388
	$\rho(-0.4)$	-0.3794	0.0460	0.0025	0.0414

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

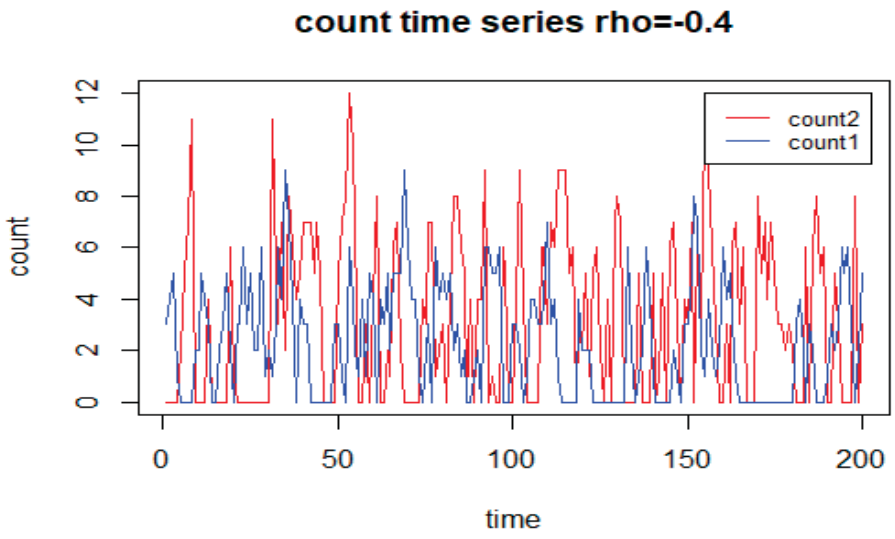

Figure 3. Plot of individual ZI count time series data with negative cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022)

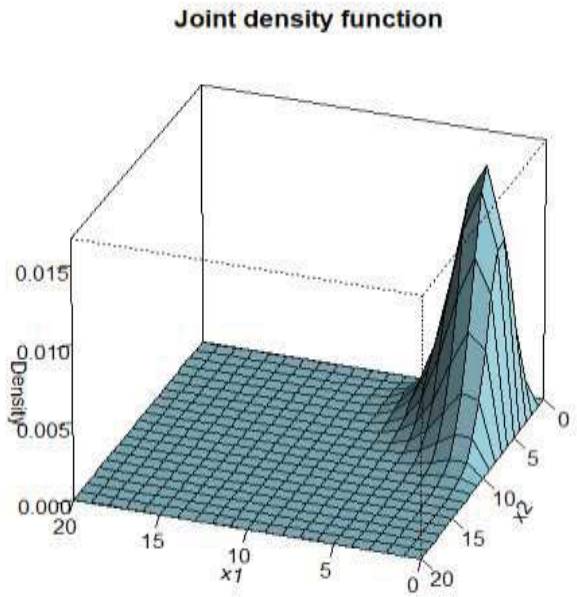

Figure 4. Joint probability density function for the bivariate ZI model with negative cross correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

Applications

The proposed class of method can be applied to model bivariate zero inflated count time series data in the presence of both temporal dependence and cross correlation.

Wang et al. (2013) proposed a bivariate zero inflated poison model to analyze occupational injuries. Alqawba et al. (2021) applied this framework to model monthly counts of forgery and fraud in the 61st police car beat in Pittsburgh, PA. Two count time series were selected to fit the proposed bivariate Poisson class of models under the clear evidence

of the presence of serial dependence and cross correlation.

4. Extensions of the Bivariate Copula for Count Time Series Data

Many copulas have been proposed in the literature for the bivariate and multivariate distributions. The choice of the copula is mainly dictated by the dependence structure.

As shown in Größer and Okhrin (2021), the research on time series dependence and copula direction is productive and has numerous applications. They showed examples of bivariate copulas. Count time series data are observed in several applied disciplines such as environmental science, biostatistics, economics, public health, and finance. Sometimes, a specific count, usually zero, may occur more often than other counts. Moreover, overlooking the frequent occurrence of zeros could result in misleading inferences. A copula-based time series regression model for zero-inflated counts is developed. Applying ordinary Poisson and Negative Binomial distributions to these time series of counts may not be appropriate due to the frequent occurrence of zeros. A new form of ZI is called the Conway-Maxwell Poisson (CMP).

Alqawba et al. (2021) have extended the work done by Masarotto (2012) to include a class of models that accounts for ZI. The marginals are assumed to follow one of the ZIP, ZINB, and ZICMP distributions, and the serial dependence was modeled by a Gaussian copula with a correlation matrix that of a stationary ARMA process. Likelihood inference was carried out using sequential importance sampling. Simulated studies were conducted to evaluate the parameter estimation procedures. Model description and parameter misspecification or unidentifiability are always concerns from the data generation to real data analysis (Faugeras, 2017). Model assessment to check the goodness of fit for the proposed models was done via residual analysis. The proposed models were applied to the occupational health data. According to the residual analysis, the model fits the data adequately, but both ZINB and ZICMP seem to have a slight advantage over ZIP distribution. Future direction is to consider different model construction methods from the marginal regression such as Markov models to handle zero-inflated count time series data. Recently, the use of copula-based time series for ZI counts in the presence of covariates has been proposed in Alqawba et al. (2019) and Alqawba and Diawara (2020). The work considered the cases of ZIP, ZINB, and ZICMP distributed marginals. Likelihood-based inference is considered under a sequential sampling method to estimate both the marginal regression parameters and copula parameters. Improvements in the Bayesian Information Criteria were noted, as discussed in Joe (2014) and Dalla Valle et al. (2018). The applications of these models include occupational injury counts, arson counts, and sandstorm counts.

5. Further Developments and Conclusion

Several high-dimensional copulas are obtained from the bivariate version seen in the previous section. The bivariate time series copula becomes then very important. The vine copula is built from blocks of bivariate version of higher dimension (Acar et al. 2019, Czado). We will only mention the Hierachical Archimedean copula, the Multivariate Archimax copula, the Factor copula, and the Vine copula. Copula functions are particularly interesting in capturing dependence with pairwise Kendall's correlations for invariance to monotonic transformations of marginal distributions. The copula is Archemedian and is applicable for higher than bivariate dimensions of the correlation between marginals (McNeil and Nešlehová, 2009). There is research on the symmetry of copula, and the family of measures under non-degenerate asymptotic distributions (Quessy and Bahraoui, 2018). The disentangling of features with copula transformation is also gaining popularity in so called deep Information bottleneck (DIB) to yield higher convergence rates (Wieczorek et al. 2018, Wieczorek and Roth 2020). As a measure, the copula can be thought as a transformation on a set, which is also a measure preserving transformation. Copulas are also obtained under non-monotonic transformations. Bardossy and Li (2008) proposed a v-transformed copula.

The ideas of Levy processes modelled via copula offer many areas of research (Liu et al., 2021).

The spatio-temporal dependence will become more of a priority as the research evolves. See more in Krupskii and Genton (2017). Bivariate time varying copulas are proposed in Acar et al. (2019). The dynamic vine copula is also adapted to the Bayesian inference (Kreuzer and Czado, 2019).

In this review, we have shown statistical and computational methods for bivariate count time series data analyses using copula distributions. The general framework for discrete count data and the bivariate nature of data are presented. The copula structure is described with details on its analytic perspectives. The identifiability and the choice of copula are very challenging in any discrete data setting and in the case of negative associations between components. As mentioned in Genest et al. (2011), Faugeras (2017) and in Trivedi and Zimmer (2007, 2017), the copula may not generate the perfect data distributions. Such concern is also pointed out in Durante and Sempi (2016). Copula can model bivariate dependence that are invariant under monotonic transformation only (Größer and Okhrin, 2021). When the dependence is weak, the FGM copula offers great alternative, but determining the most appropriate type of FGM copula to fit data is an open problem. Trivedi and Zimmer (2017) proposed several simulations to show these concerns.

Similar to any other functions, the copula functions cannot be deemed as the solution to all data problems. However,

they offer a valuable alternative, especially in the case of discrete data. The research on discrete time series data is more important in this class of functions, especially for bivariate cases as the characterization of bivariate count dependence structure provides tools for may applied problems.

Conflict of Interest

We attest that the manuscript titled "Review of copula for bivariate distributions of zero-inflated count time series data" is original and has not been submitted to or considered for publication elsewhere. The authors declare that they have no competing or conflicts of interest with regard to this publication.

Funding Information

This study received no external funding.

References

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, *44*(2), 182-198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Acar, E. F., Czado, C., & Lysy, M. (2019). Flexible dynamic vine copula models for multivariate time series data. *Econometrics and Statistics*, *12*, 181-197. https://doi.org/10.1016/j.ecosta.2019.03.002

Alqawba, M., & Diawara, N. (2021). Copula-based Markov zero-inflated count time series models with application. *Journal of Applied Statistics*, *48*(5), 786-803. https://doi.org/10.1080/02664763.2020.1748581

Alqawba, M., Diawara, N., & Chaganty, R. (2019). Zero-Inflated Count Time Series Models Using Gaussian Copula. *Sequential Analysis Journal: Design methods and Applications, 38*(3), 342–357. https://doi.org/10.1080/07474946.2019.1648922

Alqawba, M., Fernando, D., & Diawara, N. (2021). A Class of Copula-Based Bivariate Poisson Time Series Models with Applications. *Computation, 9*(10), 108. https://doi.org/10.3390/computation9100108

Armillotta, M., & Fokianos, K. (2021). Poisson network autoregression. *arXiv preprint arXiv:2104.06296*.

Bahraoui, T., Bouezmarni, T., & Quessy, J. F. (2018). Testing the symmetry of a dependence structure with a characteristic function. *Dependence Modeling*, *6*(1), 331-355. https://doi.org/10.1515/demo-2018-0019

Bedford, T., & Cooke, R. M. (2002). Vines--a new graphical model for dependent random variables. *The Annals of Statistics*, *30*(4), 1031-1068. https://doi.org/10.1214/aos/1031689016

Cook, R. D., & Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(2), 210-218. https://doi.org/10.1111/j.2517-6161.1981.tb01173.x

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, *222*. https://doi.org/10.1007/978-3-030-13785-4

Czado, C., Schepsmeier, U., & Min, A. (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, *12*(3), 229-255. https://doi.org/10.1177/1471082X1101200302

Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., ... & Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association*, *116*(535), 1533-1547. https://doi.org/10.1080/01621459.2021.1904957

Davis, R. A., Holan, S. H., Lund, R., & Ravishanker, N. (Eds.). (2016). *Handbook of discrete-valued time series*. CRC Press. https://doi.org/10.1201/b19485

Deng, Y., & Chaganty, N. R. (2021). Pair-copula models for analyzing family data. *Journal of Statistical Theory and Practice*, *15*(1), 1-12. https://doi.org/10.1007/s42519-020-00146-z

Durante, F., & Sempi, C. (2016). *Principles of copula theory* (Vol. 474). Boca Raton, FL: CRC press. https://doi.org/10.1201/b18674

Durante, F., Sánchez, J. F., & Sempi, C. (2018). A note on bivariate Archimax copulas. *Dependence Modeling*, *6*(1), 178-182. https://doi.org/10.1515/demo-2018-0011

Fatahi, A. A., Noorossana, R., Dokouhaki, P., & Moghaddam, B. F. (2012). Copula-based bivariate ZIP control chart for monitoring rare events. *Communications in Statistics-Theory and Methods*, *41*(15), 2699-2716. https://doi.org/10.1080/03610926.2011.556296

Fokianos, K. (2021). Multivariate count time series modelling. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.11.006

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, *74*(3), 549-555. https://doi.org/10.1093/biomet/74.3.549

Genest, C., & MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, *40*(4), 280-283. https://doi.org/10.1080/00031305.1986.10475414

Genest, C., Nešlehová, J., & Ziegel, J. (2011). Inference in multivariate Archimedean copula models. *Test*, *20*(2), 223-256. https://doi.org/10.1007/s11749-011-0250-6

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, *10*, 87-102. https://doi.org/10.1016/j.spasta.2014.01.001

Größer, J., & Okhrin, O. (2022). Copulae: An overview and recent developments. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(3), e1557. https://doi.org/10.1002/wics.1557

Gumbel, E. J. (1958). Distributions à plusieurs variables dont les marges sont données. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, *246*(19), 2717-2719.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, *56*(4), 1030-1039. https://doi.org/10.1111/j.0006-341X.2000.01030.x

Han, Z., & De Oliveira, V. (2016). On the correlation structure of Gaussian copula models for geostatistical count data. *Australian & New Zealand Journal of Statistics*, *58*(1), 47-69. https://doi.org/10.1111/anzs.12140

Han, Z., & De Oliveira, V. (2020). Maximum likelihood estimation of Gaussian copula models for geostatistical count data. *Communications in Statistics-Simulation and Computation*, *49*(8), 1957-1981. https://doi.org/10.1080/03610918.2018.1508705

Irannezhad, E., Prato, C. G., Hickman, M., & Mohaymany, A. S. (2017). Copula-based joint discrete–continuous model of road vehicle type and shipment size. *Transportation Research Record*, *2610*(1), 87-96. https://doi.org/10.3141/2610-10

Joe, H., Li, H., & Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, *101*(1), 252-270. https://doi.org/10.1016/j.jmva.2009.08.002

Johnson, M. E., & Tenenbein, A. (1981). A bivariate distribution family with specified marginals. *Journal of the American Statistical Association*, *76*(373), 198-201. https://doi.org/10.1080/01621459.1981.10477628

Karlis, D., & Pedeli, X. (2013). Flexible bivariate INAR (1) processes using copulas. *Communications in Statistics-Theory and Methods*, *42*(4), 723-740. https://doi.org/10.1080/03610926.2012.754466

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14. https://doi.org/10.2307/1269547

Lin, H., & Chaganty, N. R. (2021). Multivariate distributions of correlated binary variables generated by pair-copulas. *Journal of Statistical Distributions and Applications*, *8*(1), 1-14. https://doi.org/10.1186/s40488-021-00118-z

Liu, Y., Djurić, P. M., Kim, Y. S., Rachev, S. T., & Glimm, J. (2021). Systemic risk modeling with lévy copulas. *Journal of Risk and Financial Management*, *14*(6), 251. https://doi.org/10.3390/jrfm14060251

Ma, Z., Hanson, T. E., & Ho, Y. Y. (2020). Flexible bivariate correlated count data regression. *Statistics in Medicine, 39*(25), 3476-3490. https://doi.org/10.1002/sim.8676

Masarotto, G. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics, 6*, 1517-1549. https://doi.org/10.1214/12-EJS721

McNeil, A. J., & Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ1-norm symmetric distributions. *The Annals of Statistics*, *37*(5B), 3059-3097. https://doi.org/10.1214/07-AOS556

Nikoloulopoulos, A. K., & Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics, 37*(9), 1555-1568. https://doi.org/10.1080/02664760903093591

Nikoloulopoulos, A. K., & Moffatt, P. G. (2019). Coupling couples with copulas: analysis of assortative matching on risk attitude. *Economic Inquiry, 57*(1), 654-666. https://doi.org/10.1111/ecin.12726

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, *107*(499), 1063-1072. https://doi.org/10.1080/01621459.2012.682850

Rao, M. B., & Subramanyam, K. (1990). The structure of some classes of bivariate distributions and some applications. *Computational Statistics & Data Analysis*, *10*(2), 175-187. https://doi.org/10.1016/0167-9473(90)90063-N

Ridout, M., Hinde, J., & Demétrio, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, *57*(1), 219-223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press. https://doi.org/10.1201/9780429298547

Safari-Katesari, H., Samadi, S. Y., & Zaroudi, S. (2020). Modelling count data via copulas. *Statistics*, *54*(6), 1329-1355. https://doi.org/10.1080/02331888.2020.1867140

Sellers, K. F., & Raim, A. (2016). A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, *99*, 68-80. https://doi.org/10.1016/j.csda.2016.01.007

Shamma, N., Mohammadpour, M., & Shirozhan, M. (2020). A time series model based on dependent zero inflated counting series. *Computational Statistics*, *35*(4), 1737-1757. https://doi.org/10.1007/s00180-020-00982-4

Shi, P., & Zhang, W. (2015), Private information in healthcare utilization: specification of a copula-based hurdle model. *J. R. Stat. Soc. A, 178,* 337-361. https://doi.org/10.1111/rssa.12065

Sklar, A. (1973). Random variables, joint distribution functions and copulas. *Kybernetika, 9,* 449–460.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, *8*, 229-231.

Trivedi, P. K., & Zimmer, D. M. (2007). *Copula Modeling: An Introduction for Practitioners*. Foundations and Trends in Econometrics. https://doi.org/10.1561/0800000005

Trivedi, P., & Zimmer, D. (2017). A note on identification of bivariate copulas for discrete count data. *Econometrics*, *5*(1), 10. https://doi.org/10.3390/econometrics5010010

Valle, L. D., Leisen, F., & Rossini, L. (2018). Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(3), 523-548. https://doi.org/10.1111/rssc.12237

van den Heuvel, E. R., van Driel, S. A., & Zhan, Z. (2022). A bivariate zero-inflated Poisson control chart: Comments and corrections on earlier results. *Communications in Statistics-Theory and Methods*, *51*(10), 3438-3445. https://doi.org/10.1080/03610926.2020.1736304

Wang, K., Lee, A. H., Yau, K. K., & Carrivick, P. J. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*, *35*(4), 625-629. https://doi.org/10.1016/S0001-4575(02)00036-2

Weiß, C. H., Möller, T., & Kim, H. Y. (2020). Modelling counts with state-dependent zero inflation. *Statistical modelling*.

Wieczorek, A., & Roth, V. (2020). On the difference between the information bottleneck and the deep information bottleneck. *Entropy*, *22*(2), 131. https://doi.org/10.3390/e22020131

Wieczorek, A., Wieser, M., Murezzan, D., & Roth, V. (2018). Learning sparse latent representations with the deep copula information bottleneck. *arXiv preprint arXiv:1804.06216*.

Yang, Q., Xu, M., Lei, X., Zhou, X., & Lu, X. (2014). A Methodological Study on AMH Copula-Based Joint Exceedance Probabilities and Applications for Assessing Tropical Cyclone Impacts and Disaster Risks (Part I). *Tropical Cyclone Research and Review*, *3*(1), 53-62.

Young, D. S., Roemmele, E. S., & Shi, X. (2022). Zero-inflated modeling part II: Zero-inflated models for complex data structures. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(2), e1540.

Yu, R., Yang, R., Zhang, C., Špoljar, M., Kuczyńska-Kippen, N., & Sang, G. (2020). A vine copula-based modeling for identification of multivariate water pollution risk in an interconnected river system network. *Water*, *12*(10), 2741. https://doi.org/10.3390/w12102741

Zhang, Y., & Lam, J. S. L. (2016). A copula approach in the point estimate method for reliability engineering. *Quality and Reliability Engineering International*, *32*(4), 1501-1508. https://doi.org/10.1002/qre.1860

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Simple Sampling for SARS-CoV-2 Infection in Hidalgo

Lucia V. P. Torres[1] & Juan B. Guerrero Escamilla[1]

[1] Institute of Social Sciences and Humanities, Autonomous University of the State of Hidalgo, Hidalgo, Mexico.

Correspondence: Lucia V. P. Torres, Doctorate in Public Policy, Institute of Social Sciences and Humanities, Autonomous University of the State of Hidalgo, Hidalgo, Mexico.

Received: September 8, 2022 Accepted: October 10, 2022 Online Published: October 26, 2022

doi:10.5539/ijsp.v11n6p41 URL: https://doi.org/10.5539/ijsp.v11n6p41

Abstract

Throughout the history of our country, different policies have left an incentive for favorable changes, however, none by itself has managed to combat the problems of chronic malnutrition, to which the current pandemic is added. The state of Hidalgo is in a nutritional transition, with persistent child undernutrition and the predominance of chronic diseases associated with malnutrition (undernutrition, overweight and obesity). Part of this research aims to contribute (in a second phase) to the adequacy of current public policy in the fight against malnutrition and, of course, to the current needs experienced by the SARS-CoV-2 infection contingency. This work develops the application of simple sampling and the stages involved in this statistical tool, whose objective is to establish which part of the reality under study should be studied in order to make inferences about a given population. From the period contemplated between April 28, 2020 and March 8, 2022, the 84 municipalities of the state of Hidalgo reported a total of 86,124 confirmed cases of SARS-CoV-2 infection, from which a sample size of 1,054 subjects has been calculated (representativeness of 91.35% of the target population). The correct application of mathematics in the context of health should allow us to enjoy good health, especially if these results are focused on the promotion and prevention of diseases and their complications; mathematics has surpassed the frontiers of knowledge in various areas and its implementation in this case with respect to public policy and nutrition.

Keywords: cases, diseases, infections, malnutrition and public policy.

1. Introduction

Health problems have a multifactorial character that allows science, society, health professionals and other areas to contribute their multidisciplinary and transdisciplinary perspectives (Salas-Perea, 2003) in the search for strategies to combat diseases, which require compliance with ethical, social, economic and scientific aspects (Cortés et al., 2020).

Malnutrition (which includes obesity, overweight and desnutrition) represents a serious health problem that not only has biological repercussions, unfortunately Mexico faces the consequences of these diseases because it is the first place in overweight and obesity in adults and children, although undernutrition has not been fought either.

As the quarantine period ascended due to SARS-CoV-2 infection, social distancing and isolation, generated negative changes in healthy eating; body weight and body mass index increased, which requires informing people about proper nutrition management and the importance of regular exercise (Ateş & Yeşilkaya, 2021).

It has been described that the high risk of severe manifestations and mortality due to SARS-Cov-2 infection is presented mainly by patients with chronic underlying diseases (although they have also been reported in any age, without previous comorbidities), such as cardiovascular disease, diabetes, chronic kidney disease, obesity (Antezana Llaveta & Arandia-Guzmán, 2020), arterial hypertension and immunosuppression (lymphomas, active tumors or under chemotherapy regimen) (Zetina-Tun & Careaga-Reyna, 2022).

In April 2022, the state of Hidalgo ranked ninth in national mortality, with a rate of 272 deaths per 100,000 inhabitants; 3 confirmed cases per 100,000 inhabitants (population size: 3,086,414) and a cumulative 93,111 confirmed cases related to SARS-CoV-2 infection (Secretaría de Salud, 2021).

Long-term complications of this infection are described, including altered insulin sensitivity, pancreatic islet damage with decreased insulin secretion, muscle weakness and atrophy with altered exercise capacity, changes in body composition with increased fat mass and elevated triglycerides and circulating fatty acids, which could ultimately lead to increased risk of future cardiovascular events (Ayres, 2020).

Various investigations in the world and national literature continue to provide valuable information on this historical

pandemic event, but none specifically has characterized the population of Hidalgo in relation to malnutrition as a risk factor for this infection, so this study is considered of great impact for society and its government.

A universe or population is the set of total elements that make up the interest of an analysis and on which inferences and conclusions are made (López-Roldán & Fachelli, 2017).

In this context, the objective is to choose the size of the representative sample of the universe generated by the 84 municipalities of Hidalgo, corresponding to the subjects confirmed with SARS-CoV-2 coronavirus infection, using the simple sampling technique having as reference the state database belonging to the state of Hidalgo and considering a given period of time.

The usefulness of a representative sample allows the study subjects to have the same opportunity to be chosen and therefore, to be included in a study, achieving that the researcher extrapolates and extends his/hers results to a given population, understanding that those selected are a numerical representation of the universe from which they come (Otzen & Manterola, 2017).

The hypothesis of this exercise is that the greater the reduction of the dimension of the universe studied, the greater the understanding of the phenomenon under study.

Understanding sampling as a scientific research tool whose objective is to determine that part of the population worthy of study (Hernández & Carpio, 2019), feeds a transcendental part in the research exercise of the next phase of this publication called: evaluation of public policy in relation to malnutrition as a risk of SARS-CoV-2 in Hidalgo, describing the hypothesis that malnutrition is an element that influences the mechanics of the disease, with the vision of obtaining the necessary information to analyze, study and evaluate the current policy in the field of nutrition and food, highlighting that illness and death affect the family economy, that of health systems and that of governments.

2. Method

Sampling makes it possible to analyze fragments of a phenomenon with the advantages of reduced costs and more accurate, faster, flexible and more supervised results. Simple sampling is a method of selecting n units in a set of N so that each of the NCn different samples has the same possibility of being elected. In practice, random sampling is performed unit by unit, that is, the units from 1 to N are listed, then a series of n random numbers between 1 and N is extracted, because through a computer program (R, Python or Julia) a table of random numbers is created, where each extraction is chosen randomly, the units that carry these n numbers constitute the sample (Cochran, 1977).

The sample size, a guide to the follow-up of a certain procedure described below (Portela & Villeta, 2007).

Stage 1. Approach to the problem (in which the phenomenon to be studied is identified, raising all the characteristics that encompass it).

Stage 2. Sample frame (a list of elements that make up the population of the phenomenon under study, known as sample units, is outlined).

Stage 3. Selection of the sampling technique (from a sample frame, the ideal technique is chosen to estimate the sample size).

Stage 4. Sample size (based on the sampling technique, the sample size and its proportional distribution for each of its elements are calculated).

Stage 5. Feasibility of the sample size (which means determining the degree of reliability of the sampling).

2.1 Sample Frame

The complexity of the universe under study, due to the large amount of data emanating from it, requires the selection of a sample, which reduces the use of resources such as financial, human, material and intangible resources such as time. By simplifying the size of the population from which we wish to analyze a series of variables, the time in which data are generated that contribute to a more accurate knowledge of a phenomenon, its behavior and prevention in terms of health, is compromising; the pandemic has given us several lessons on the right or wrong actions of governments and their effect on citizens; numbers have that power.

The size of the reported population corresponds to 86,124 subjects, confirmed with SARS-CoV-2 infection, according to the state database, collected thanks to the Epidemiology area of the State Health Secretariat (Table 1).

Table 1. Confirmed cases of SARS-CoV-2 by municipality in the state of Hidalgo

Municipalities	Registered cases
1. Acatlán	26
2. Acaxochitlán	2
3. Actopan	2,506
4. Agua Blanca de Iturbide	0
5. Ajacuba	214
6. Alfajayucan	132
7. Almoloya	96
8. Apan	1,851
9. El Arenal	24
10. Atitalaquia	231
11. Atlapexco	93
12. Atotonilco el Grande	172
13. Atotonilco de Tula	1,279
14. Calnali	156
15. Cardonal	174
16. Cuautepec de Hinojosa	495
17. Chapantongo	270
18. Chapulhuacan	48
19. Chilcuautla	117
20. Eloxochitlán	25
21. Emiliano Zapata	261
22. Epazoyucan	3
23. Francisco I. Madero	25
24. Huasca de Ocampo	57
25. Huautila	0
26. Huazalingo	75
27. Huehuetla	289
28. Huejutla de Reyes	2,275
29. Huichapan	1,602
30. Ixmiquilpan	1,831
31. Jacala de Ledezma	78
32. Jaltocán	14
33. Juárez Hidalgo	5
34. Lolotla	29
35. Metepec	367
36. San Agustín Metzquititlán	49
37. Metztitlán	208
38. Mineral del Chico	37
39. Mineral del Monte	266
40. La Misión	33
41. Mixquiahuala de Juárez	1,682
42. Molango de Escamilla	384
43. Nicolás Flores	47
44. Nopala de Villagrán	225
45. Omitlán de Juárez	80
46. San Felipe Orizatlán	1

47. Pacula	44
48. Pachuca de Soto	35,433
49. Pisaflores	41
50. Progreso de Obregón	30
51. Mineral de la Reforma	1,433
52. San Agustín Tlaxiaca	4
53. San Bartolo Tutotepec	99
54. San Salvador	52
55. Santiago de Anaya	75
56. Santiago Tulantepec de Lugo de Guererero	1,900
57. Singilucan	0
58. Tasquillo	225
59. Tecozautla	206
60. Tenango de Doria	581
61. Tepeapulco	4,153
62. Tepehuacán de Guerrero	108
63. Tepeji del Río de Ocampo	3,299
64. Tepetitlán	79
65. Tetepango	60
66. Villa de Tezontepec	64
67. Tezontepec de Aldama	151
68. Tianguistengo	43
69. Tizayuca	5,523
70. Tlahuelilpan	127
71. Tlahuiltepa	51
72. Tlanalapa	76
73. Tlanchinol	165
74. Tlaxcoapan	1,492
75. Tolcayuca	173
76. Tula de Allende	5,360
77. Tulancingo de Bravo	5,996
78. Xochiatipan	6
79. Xochicoatlán	77
80. Yahualica	0
81. Zacualtipán de Ángeles	641
82. Zapotlán de Juárez	61
83. Zempoala	72
84. Zimapán	390
Total	86,124

Note. Period contemplated from April 28, 2020 to March 8, 2022; personal elaboration.

2.2 Selection of the Sampling Technique

Assuming that the target population is finite (since the total number of observation units that compose it is known), we have that (Aguilar-Barojas, 2005):

$$n = \frac{N * Z_{\alpha}^{2} PQ}{E^{2}(N-1) + Z_{\alpha}^{2} PQ} \quad (1)$$

Where:

- n = Sample size.
- N = Total population size.
- Zα = Confidence level at 0.95 and with a significance level at 0.05. Below the curve of the normal distribution is 1.96.
- P = Probability of success.
- Q = (1 – P) = Probability of failure.
- E = Error admitted in the sample.

It is important to clarify that *N* is the 86,124 subjects and *n*, the revealing sample size calculation; *P*, explains the possibility of being selected as part of the sample and that *Q* is the probability of not being selected (or known as failure), so it assigns 50% versus 50% (0.5+0.5=1); that is, both P and Q have the same probability of being selected.

Its main estimators are the following (Pérez, 2005):

- Sample size by item:

$$n_i = \left(\frac{N_i}{N}\right) * n\,; \quad i = 1,2,3,\ldots,k \tag{2}$$

Where:

- N_i = Any of the states, i.e., the size of the population of each municipality.
- N = Total population size.
- n = Sample size.
- k =1,2,3,4...k As the total number of municipalities.
- Estimator of the total of the sample:

$$Y = n\bar{Y} \tag{3}$$

Where:

$\overline{Y}$ = Population size of each municipality.

Sample meaner:

$$\bar{Y} = \sum_{i=1}^{K} \frac{Y_i}{n} \tag{4}$$

Σ^k = The average number of objects within each sample, starting from municipality 1 to minucipality 84.

Y_i = Sample size in any municipality.

- Confidence intervals:

$$\overline{Y} - (Z_\alpha)\left(\sqrt{Var(\overline{Y})}\right) < \overline{Y} < \overline{Y} + (Z_\alpha)\left(\sqrt{Var(\overline{Y})}\right) \tag{5}$$

- Expansion factor (Ackoff & Sasieni, 1977):

$$Fx = \frac{N}{n} \tag{6}$$

- Variance of the sample:

$$S^2 = \frac{\sum_{i=1}^{K}(Y_i - \overline{Y})^2}{n-1} \tag{7}$$

- Variance of the average:

$$Var(\overline{Y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right) \tag{8}$$

- Absolute error:

$$Ea(\overline{Y}) = \left(\frac{\sqrt{Var(\overline{Y})}}{\overline{Y}}\right) * 100 \tag{9}$$

- Degree of adjustment:

$$\mathrm{Gr} = 100 - \mathrm{Ea}(\overline{\mathrm{Y}}) \tag{10}$$

For the calculation of the sample and its estimators, we start from a confidence level of 0.95 and a significance level of 0.05, with an error of 0.03.

2.3 Sample Size

Based on the algebraic expression (3) and based on the following data:

- N = 86124
- Zα= 1.96
- P = 0.50
- Q = 0.50
- E = 0.03

Substituting in the algebraic expression (1):

$$\mathrm{n} = \frac{\mathrm{N}*\mathrm{Z}_{\alpha}^{2}\mathrm{PQ}}{\mathrm{E}^{2}(\mathrm{N}-1)+\mathrm{Z}_{\alpha}^{2}\mathrm{PQ}} = \frac{86{,}124+(1.96)^{2}(0.50)(0.50)}{(0.03)^{2}(86{,}124-1)+(1.96)^{2}(0.50)(0.50)} \tag{11}$$

Therefore, the sample size is:

$$\mathrm{n} = \frac{82{,}713.4896}{78.4711} = 1054.1 \sim 1054 \tag{12}$$

Based on the sample size and applying the algebraic expression (6), the results are shown in Table 2 below.

Table 2. Sample size of confirmed SARS-CoV-2 cases in each municipality of the state of Hidalgo.

Municipalities	Sampling
1. Acatlán	0
2. Acaxochitlán	0
3. Actopan	31
4. Agua Blanca de Iturbide	0
5. Ajacuba	3
6. Alfajayucan	2
7. Almoloya	1
8. Apan	23
9. El Arenal	0
10. Atitalaquia	3
11. Atlapexco	1
12. Atotonilco el Grande	2
13. Atotonilco de Tula	16
14. Calnali	2
15. Cardonal	2
16. Cuautepec de Hinojosa	6
17. Chapantongo	3
18. Chapulhuacan	1
19. Chilcuautla	1
20. Eloxochitlán	0
21. Emiliano Zapata	3
22. Epazoyucan	0
23. Francisco I. Madero	0
24. Huasca de Ocampo	1
25. Huautila	0

26. Huazalingo	1
27. Huehuetla	4
28. Huejutla de Reyes	28
29. Huichapan	20
30. Ixmiquilpan	22
31. Jacala de Ledezma	1
32. Jaltocán	0
33. Juárez Hidalgo	0
34. Lolotla	0
35. Metepec	4
36. San Agustín Metzquititlán	1
37. Metztitlán	3
38. Mineral del Chico	0
39. Mineral del Monte	3
40. La Misión	0
41. Mixquiahuala de Juárez	21
42. Molango de Escamilla	5
43. Nicolás Flores	1
44. Nopala de Villagrán	3
45. Omitlán de Juárez	1
46. San Felipe Orizatlán	0
47. Pacula	1
48. Pachuca de Soto	434
49. Pisaflores	1
50. Progreso de Obregón	0
51. Mineral de la Reforma	18
52. San Agustín Tlaxiaca	0
53. San Bartolo Tutotepec	1
54. San Salvador	1
55. Santiago de Anaya	1
56. Santiago Tulantepec de Lugo de Guererero	23
57. Singilucan	0
58. Tasquillo	3
59. Tecozautla	3
60. Tenango de Doria	7
61. Tepeapulco	51
62. Tepehuacán de Guerrero	1
63. Tepeji del Río de Ocampo	40
64. Tepetitlán	1
65. Tetepango	1
66. Villa de Tezontepec	1
67. Tezontepec de Aldama	2
68. Tianguistengo	1
69. Tizayuca	68
70. Tlahuelilpan	2
71. Tlahuiltepa	1
72. Tlanalapa	1
73. Tlanchinol	2

74.	Tlaxcoapan	18
75.	Tolcayuca	2
76.	Tula de Allende	66
77.	Tulancingo de Bravo	73
78.	Xochiatipan	0
79.	Xochicoatlán	1
80.	Yahualica	0
81.	Zacualtipán de Ángeles	8
82.	Zapotlán de Juárez	1
83.	Zempoala	1
84.	Zimapán	5
	Total	1,054

Note. Random selection, personal elaboration.

Calculating the variance of the sample:

$$S^2 = \frac{\sum_{i=1}^{K}(Y_i - \overline{Y})^2}{n-1} = 89.256 \quad (13)$$

Based on the above, the variance of the sample mean:

$$\mathrm{Var}(\overline{Y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{188.71}{1054}\left(\frac{86{,}124-1054}{86{,}124}\right) \quad (14)$$

$$\mathrm{Var}(\overline{Y}) = \frac{(188.71)(0.9878)}{1{,}054} = 1.1768 \quad (15)$$

By Calculating your confidence interval, you have to:

- S2 = 188.71.
- $Z\alpha$ = 1.96.
- $\overline{Y}$ = 12.55.
- $\sqrt{\mathrm{Var}(\overline{Y})}$ = 1.085.

Replacing:

$$12.55 - (1.96)(1.085) < \overline{Y} < \overline{Y} + 12.55 + (1.96)(1.085) \quad (16)$$

Such that:

$$875 < \overline{Y} < 1232 \quad (17)$$

With a confidence level of 0.95 and a significance level of 0.05, the sample size will range between 875 and 1232.

Feasibility of sample size

Based on the size of the population and the sample, the expansion factor of the selected units is 82, that is:

$$\frac{N}{n} = \frac{86{,}124}{1{,}054} = 81.71(18)$$

Each individual who is randomly selected has the ability to answer 82 individuals in the population.

Calculating the relative error from the algebraic expression (9):

$$\mathrm{Ea}(\overline{Y}) = \left(\frac{1.085}{12.55}\right) * 100 = 8.65\% \quad (19)$$

Obtained the degree of adjustment:

$$Gr = 100- Ea(\bar{Y})=100 - 8.65 = 91.35\ \% \quad (20)$$

3. Results

The universe constituted by the 84 municipalities of the state of Hidalgo, corresponds to the subjects confirmed with SARS-CoV-2 coronavirus infection, in a period contemplated between April 28, 2020 and March 8, 2022, according to the database of the Epidemiology area of the State Health Secretariat.

Regarding the calculation of the sample and its estimators, a confidence level of 0.95 and a significance level of 0.05 were used, with an error of 0.03, which, after calculation, yields a sample size of 1,054 subjects.

By applying the simple random sampling technique, the individuals in the study population will all have the same probability of being selected.

Regarding the expansion factor, this describes that each randomly selected subject has the power to respond to 82 individuals in the population, therefore, "it does not matter which subject is chosen", as long as it meets the selection criteria of the population to be sampled, i.e., belonging to the state records referred to above and also includes those subjects for whom there is a total number of responses for each of the variables of interest for the study.

The relative error of 8.65% describes that this percentage of the selected sample would not provide the relevant information for the study. Finally, the sample size can range between 875 and 1232 subjects, having a representativeness of 91.35% of the target population.

For the purposes of this research, it has been decided to obtain the greatest possible representativeness of the sample, so the upper limit range of subjects under study will be taken, that is, 1232, in addition to not excluding any municipality in the same, adjusting the sampling revealed here (except those municipalities that report with 0 registered cases).

The realization of this sampling implies the construction of two other models where the first one focuses on the reduction of the dimension of the variables and the second one in reference to the estimation of the evolution of the health status of the subjects.

4. Discussion

Simple sampling is a vital tool for the research proposed here; it is necessary to evaluate public policies focused on malnutrition in Mexico and the state of Hidalgo, in order to obtain the necessary information to optimize, propose and act on real proposals, generated precisely in the heart of this population.

In Mexico, a double burden of disease effect has been described, where poor diet quality is responsible for both obesity and malnutrition (Barquera et al., 2001).

Throughout the history of our country, different policies and proposals have left an incentive for favorable changes, however, none alone has managed to combat the problems of malnutrition that to date have worsened and moved to a state of chronicity, which imposes the current challenges to which are added adversities such as the current pandemic.

Nutritional status and diet are determinants of health and, in the case of SARS-CoV-2, could play a transcendental role in the prevention and development of complications, since, in recent decades, there have been multiple changes in the dietary patterns of Mexican families. Together with sociocultural transformations, product of a globalized economic model, occupations and physical activity habits have been modified, people have become more sedentary and devote much of their time to television and device screens (Castro et al., 2020).

Despite the fact that 1 in 5 people who contract SARS-CoV-2 ends up presenting a severe picture and experiencing respiratory difficulties, it is known that in our country a large part of the population is within the risk group not necessarily because they are older people, but because of the presentation of previous medical conditions such as arterial hypertension, cardiac or pulmonary problems, diabetes or cancer (Rivas et al., 2020).

Poor metabolic control, together with an elevated body mass index or excess adipose tissue, appear to be risk factors for SARS-CoV-2 complications. Prevention is mainly based on the promotion of healthy habits and the effective and persistent control of these behaviors. In this population, inadequate or insufficient nutrition may result in increased susceptibility to infection (Vecilla et al., 2020).

Our country is facing a pandemic that puts at risk the advances in social development and with it the health of the population, which are affecting various sectors of the population to a greater or lesser degree, some due to economic deprivation, others due to their health status, loss of employment, food insecurity, educational deficiencies, inequity and lack of equality, in addition to a long list of social afflictions. The question is, How long will it take for the country to recover, or even to think about whether or not it will be able to overcome the situation it is going through?; isolated solutions are unthinkable, coordination between sectors is needed.

5. Conclusion

The proper application of mathematics in the context of health should allow us to enjoy good health and therefore a positive economic impact (Serrano et al., 2020), especially if these results were focused on the promotion and prevention of diseases and their complications. The research and execution of medicine in conjunction with mathematics has contributed to the knowledge of risk factors and the way in which various pathologies behave (Olmedo & Ariza, 2012), including the chronic conditions highlighted here. For example, the use of mathematical models makes it possible to pose and test hypotheses about the use of certain treatments or to personalize therapies, run simulations and predict the behavior of human biology (Pérez-García et al., 2016).

The initial objective of choosing a representative sample size has been achieved; mathematics has surpassed the frontiers of knowledge in different areas and its application in this case to public policy, medicine and nutrition; the mathematical models of the second phase will be a fundamental part in the continuity of this research that in the near future will characterize the population of Hidalgo in terms of malnutrition and SARS-CoV-2 infection, factor analysis and multidisciplinary scaling are made visible.

Acknowledgment

To the Epidemiology area of the Secretariat of Health of the State of Hidalgo.

References

Ackoff, R., & Sasieni, M. (1977). *Fundamentos de investigación de operaciones* (1a ed.). Limusa. http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=librosslp.xis&method=post&formato=2&cantidad=1&expresion=mfn=000936

Aguilar-Barojas, S. (2005). Fórmulas para el cálculo de la muestra en investigaciones de salud. *Salud En Tabasco*, *11*(1–2), 333–338. http://www.redalyc.org/articulo.oa?id=48711206

Antezana Llaveta, G., & Arandia-Guzmán, J. (2020). SARS-CoV-2: estructura, replicación y mecanismos fisiopatológicos relacionados con COVID-19. *Gaceta Médica Boliviana*, *43*(2), 172–178. https://doi.org/10.47993/gmb.v43i2.85

Ateş, B., & Yeşilkaya, B. (2021). Adverse Effect of Emotional Eating Developed during the COVID-19 Pandemic on Healthy Nutrition, a Vicious Circle: A cross-sectional descriptive study. *Revista Española de Nutrición Humana y Dietética*, *25*(2), 1–25. https://doi.org/10.14306/renhyd.25.S2.1144

Ayres, J. (2020). A metabolic handbook for the COVID-19 pandemic. *Nature Metabolism*, *2*(7), 572–585. https://doi.org/10.1038/s42255-020-0237-2

Barquera, S., Rivera-Dommarco, J., & Gasca-García, A. (2001). Food and nutrition policies and programs in Mexico. *Salud Publica de Mexico*, *43*(5), 464–477. https://doi.org/10.1590/s0036-36342001000500011

Castro, I., Flores, E., & Ochoa, H. (2020). Malnutrición y covid-19. *Ecofronteras*, *24*(69), 22–24. https://revistas.ecosur.mx/ecofronteras/index.php/eco/article/view/1917

Cochran, W. (1977). *Técnicas de Muestreo*. CECSA.

Cortés, M., Mur, N., Iglesias, M., & Cortés, M. (2020). Algunas consideraciones para el cálculo del tamaño muestral en investigaciones de las Ciencias Médicas. *MediSur*, *18*(5), 937–942. https://www.medigraphic.com/pdfs/medisur/msu-2020/msu205x.pdf

Hernández, C., & Carpio, N. (2019). Introducción a los tipos de muestreo. *Revista ALERTA*, *2*(1), 76–79. https://doi.org/10.5377/alerta.v2i1.7535

López-Roldán, P., & Fachelli, S. (2017). El diseño de la muestra. In *Metodología de la Investigación Social Cuantitativa: Vol. 1° edición* (1a ed., pp. 19–43). Universidad Autónoma de Barcelona. https://ddd.uab.cat/pub/caplli/2017/185163/metinvsoccua_cap2-4a2017.pdf

Olmedo, V., & Ariza, R. (2012). Matemáticas en medicina: una necesidad de capacitación. *Medicina Interna de México*, *28*(3), 278–281. https://www.medigraphic.com/pdfs/medintmex/mim-2012/mim123l.pdf

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. *International Journal of Morphology*, *35*(1), 227–232. https://doi.org/10.4067/S0717-95022017000100037

Pérez-García, V. M., Fitzpatrick, S., Pérez-Romasanta, L. A., Pesic, M., Schucht, P., Arana, E., & Sánchez-Gómez, P. (2016). Applied mathematics and nonlinear sciences in the war on cancer. *Applied Mathematics and Nonlinear Sciences*, *1*(2), 423–436. https://doi.org/10.21042/amns.2016.2.00036

Pérez, C. (2005). *Muestreo Estadístico. Conceptos y Problemas Resueltos*. Pearson Prentice Hall.

https://es.scribd.com/document/379739868/Perez-Lopez-Cesar-Muestreo-Estadistico-Conceptos-Y-Problemas-Resueltos-pdf

Portela, J., & Villeta, M. (2007). *Técnicas básicas de Muestreo con SAS* (1a ed.). Universidad Computense de Madrid. https://eprints.ucm.es/id/eprint/47107/2/Técnicas básicas de muestreo con SAS. J. Portela%2C M. Villeta.pdf

Rivas, J., Callejas, R., & Nava, D. (2020). COVID-19: Desafíos y estrategias para el sistema de salud mexicano. In *Factores críticos y estratégicos en la interacción territorial. Desafíos actuales y escenarios futuros* (pp. 565–582). Universidad Nacional Autónoma de México y Asociación Mexicana de Ciencias para el Desarrollo Regional A.C, Coeditores,. http://ru.iiec.unam.mx/5070/1/1-195-Rivas-Callejas-Nava.pdf

Salas-Perea, R. (2003). La identificación de necesidades de aprendizaje. *Revista Cubana de Educación Médica Superior*, *17*(1), 25–38. http://scielo.sld.cu/pdf/ems/v17n1/ems03103.pdf

Secretaría de Salud. (2021). *Dirección General de Epidemiología*. https://datos.covid-19.conacyt.mx/

Serrano, B., Aguilar, C., Cervantes, A., Molina, M., & Trujillo, V. (2020). Las matemáticas aplicadas como una oportunidad para preservar la salud. *Revista Conrado*, *16*(75), 272–279.

Vecilla, J., Torres, Y., Beltrán, J., Sánchez, S., Yépez, G., Arias, R., Charcopa, V., & Morales, M. (2020). Importancia del control del control nutricional en los pacientes diabéticos durante la pandemia de COVID-19. *Diabetes Internacional y Endocrinología*, *XII*(1), 39–43. https://doi.org/10.5281/zenodo.4381065

Zetina-Tun, H., & Careaga-Reyna, G. (2022). Infección por SARS-CoV-2 en pacientes trasplantados de corazón. Experiencia en México. *Cirugía Cardiovascular*, *29*(1), 21–24. https://doi.org/10.1016/j.circv.2021.06.005

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Review of Copula for Bivariate Distributions of Zero-Inflated Count Time Series Data

Dimuthu Fernando[1], Mohammed Alqawba[2], Manar Samad[3], Norou Diawara[1]

[1.] Department of Mathematics & Statistics, College of Sciences, Old Dominion University, Norfolk VA, USA

[2.] Department of Mathematics, College of Science and Arts, Qassim University, Ar Rass 51452, Saudi Arabia

[3.]Dept. of Computer Science, Tennessee State University, Nashville, TN, USA

Correspondence: Norou Diawara, Department of Mathematics & Statistics, College of Sciences, Old Dominion University, Norfolk VA, USA

Received: September 8, 2022 Accepted: October 9, 2022 Online Published: October 30, 2022

doi:10.5539/ijsp.v11n6p52 URL: https://doi.org/10.5539/ijsp.v11n6p52

Abstract

The class of bivariate integer-valued time series models, described via copula theory, is gaining popularity in the literature because of applications in health sciences, engineering, financial management and more. Each time series follows a Markov chain with the serial dependence captured using copula-based distribution functions from the Poisson and the zero-inflated Poisson margins. The copula theory is again used to capture the dependence between the two series.

However, the efficiency and adaptability of the copula are being challenged because of the discrete nature of data and also in the case of zero-inflation of count time series. Likelihood-based inference is used to estimate the model parameters for simulated and real data with the bivariate integral of copula functions. While such copula functions offer great flexibility in capturing dependence, there remain challenges related to identifying the best copula type for a given application. This paper presents a survey of the literature on bivariate copula for discrete data with an emphasis on the zero-inflated nature of the modelling. We demonstrate additional experiments on to confirm that the copula has potential as greater research area.

Keywords: count time series, copula, Zero-Inflated, count data, Poisson distribution

Subject Classification: 62H05, 62H10

1. Introduction

In the study of multivariate distributions, copula functions are gaining popularity in recent years. They are attractive as they can handle internal and mutual dependences among variables. The copula was first introduced in the Sklar (1954) paper, a paper that Frechet helped publish. Hoffding (1940) is also credited for almost innovating the concept of copula. Many problems in practical situations are modeled under related distributions using copula functions, in contrast to classical multivariate (Gaussian) distributions for count data. As such, the literature shows a growing interest in the investigation of dependence for sequences of counts in time series cases. The simplest of such sequences are bivariate count time series data. Copula functions have gained popularity in building such bivariate and multivariate distributions as the desire to understand the structure in massive time series count data is becoming more common. For diseases and rare events, observed counts over time appear in a high frequency of zeros (zero inflation), which is discussed in MÖller et al. (2020) and Young et al. (2020).

Sklar (1959) introduced a method to build in the bivariate and multivariate distributions for two random variables. The idea of joint distribution, especially in the bivariate case can be traced back to Frechet (1951, 1956, 1958). Morgensetrn (1956), Plackett (1965), Farlie (1969) and many other authors could be included in this systematic approach of constructing bivariate distributions with specific marginals and different dependence measures. See examples such as Gumbel (1958) or Johnson and Tenenbein (1981). In that same line of thought, Cook and Johnson (1981) asked two questions that are still of relevance. The questions are: 1) “Is there a distribution that appears to be the most promising candidate for non-normal types of data?” 2) “Is the resulting distribution or model fit significantly better than that obtained from the multivariate normal distribution?”

Finding a unique copula for a joint distribution requires one to know the form of the joint distribution. When using copula, one can separately model the marginal distributions and the dependence structure, which makes the copula

approach unique. Choosing the appropriate copula for a particular scenario means finding the one that best captures the dependence in data. Many variants of copulas have been proposed in the literature where each of these is suitable for different dependence structures. For example, Gaussian copula is flexible, and it allows for equally positive and negative dependence. The Clayton copula cannot account for negative dependence, and it exhibits strong left tail dependence. Similar to Gaussian copula, Frank copula allows for both positive and negative dependence between the marginals.

Copulas offer a flexible framework to combine distributions. It is unique if marginal densities are continuous. However, if some of the marginal distributions are discrete, the unicity cannot be obtained automatically.

Many copula functions have been identified, from the extreme of independent variables (the so-called independent copula or the product) to the max or min copula. The dependence is then captured by a selection of parameters and criteria associated with the range and properties of model parameters.

Moreover, high dimensional copulas have been introduced via bivariate copulas, under different decompositions and structures. These structures are known as the canonical vine (C-vine) or drawable vine (D-vine). References to C and D vines can be found in Bedford and Cooke (2002), Joe et al. (2010), and Aas et al. (2009). Gräler (2014) proposed the convex combination of bivariate copula densities incorporating the distance [between what?] as a parameter in the spatial setting. The application of copula functions can be found in finances (Czado et al, 2012), hydrology (Yu et al., 2020), transportation (Irannezhad et al., 2017), health care (Shi and Zhang, 2015), and more. The Farlie-Gumbel-Morgenstern (FGM) family of copula can be used to establish relationship between predictors (Durante and Sempi, 2016)).

Within the count time series, if we look at the binary data, there is a growing interest in the description of multivariate distributions under pair copulas (Lin and Chaganty, 2021). Panagiotelis et al. (2012) presented pair copula constructions for discrete multivariate data. Their algorithm is explained as a product of bivariate pair copula, demonstrating the great potential of vine copula approaches. They stated that the model selection for C or D vine remains an important open problem, with a particular emphasis on the conditional independence identification (Czado, 2019, Deng and Chaganty, 2021). From there, the idea of using the D vine for modeling counts with excess zeros and temporal dependence is presented in Sefidi et al. (2020). Perrone and Durante (2021) highlighted the link between the extreme discrete copula and mathematical concept of convex polytope, which is an idea spinning from the class of bivariate distributions (Rao and Subramanyam (1990).

There are numerous problems and interesting challenges related to time series of counts. Davis et al. (2016, 2021) presented extensive literature and many examples of count time series. Fokianos (2021) and Armillota and Fokianos (2021) presented a Poison network autoregression for counts. In the statistical process control, Fatahi et al. (2012) proposed the monitoring of rare events under the copula based bivariate zero-inflated Poisson. van Den Heuvel et al. (2020) proposed corrections to such results adding the negative correlation option.

With these studies and observations in mind, this paper presents reviews and updates related to the copula for bivariate distributions of zero-inflated count time series and highlights research directions. Motivated by multivariate datasets acquired using correlation structures, our goal is to review the bivariate count and zero-inflated count time series for inference and application purposes under copula modeling. We give some insights into the bivariate count copula and its recent developments. We organize our discussion as follows. In Section 2, copulas for discrete count and zero-inflation of discrete count time series data are described. The use of univariate and bivariate copula for discrete data is discussed in Section 3. Extensions of discrete bivariate copulas are described in Section 4. We conclude this paper with an extended discussion on future work.

2. Copula for Zero-inflated of Discrete and Count Time Series Data

This section introduces the general form for multivariate copula, and its Gaussian representation. We also give an explicit definition of the zero inflated counts time series data.

2.1 Simple Gaussian Copula Example

Masarotto and Varin (2012) introduced a Gaussian copula model which can be used to model time series data in the presence of covariates. The corresponding regression model can be written as follows.

$$\mathrm{Y_t} = \mathrm{g}(X_t, \epsilon_\mathrm{t}\, \theta), \text{for t} = 1, .., \mathrm{n},$$

where $g(.)$ is a function of the covariates X_t and ϵ_t, which capture the serial dependence. The parameter $\boldsymbol{\theta}$ is a vector of marginal regression coefficients. The joint distribution function of the time series $\{Y_t\}\, for\, t = 1, \dots, n$ can be constructed using the Gaussian copula as follows.

$$F(y_1, y_2, \dots, y_n) = P(Y_1 \le y_1, Y_2 \le y_2, \dots, Y_n \le y_n) = \Phi_{R(\rho)}(\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2)) \dots, \Phi^{-1}(F_n(y_n))) \quad (1)$$

Here, Φ^{-1} is the inverse CDF of standard normal distribution, and $\Phi_{R(\rho)}$ is the joint CDF of a multivariate normal distribution with a mean vector of zeros and covariance matrix *R*.

2.2 Review of Copula for Discrete Data

Copula distributions are becoming increasingly popular in many areas of statistical data sciences. For example, in engineering, copula distributions are used to model the shear fore for cantilever beams and for beams with multiple point loads (Zhang and Lam, 2016). In pharmaceutical quality control, two correlated characteristics sample data are presented in Fatahi et al. (2012). The authors describe the bivariate Poisson distribution with the evidence of zero-inflation. Sukparungsee et al. (2021) developed a bivariate copula for control chart effectiveness. They show the bivariate copula distribution on Hotelling's T^2 over the multivariate cumulative sum for positive, negative, weak, moderate, and strong correlations when the assumption of multivariate normality is violated. Van den Heuvel et al. (2020) extended the idea from Fatiha et al. (2012) and included negative correlation case, and an upper control limit on the sum of bivariate random variables. Copulas are elegantly captured in the Genest and MacKay (1986), Genest (1987) and also in Han and De Oliveira (2016 and 2020), among others. In the financial sector, a recent work by Nikoloulopoulos and Moffatt (2019) reminds us of the need to study dependence structures. There are also more general ambitions for the bivariate copula from a bigger perspective than we expect to show the aggregated effects in many other areas.

The list of copula functions is very large. The work of Größer and Okhrin (2021) presents a summary of bivariate copula followed by the construction of multivariate copula using pair copula decompositions. They provide examples for each copula family and provide an overview of how copula theory can be used in various fields of data science.

Yang et al. (2014) proposed the Ali-Mikhael-Haq (AMH) copula-based function to investigate the joint risk probabilities of rainstorms, wind speeds, and storm surges. The proposed model was developed to assess the impact based on marginal distributions of maximum daily rainfall and extreme gust velocity. Alqawba et al. (2021) constructed a class of bivariate integer-valued time series models using copula theory. Applying either the bivariate Gaussian copula or the bivariate t copula functions, they jointly modeled two copula-based Markov time series models. They applied their method on bivariate count time series data, where the marginals follow either a Poisson or zero-inflated Poisson distribution.

Safari et al. (2020) proposed a bivariate copula regression model to analyze cervical cancer data. They applied a bivariate copula to model and estimate joint distribution parameters. Nikoloulopoulos and Moffatt (2019) used bivariate copulas to jointly model bivariate ordinal time-series responses with covariates for risks assessment of married couples. They proposed a copula-based Markov modelling of ordinal time-series responses and used another copula to couple their conditional (on the past) distributions at each time point. Copula families such as the Bivariate normal (BVN), Frank, Gumbel and bivariate t-copula were used to model the univariate time series as well as to couple them together.

The work of Nikoloulopoulos & Karlis (2010) presents a regression copula-based model where covariates are used not only for the marginal but also for the copula parameters. They measured the effect of covariates on dependence structure by building a fully parametric copula-based model while considering six one-parameter copula families, namely Frank, Galambos, Gumbel, Mardia–Takahasi (M-T), and normal to build the dependence structure.

Karlis & Pedeli (2013) presented a bivariate integer-valued autoregressive process (BINAR(1)) in which the cross-correlation was modeled using a copula to accommodate both positive and negative correlation. They presented an application of the Frank and Gaussian copula to model dependence, and marginal time series were modeled using Poisson and negative binomial INAR(1) distributions.

Ma et al. (2020) proposed a copula approach utilizing a Gaussian copula with random effects to model correlated bivariate count data regression.

2.3 The Zero-Inflated Discrete Data

Zero inflation models can be found in many studies from Lambert (1992) to Hall (2000) and recently in Rigby et al. (2019). The zero-inflated count regression models are described as follows.

- Zero-Inflated Poisson (ZIP) Distribution (Lambert, 1992):

$$F_{Y_t}(m) = \omega_t + (1-\omega_t)e^{-\lambda_t}\sum_{y_t=0}^{m}\frac{\lambda_t^{y_t}}{y_t!}. \quad (2)$$

- Zero-Inflated Negative Binomial (ZINB) Distribution (Ridout et al, 2001):

$$F_{Y_t}(m) = \omega_t + \frac{(1-\omega_t)}{\Gamma(\kappa_t)}\left(\frac{\kappa_t}{\kappa_t+\lambda_t}\right)^{\kappa_t} \sum_{y_t=0}^{m} \frac{\Gamma(\kappa_t+y_t)}{y_t!} \left(\frac{\lambda_t}{\kappa_t+\lambda_t}\right)^{y_t}.$$

- Zero-Inflated Conway-Maxwell-Poisson (ZICMP) Distribution (Sellers and Raim, 2016):

$$F_{Y_t}(m) = \omega_t + \frac{(1-\omega_t)}{Z(\lambda_t+\kappa_t)} \sum_{y_t=0}^{m} \frac{\lambda_t^{y_t}}{(y_t!)^{\kappa_t}},$$

where $\lambda_t = \exp(\boldsymbol{X}_t'\beta), \omega_t = \frac{\exp(\boldsymbol{z}_t'\gamma)}{1+\exp(\boldsymbol{z}_t'\gamma)}$, and $\kappa_t = \exp(\boldsymbol{W}_t'\alpha)$

are the associated covariate vectors affecting the intensity parameter λ_t, the zero-inflation parameter ω_t and the dispersion parameter κ_t, respectively.

The term $\sum_{y_t=0}^{m} \frac{\lambda_t^{y_t}}{(y_t!)^{\kappa_t}}$ is the normalizing function of the CMP.

Different variants of similar regression models have been proposed in the literature. A noteworthy use of copula for zero-inflated data is studied in Shamma et al. (2020), where the inflation is built from a geometric count time series in an integer-valued autoregressive (INAR) process.

3. Univariate and Bivariate Copula Models for Count Time Series Data

3.1 Univariate Copula-Based Model for Count Time Series Data

First order Markov model

Alqawba, & Diawara (2021) introduced a class of Markov zero inflated count time series model where the joint distribution function of the consecutive observations is constructed through copula functions. Suppose $\{Y_t\}$ zero-inflated count time series first order Markov chains the multivariate joint density distribution of $Y_1, Y_2, \ldots, Y_n$ can be constructed as below.

$$Pr(Y_1 = y_1, \ldots, Y_n = y_n) = Pr(Y_1 = y_1) \prod_{t=2}^{n} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1})$$

Using the copula theory, the joint distribution function of Y_t, Y_{t-1} can be written as below.

$F_{12}(y_t, y_{t-1}) = C(F_t(yt), F_{t-1}(y_{t-1}); \boldsymbol{\delta})$ *where* $\boldsymbol{\delta}$ *is bivariate copula parameter vector.*

Hence, we can calculate the transition probability as below.

$$\Pr(Y_t = y_t | Y_{t-1} = y_{t-1}) = \frac{\Pr(Y_t = y_t, Y_{t-1} = y_{t-1})}{f_{t-1}(y_{t-1})}$$

Where

$$Pr(Y_1 = y_1, Y_{t-1} = y_{t-1}) = F_{12}(y_t, y_{t-1}) - F_{12}(y_t - 1, y_{t-1}) - F_{12}(y_t, y_{t-1} - 1) + F_{12}(y_t - 1, y_{t-1} - 1)$$

Likelihood and parameter estimation under first order Markov model

The likelihood function of the first order Markov model is given by

$$L(\vartheta, y) = \Pr(Y_1 = y_1; \theta) \prod_{t=2}^{n} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}; \vartheta) \quad (3)$$

The log likelihood function ($l(\vartheta; y)$) is given by

$$l(\vartheta; y) = \log \Pr(Y_1 = y_1; \theta) + \sum_{t=2}^{n} log \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}; \vartheta)$$

Where θ *and* δ are the parameter vectors of the marginals and the dependence structure, respectively. For the Gaussian copula family, the likelihood function involves a bivariate integral of the normal probability in $C(.;\delta)$ which means that the function is not in a closed form and we need approximations for the rectangle probabilities.

The simulation study was conducted using the **R software** by the **'optim'** function in the **"stats" package.** We simulate first order stationary Markov processes with joint distribution of consecutive observations following the bivariate Gaussian copula. The marginal distributions are chosen to be the Poisson and ZIP distributions. We present the simulation results for a first order Markov model with Poisson marginals. The parameter λ represents the mean of a marginal Poisson, ω is the measure of zero inflation, and δ is the serial dependence associated with time series data.

We found that the estimate of these parameters is fairly stable where the precision increases with increasing sample size. Table 1 and Table 2 show the estimates of copula parameters for positive and negative autocorrelations, respectively. The estimates are described by standard measures of variation, including standard deviation, mean square error and mean absolute error.

Univariate ZI count time series models

For positive serial dependence with λ=3, ω=0.3, δ =0.6

Table 1. Parameter estimates for the univariate ZI Poisson model with positive autocorrelation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda(3)$	2.990	0.347	0.1200	0.282
	$\omega(0.3)$	0.288	0.083	0.0070	0.006
	$\delta(0.6)$	0.577	0.091	0.0080	0.073
300	$\lambda(3)$	3.013	0.192	0.037	0.152
	$\omega(0.3)$	0.293	0.046	0.002	0.037
	$\delta(0.6)$	0.596	0.046	1.433	1.196
500	$\lambda(3)$	3.006	0.154	0.024	0.120
	$\omega(0.3)$	0.295	0.035	0.001	0.028
	$\delta(0.6)$	0.596	0.037	0.001	0.028

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

For negative serial dependence with λ=3, ω=0.3, δ =-0.6

Table 2. Parameter estimates for the univariate ZI Poisson model with negative autocorrelation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda(3)$	3.045	0.280	0.080	0.234
	$\omega(0.3)$	0.299	0.046	0.002	0.036
	$\delta(-0.6)$	-0.618	0.087	0.0070	0.072
300	$\lambda(3)$	3.019	0.152	0.023	0.119
	$\omega(0.3)$	0.298	0.030	0.0007	0.002
	$\delta(-0.6)$	-0.605	0.050	0.003	0.040
500	$\lambda(3)$	3.014	0.112	0.0127	0.009
	$\omega(0.3)$	0.299	0.019	0.0004	0.015
	$\delta(-0.6)$	-0.603	0.040	0.002	0.031

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

Applications

Alqawba & Diawara (2021) applied the proposed model to analyze monthly count of strong sandstorms recorded by the AQI airport station in Eastern Province, Saudi Arabia. The data set consists of 348 monthly counts of strong sandstorms, starting from January 1978 to December 2013. The bar plots suggest that both counts follow Zero inflated Poisson distribution, whereas the ACFs indicate that the counts are serially dependent. Finally, to illustrate the superiority of the proposed method they compare the method with zero-inflated integer-valuedautoregressive (ZIINAR) models.

3.2 Bivariate Copula-Based Model for Count Time Series Data

Copula based bivariate model

Suppose we have $\{Y_{1t}\}$ and $\{Y_{2t}\}$ jointly observed at timepoints t=1, 2, …, n, with the assumption that each series $\{Y_{1t}\}$ and $\{Y_{2t}\}$ follows a copula-based Markov process described on section 3.1. Let's mean vector, correlation matrix of the bivariate series as μ_t

and $\tau(t, t-1)$ which are described as below.

$$\mu_t = E(Y_t) = \begin{bmatrix} E(Y_{1t}) \\ E(Y_{2t}) \end{bmatrix}$$

$$\tau(t, t-1) = COV(Y_t, Y_{t-1}) \begin{bmatrix} COV(Y_{1t}, Y_{1,t-1}) & COV(Y_{1t}, Y_{2,t-1}) \\ COV(Y_{2t}, Y_{1,t-1}) & COV(Y_{2t}, Y_{2,t-1}) \end{bmatrix}$$

Here the diagonal elements of the matrix represent the serial dependence between two series, while the off-diagonal elements describe the cross-correlation between two time series.

The joint distribution of Y_{1t} and Y_{2t} given Y_{1t-1}, Y_{2t-1} for t=1, 2,…,n is given by

$$\mathrm{f}(y_{1t}, y_{2t} | y_{1t-1}, y_{2t-1}) = \int_{V^{-1}(F_{1,t}^-)}^{V^{-1}(F_{1,t}^+)} \int_{V^{-1}(F_{2,t}^-)}^{V^{-1}(F_{2,t}^+)} V_2(z_1, z_2, R) d z_2 \, dz_1$$

where V^{-1} is either the inverse cdf (Cumulative distribution function) of the normal distribution or the t-distribution with $V_2(., R)$ being the bivariate normal or t-distribution, respectively. R is correlation matrix capturing the cross correlation between two time series which is described below.

$$\mathrm{R}=\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The limits of the bivariate integral can be calculated as below.

$F_{i,t}^+ = F(y_{i,t} | y_{i,t-1})$ and $F_{i,t}^- = F(y_{i,t} - 1 | y_{i,t-1})$, for i= 1,2 where,

$$F(y_{i,t} | y_{i,t-1}) = \frac{F_{12}(y_{i,t}, y_{i,t-1}) - F_{12}(y_{i,t}, y_{i,t-1} - 1)}{f_{t-1}(y_{i,t-1}; \boldsymbol{\theta})}$$

and

$$F_{12}(y_{i,t}, y_{i,t-1}) = C(F_t(y_{i,t}), F_{t-1}(y_{i,t-1} - 1); \delta)$$

$C(.; \delta)$ represents the bivariate copula function with dependence parameter δ, describing the serial dependence in a single series, and $\boldsymbol{\theta}$ is a vector of the marginal parameters.

Likelihood and parameter estimation for the bivariate model

Likelihood based inference were conducted with maximizing the log-likelihood function of the bivariate distribution. The corresponding likelihood function for the joint distribution is given by,

$$L(\vartheta, y) = f(Y_{11}, Y_{21}). \prod_{t=2}^{n} f(Y_{1t}, Y_{2t} | Y_{1t-1}, Y_{2t-1}) \quad (4)$$

Where $\vartheta = (\theta', \delta_1, \delta_2, \rho)'$,where θ is the marginal parameter vector and δ_1, δ_2 are parameters associated with the serial dependence in each time series respectively. The cross correlation between the two-time series is captured by ρ.

We can construct the log -likelihood function $l(\vartheta, y)$ as below.

$$l(\vartheta, y) = \log(f(Y_{1t}, Y_{2t})) + \sum_{t=2}^{n} \log f(Y_{1t}, Y_{2t} | Y_{1t-1}, Y_{2t-1}).$$

The likelihood function ($l(\vartheta, y)$) contains either a bivariate normal or t-integral function which unable us to use the standard maximization procedures to get the ML estimates. Due to this reason, we evaluated the bivariate integral function using the standard randomized importance sampling method.

We present simulation results for the proposed bivariate model in Section 3.1 after expanding from univariate to bivariate model. For each univariate time series, we considered a copula-based Markov model, where a copula family was used for the joint distribution of subsequent observations, and then, coupled these two-time series using another copula at each time point.

The parameters of the marginal Poisson distribution are shown in Table 3 and Table 4 for positive and negative cross correlations, respectively. Here $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ denote the means, $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ denote zero inflation parameters, $\boldsymbol{\delta_1}$ and $\boldsymbol{\delta_2}$ denote the serial dependence of marginal distributions. ρ is measure of the cross correlation between the two time series distributions.

The Gaussian copula was used to construct marginal distributions for 300 replicates with sample sizes of 100,300 ,500 and the true parameter values are presented in brackets. The count time series with positive cross correlation is presented in Figure 1, and the joint density is shown in Figure 2. When observing the parameter estimates displayed in Table 3, we can state that the estimated values are more precise and converges to the true parameter values as the

sample size increases.

Bivariate ZI count time series models

Table 2:Parameter estimates for the bivariate ZI Poisson model with positive cross correlation

Sample Size	Parameters	Estimate	SE	MSE	MAE
100	$\lambda_1(3)$	3.4021	0.3887	0.3123	0.4599
	$\omega_1(0.3)$	0.3333	0.0835	0.0081	0.0701
	$\lambda_2(5)$	5.1993	0.3832	0.1860	0.3337
	$\omega_2(0.4)$	0.4026	0.0686	0.0047	0.0537
	$\delta_1(0.6)$	0.5425	0.0837	0.0103	0.0788
	$\delta_2(0.4)$	0.3628	0.0963	0.0106	0.0806
	$\rho(0.5)$	0.4822	0.0911	0.0086	0.0748
300	$\lambda_1(3)$	3.4051	0.1974	0.2030	0.4082
	$\omega_1(0.3)$	0.3380	0.0447	0.0034	0.0471
	$\lambda_2(5)$	5.1816	0.2097	0.0768	0.2226
	$\omega_2(0.4)$	0.4065	0.0386	0.0015	0.0309
	$\delta_1(0.6)$	0.5540	0.0433	0.0040	0.0524
	$\delta_2(0.4)$	0.3669	0.0544	0.0040	0.0492
	$\rho(0.5)$	0.4711	0.0493	0.0033	0.0441
500	$\lambda_1(3)$	3.4105	0.1721	0.1980	0.4108
	$\omega_1(0.3)$	0.3408	0.0365	0.0030	0.0456
	$\lambda_2(5)$	5.1843	0.1622	0.0602	0.2028
	$\omega_2(0.4)$	0.4084	0.0293	0.0009	0.0246
	$\delta_1(0.6)$	0.5558	0.0320	0.0030	0.0465
	$\delta_2(0.4)$	0.3700	0.0430	0.0027	0.0413
	$\rho(0.5)$	0.4720	0.0392	0.0023	0.0379

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

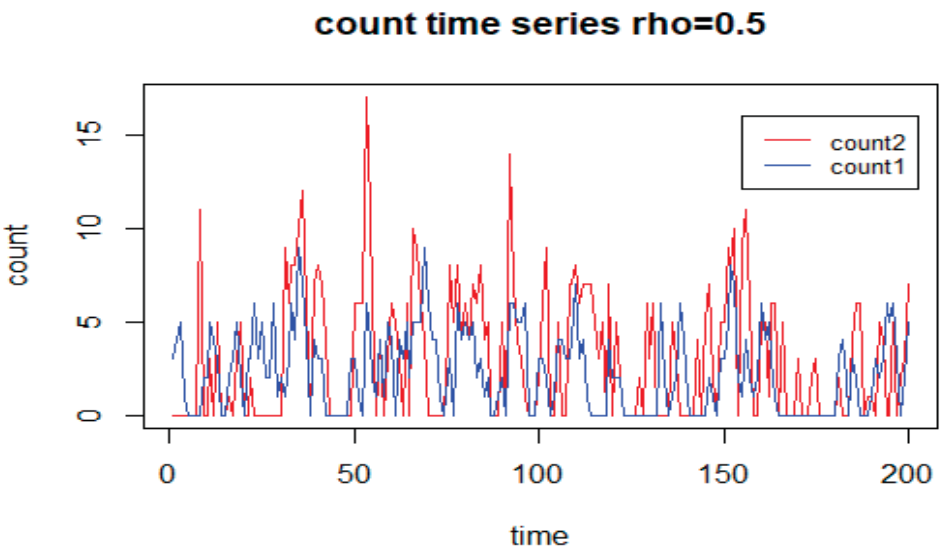

Figure 1. Plot of individual ZI count time series with positive cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

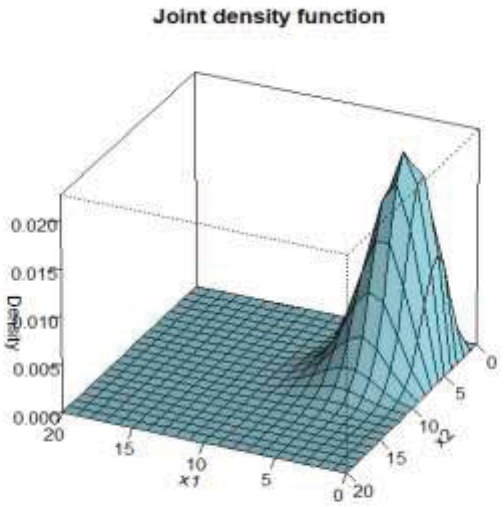

Figure 2. Joint probability density function for the bivariate ZI model with positive cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

There are times when the correlation is negative and table 4 shows the parameter estimates for such scenarios. The Gaussian copula was again used in constructing marginal distributions for 300 replicates with sample sizes of 100, 300, 500 and the true parameter values are presented in brackets. The count time series with negative cross correlation is illustrated in Figure 3, and the joint density is shown in Figure 4. The estimated parameters in Table 4 are more precise and converge to the true parameter values with increasing sample size as observed before.

These results are new because a large body of the literature focuses on positive correlations. Therefore, our proposed algorithm can handle less restrictive cases of ZI count time series data.

Table 3. Parameter estimates for the bivariate ZI Poisson model with negative cross correlation

Sample Size	Parameters	Estimate	St_Dev	MSE	MAE
100	$\lambda_1(3)$	3.417	0.388	0.324	0.474
	$\omega_1(0.3)$	0.341	0.084	0.074	0.074
	$\lambda_2(5)$	5.225	0.382	0.196	0.354
	$\omega_2(0.4)$	0.408	0.070	0.056	0.056
	$\delta_1(0.6)$	0.549	0.085	0.010	0.077
	$\delta_2(0.4)$	0.368	0.103	0.012	0.086
	$\rho(-0.4)$	-0.391	0.104	0.011	0.081
300	$\lambda_1(3)$	3.4072	0.2016	0.2063	0.4110
	$\omega_1(0.3)$	0.3378	0.0455	0.0035	0.0477
	$\lambda_2(5)$	5.2100	0.1965	0.0826	0.2331
	$\omega_2(0.4)$	0.4077	0.0379	0.0015	0.0313
	$\delta_1(0.6)$	0.5529	0.0458	0.0043	0.0534
	$\delta_2(0.4)$	0.3683	0.0537	0.0039	0.0499
	$\rho(-0.4)$	-0.3815	0.0559	0.0035	0.0465
500	$\lambda_1(3)$	3.4181	0.1727	0.2045	0.4182
	$\omega_1(0.3)$	0.3364	0.0348	0.0025	0.0412
	$\lambda_2(5)$	5.1984	0.1575	0.0641	0.2138
	$\omega_2(0.4)$	0.4094	0.0304	0.0010	0.0254
	$\delta_1(0.6)$	0.5524	0.0321	0.0033	0.0493
	$\delta_2(0.4)$	0.3731	0.0417	0.0025	0.0388
	$\rho(-0.4)$	-0.3794	0.0460	0.0025	0.0414

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

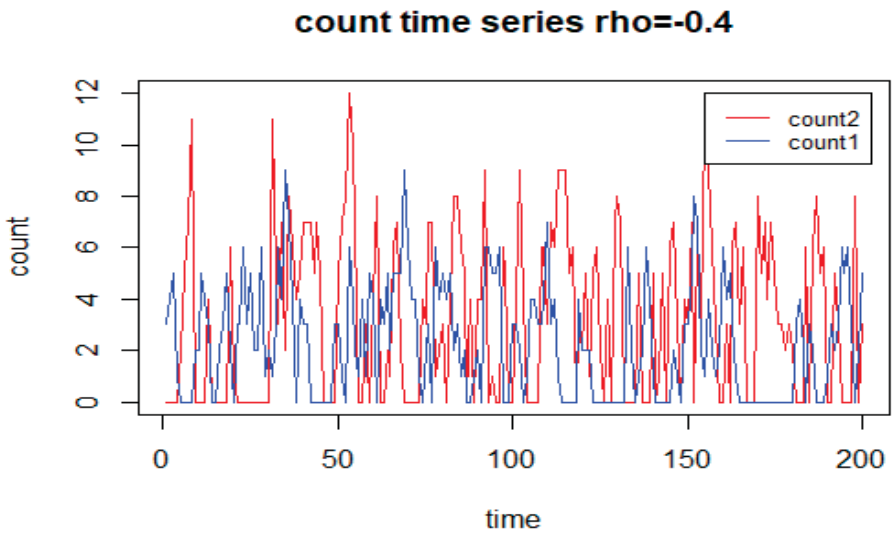

Figure 3. Plot of individual ZI count time series data with negative cross-correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022)

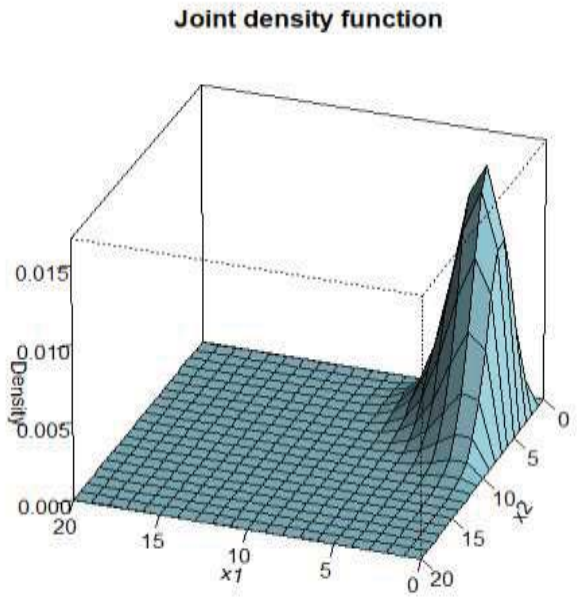

Figure 4. Joint probability density function for the bivariate ZI model with negative cross correlation

Source: Fernando, D., Alqawba, M., Fernando, D., Diawara, N.& Samad, M. (2022).

Applications

The proposed class of method can be applied to model bivariate zero inflated count time series data in the presence of both temporal dependence and cross correlation.

Wang et al. (2013) proposed a bivariate zero inflated poison model to analyze occupational injuries. Alqawba et al. (2021) applied this framework to model monthly counts of forgery and fraud in the 61st police car beat in Pittsburgh, PA. Two count time series were selected to fit the proposed bivariate Poisson class of models under the clear evidence

of the presence of serial dependence and cross correlation.

4. Extensions of the Bivariate Copula for Count Time Series Data

Many copulas have been proposed in the literature for the bivariate and multivariate distributions. The choice of the copula is mainly dictated by the dependence structure.

As shown in Größer and Okhrin (2021), the research on time series dependence and copula direction is productive and has numerous applications. They showed examples of bivariate copulas. Count time series data are observed in several applied disciplines such as environmental science, biostatistics, economics, public health, and finance. Sometimes, a specific count, usually zero, may occur more often than other counts. Moreover, overlooking the frequent occurrence of zeros could result in misleading inferences. A copula-based time series regression model for zero-inflated counts is developed. Applying ordinary Poisson and Negative Binomial distributions to these time series of counts may not be appropriate due to the frequent occurrence of zeros. A new form of ZI is called the Conway-Maxwell Poisson (CMP).

Alqawba et al. (2021) have extended the work done by Masarotto (2012) to include a class of models that accounts for ZI. The marginals are assumed to follow one of the ZIP, ZINB, and ZICMP distributions, and the serial dependence was modeled by a Gaussian copula with a correlation matrix that of a stationary ARMA process. Likelihood inference was carried out using sequential importance sampling. Simulated studies were conducted to evaluate the parameter estimation procedures. Model description and parameter misspecification or unidentifiability are always concerns from the data generation to real data analysis (Faugeras, 2017). Model assessment to check the goodness of fit for the proposed models was done via residual analysis. The proposed models were applied to the occupational health data. According to the residual analysis, the model fits the data adequately, but both ZINB and ZICMP seem to have a slight advantage over ZIP distribution. Future direction is to consider different model construction methods from the marginal regression such as Markov models to handle zero-inflated count time series data. Recently, the use of copula-based time series for ZI counts in the presence of covariates has been proposed in Alqawba et al. (2019) and Alqawba and Diawara (2020). The work considered the cases of ZIP, ZINB, and ZICMP distributed marginals. Likelihood-based inference is considered under a sequential sampling method to estimate both the marginal regression parameters and copula parameters. Improvements in the Bayesian Information Criteria were noted, as discussed in Joe (2014) and Dalla Valle et al. (2018). The applications of these models include occupational injury counts, arson counts, and sandstorm counts.

5. Further Developments and Conclusion

Several high-dimensional copulas are obtained from the bivariate version seen in the previous section. The bivariate time series copula becomes then very important. The vine copula is built from blocks of bivariate version of higher dimension (Acar et al. 2019, Czado). We will only mention the Hierachical Archimedean copula, the Multivariate Archimax copula, the Factor copula, and the Vine copula. Copula functions are particularly interesting in capturing dependence with pairwise Kendall's correlations for invariance to monotonic transformations of marginal distributions. The copula is Archemedian and is applicable for higher than bivariate dimensions of the correlation between marginals (McNeil and Nešlehová, 2009). There is research on the symmetry of copula, and the family of measures under non-degenerate asymptotic distributions (Quessy and Bahraoui, 2018). The disentangling of features with copula transformation is also gaining popularity in so called deep Information bottleneck (DIB) to yield higher convergence rates (Wieczorek et al. 2018, Wieczorek and Roth 2020). As a measure, the copula can be thought as a transformation on a set, which is also a measure preserving transformation. Copulas are also obtained under non-monotonic transformations. Bardossy and Li (2008) proposed a v-transformed copula.

The ideas of Levy processes modelled via copula offer many areas of research (Liu et al., 2021).

The spatio-temporal dependence will become more of a priority as the research evolves. See more in Krupskii and Genton (2017). Bivariate time varying copulas are proposed in Acar et al. (2019). The dynamic vine copula is also adapted to the Bayesian inference (Kreuzer and Czado, 2019).

In this review, we have shown statistical and computational methods for bivariate count time series data analyses using copula distributions. The general framework for discrete count data and the bivariate nature of data are presented. The copula structure is described with details on its analytic perspectives. The identifiability and the choice of copula are very challenging in any discrete data setting and in the case of negative associations between components. As mentioned in Genest et al. (2011), Faugeras (2017) and in Trivedi and Zimmer (2007, 2017), the copula may not generate the perfect data distributions. Such concern is also pointed out in Durante and Sempi (2016). Copula can model bivariate dependence that are invariant under monotonic transformation only (Größer and Okhrin, 2021). When the dependence is weak, the FGM copula offers great alternative, but determining the most appropriate type of FGM copula to fit data is an open problem. Trivedi and Zimmer (2017) proposed several simulations to show these concerns.

Similar to any other functions, the copula functions cannot be deemed as the solution to all data problems. However,

they offer a valuable alternative, especially in the case of discrete data. The research on discrete time series data is more important in this class of functions, especially for bivariate cases as the characterization of bivariate count dependence structure provides tools for may applied problems.

Conflict of Interest

We attest that the manuscript titled “Review of copula for bivariate distributions of zero-inflated count time series data” is original and has not been submitted to or considered for publication elsewhere. The authors declare that they have no competing or conflicts of interest with regard to this publication.

Funding Information

This study received no external funding.

References

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, *44*(2), 182-198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Acar, E. F., Czado, C., & Lysy, M. (2019). Flexible dynamic vine copula models for multivariate time series data. *Econometrics and Statistics*, *12*, 181-197. https://doi.org/10.1016/j.ecosta.2019.03.002

Alqawba, M., & Diawara, N. (2021). Copula-based Markov zero-inflated count time series models with application. *Journal of Applied Statistics*, *48*(5), 786-803. https://doi.org/10.1080/02664763.2020.1748581

Alqawba, M., Fernando, D., & Diawara, N. (2021). A Class of Copula-Based Bivariate Poisson Time Series Models with Applications. *Computation, 9*(10), 108. https://doi.org/10.3390/computation9100108

Armillotta, M., & Fokianos, K. (2021). Poisson network autoregression. *arXiv preprint arXiv:2104.06296*.

Bahraoui, T., Bouezmarni, T., & Quessy, J. F. (2018). Testing the symmetry of a dependence structure with a characteristic function. *Dependence Modeling*, *6*(1), 331-355. https://doi.org/10.1515/demo-2018-0019

Bedford, T., & Cooke, R. M. (2002). Vines--a new graphical model for dependent random variables. *The Annals of Statistics*, *30*(4), 1031-1068. https://doi.org/10.1214/aos/1031689016

Cook, R. D., & Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(2), 210-218. https://doi.org/10.1111/j.2517-6161.1981.tb01173.x

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, *222*. https://doi.org/10.1007/978-3-030-13785-4

Czado, C., Schepsmeier, U., & Min, A. (2012). Maximum likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, *12*(3), 229-255. https://doi.org/10.1177/1471082X1101200302

Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., ... & Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association*, *116*(535), 1533-1547. https://doi.org/10.1080/01621459.2021.1904957

Davis, R. A., Holan, S. H., Lund, R., & Ravishanker, N. (Eds.). (2016). *Handbook of discrete-valued time series*. CRC Press. https://doi.org/10.1201/b19485

Deng, Y., & Chaganty, N. R. (2021). Pair-copula models for analyzing family data. *Journal of Statistical Theory and Practice*, *15*(1), 1-12. https://doi.org/10.1007/s42519-020-00146-z

Durante, F., & Sempi, C. (2016). *Principles of copula theory* (Vol. 474). Boca Raton, FL: CRC press. https://doi.org/10.1201/b18674

Durante, F., Sánchez, J. F., & Sempi, C. (2018). A note on bivariate Archimax copulas. *Dependence Modeling*, *6*(1), 178-182. https://doi.org/10.1515/demo-2018-0011

Fatahi, A. A., Noorossana, R., Dokouhaki, P., & Moghaddam, B. F. (2012). Copula-based bivariate ZIP control chart for monitoring rare events. *Communications in Statistics-Theory and Methods*, *41*(15), 2699-2716. https://doi.org/10.1080/03610926.2011.556296

Fokianos, K. (2021). Multivariate count time series modelling. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.11.006

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, *74*(3), 549-555. https://doi.org/10.1093/biomet/74.3.549

Genest, C., & MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, *40*(4), 280-283. https://doi.org/10.1080/00031305.1986.10475414

Genest, C., Nešlehová, J., & Ziegel, J. (2011). Inference in multivariate Archimedean copula models. *Test*, *20*(2), 223-256. https://doi.org/10.1007/s11749-011-0250-6

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, *10*, 87-102. https://doi.org/10.1016/j.spasta.2014.01.001

Größer, J., & Okhrin, O. (2022). Copulae: An overview and recent developments. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(3), e1557. https://doi.org/10.1002/wics.1557

Gumbel, E. J. (1958). Distributions à plusieurs variables dont les marges sont données. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, *246*(19), 2717-2719.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, *56*(4), 1030-1039. https://doi.org/10.1111/j.0006-341X.2000.01030.x

Han, Z., & De Oliveira, V. (2016). On the correlation structure of Gaussian copula models for geostatistical count data. *Australian & New Zealand Journal of Statistics*, *58*(1), 47-69. https://doi.org/10.1111/anzs.12140

Han, Z., & De Oliveira, V. (2020). Maximum likelihood estimation of Gaussian copula models for geostatistical count data. *Communications in Statistics-Simulation and Computation*, *49*(8), 1957-1981. https://doi.org/10.1080/03610918.2018.1508705

Irannezhad, E., Prato, C. G., Hickman, M., & Mohaymany, A. S. (2017). Copula-based joint discrete–continuous model of road vehicle type and shipment size. *Transportation Research Record*, *2610*(1), 87-96. https://doi.org/10.3141/2610-10

Joe, H., Li, H., & Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, *101*(1), 252-270. https://doi.org/10.1016/j.jmva.2009.08.002

Johnson, M. E., & Tenenbein, A. (1981). A bivariate distribution family with specified marginals. *Journal of the American Statistical Association*, *76*(373), 198-201. https://doi.org/10.1080/01621459.1981.10477628

Karlis, D., & Pedeli, X. (2013). Flexible bivariate INAR (1) processes using copulas. *Communications in Statistics-Theory and Methods*, *42*(4), 723-740. https://doi.org/10.1080/03610926.2012.754466

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14. https://doi.org/10.2307/1269547

Lin, H., & Chaganty, N. R. (2021). Multivariate distributions of correlated binary variables generated by pair-copulas. *Journal of Statistical Distributions and Applications*, *8*(1), 1-14. https://doi.org/10.1186/s40488-021-00118-z

Liu, Y., Djurić, P. M., Kim, Y. S., Rachev, S. T., & Glimm, J. (2021). Systemic risk modeling with lévy copulas. *Journal of Risk and Financial Management*, *14*(6), 251. https://doi.org/10.3390/jrfm14060251

Ma, Z., Hanson, T. E., & Ho, Y. Y. (2020). Flexible bivariate correlated count data regression. *Statistics in Medicine, 39*(25), 3476-3490. https://doi.org/10.1002/sim.8676

Masarotto, G. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics, 6*, 1517-1549. https://doi.org/10.1214/12-EJS721

McNeil, A. J., & Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ1-norm symmetric distributions. *The Annals of Statistics*, *37*(5B), 3059-3097. https://doi.org/10.1214/07-AOS556

Nikoloulopoulos, A. K., & Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics, 37*(9), 1555-1568. https://doi.org/10.1080/02664760903093591

Nikoloulopoulos, A. K., & Moffatt, P. G. (2019). Coupling couples with copulas: analysis of assortative matching on risk attitude. *Economic Inquiry, 57*(1), 654-666. https://doi.org/10.1111/ecin.12726

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, *107*(499), 1063-1072. https://doi.org/10.1080/01621459.2012.682850

Rao, M. B., & Subramanyam, K. (1990). The structure of some classes of bivariate distributions and some applications. *Computational Statistics & Data Analysis*, *10*(2), 175-187. https://doi.org/10.1016/0167-9473(90)90063-N

Ridout, M., Hinde, J., & Demétrio, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, *57*(1), 219-223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press. https://doi.org/10.1201/9780429298547

Safari-Katesari, H., Samadi, S. Y., & Zaroudi, S. (2020). Modelling count data via copulas. *Statistics*, *54*(6), 1329-1355. https://doi.org/10.1080/02331888.2020.1867140

Sellers, K. F., & Raim, A. (2016). A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, *99*, 68-80. https://doi.org/10.1016/j.csda.2016.01.007

Shamma, N., Mohammadpour, M., & Shirozhan, M. (2020). A time series model based on dependent zero inflated counting series. *Computational Statistics*, *35*(4), 1737-1757. https://doi.org/10.1007/s00180-020-00982-4

Shi, P., & Zhang, W. (2015), Private information in healthcare utilization: specification of a copula-based hurdle model. *J. R. Stat. Soc. A, 178,* 337-361. https://doi.org/10.1111/rssa.12065

Sklar, A. (1973). Random variables, joint distribution functions and copulas. *Kybernetika, 9,* 449–460.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, *8*, 229-231.

Trivedi, P. K., & Zimmer, D. M. (2007). *Copula Modeling: An Introduction for Practitioners*. Foundations and Trends in Econometrics. https://doi.org/10.1561/0800000005

Trivedi, P., & Zimmer, D. (2017). A note on identification of bivariate copulas for discrete count data. *Econometrics*, *5*(1), 10. https://doi.org/10.3390/econometrics5010010

Valle, L. D., Leisen, F., & Rossini, L. (2018). Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(3), 523-548. https://doi.org/10.1111/rssc.12237

van den Heuvel, E. R., van Driel, S. A., & Zhan, Z. (2022). A bivariate zero-inflated Poisson control chart: Comments and corrections on earlier results. *Communications in Statistics-Theory and Methods*, *51*(10), 3438-3445. https://doi.org/10.1080/03610926.2020.1736304

Wang, K., Lee, A. H., Yau, K. K., & Carrivick, P. J. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*, *35*(4), 625-629. https://doi.org/10.1016/S0001-4575(02)00036-2

Weiß, C. H., Möller, T., & Kim, H. Y. (2020). Modelling counts with state-dependent zero inflation. *Statistical modelling*.

Wieczorek, A., & Roth, V. (2020). On the difference between the information bottleneck and the deep information bottleneck. *Entropy*, *22*(2), 131. https://doi.org/10.3390/e22020131

Wieczorek, A., Wieser, M., Murezzan, D., & Roth, V. (2018). Learning sparse latent representations with the deep copula information bottleneck. *arXiv preprint arXiv:1804.06216*.

Yang, Q., Xu, M., Lei, X., Zhou, X., & Lu, X. (2014). A Methodological Study on AMH Copula-Based Joint Exceedance Probabilities and Applications for Assessing Tropical Cyclone Impacts and Disaster Risks (Part I). *Tropical Cyclone Research and Review*, *3*(1), 53-62.

Young, D. S., Roemmele, E. S., & Shi, X. (2022). Zero-inflated modeling part II: Zero-inflated models for complex data structures. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(2), e1540.

Yu, R., Yang, R., Zhang, C., Špoljar, M., Kuczyńska-Kippen, N., & Sang, G. (2020). A vine copula-based modeling for identification of multivariate water pollution risk in an interconnected river system network. *Water*, *12*(10), 2741. https://doi.org/10.3390/w12102741

Zhang, Y., & Lam, J. S. L. (2016). A copula approach in the point estimate method for reliability engineering. *Quality and Reliability Engineering International*, *32*(4), 1501-1508. https://doi.org/10.1002/qre.1860

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Negative Binomial and Geometric; Bivariate and Difference Distributions

Yusra A. Tashkandy

Correspondence: Department of Statistics and Operations Research, College of Sciences, King Saud University, Riyadh 145111, Saudi Arabia

Received: September 17, 2022 Accepted: October 18, 2022 Online Published: October 30, 2022

doi:10.5539/ijsp.v11n6p65 URL: https://doi.org/10.5539/ijsp.v11n6p65

Abstract

A similarity and a difference between bivariate negative binomial distribution and bivariate geometric distribution is presented. The distribution of negative binomial difference and geometric difference and the corresponding characteristic function are presented.

Keywords: Negative Binomial, Geometric, Bivariate, Difference

1. Introduction

As a bivariate extension of two exponential distributions, Freund (1961) created his model. A family of bivariate distributions produced by the bivariate Bernoulli distributions were explored by Marshall and Olkin (1985). Bivariate exponential and geometric distributions were explored by Nair and Nair (1988). In Basu and Dhar (1995) presented the BGD (B&D) bivariate geometric model, which is comparable to the Marshall and Olkin (1985) bivariate distribution. A new discrete analog of Freund's model, called BGD (F), was developed by Dhar (1998).

In their 2008 study, Ong et al. studied at the distribution of two discrete random variables from the Panjer family. A skewed distribution known as the generalized discrete Laplace distribution was introduced by Lekshmi and Sebastian (2014). In their 2014 study, Nastic et al. presented the negative binomial difference distribution with an equal chance of success using the INAR model with discrete Laplace marginal distribution. The difference between two independent negative binomial random variables with various parameters was taken into consideration by Song and Smith (2011).

The distribution of $Z = X_1 - X_2$ when X_1 and X_2 are drawn from one of the following bivariate negative binomial distributions or one of the following bivariate geometric distributions is what we are examining in this paper.

2. Bivariate Negative Binomial Distributions

2.1 Double Negative Binomial

The probability density function for the bivariate negative binomial distribution of X_1 and X_2 is given by

$$f(x_1, x_2) = \binom{x_1 + r_1 - 1}{x_1}\binom{x_2 + r_2 - 1}{x_2}(1-p_1)^{r_1}(1-p_2)^{r_2}{p_1}^{x_1}{p_2}^{x_2};\ x_1, x_2 = 0,1,\ldots$$

where $r_1, r_2 > 0, 0 \le p_1, p_2 \le 1$.

where the probability distribution of X_i is given by

$$f(x_i) = \binom{x_i + r_i - 1}{x_i} p_i^{x_i}(1-p_i)^{r_i}; x_i = 0,1,2,\ldots$$

and X_1 ,X_2 be two independent random variables with negative binomial distributions.

The characteristic function provided by $\varphi_{X_1,X_2}(t,s) = \left(\frac{1-p_1}{1-p_1 e^{it}}\right)^{r_1}\left(\frac{1-p_2}{1-p_2 e^{is}}\right)^{r_2}$

2.2 Chou Bivariate Negative Binomial

A bivariate negative binomial distribution is proposed by Chou et al. (2011) as a combination of bivariate Poisson and Gamma distributions. Given by is the joint probability density function.

$$f(x_1, x_2) = \frac{\Gamma(x_1 + x_2 + r)}{x_1!\, x_2!\, \Gamma(r)} \frac{r^r {p_1}^{x_1}{p_2}^{x_2}}{(r + p_1 + p_2)^{r + x_1 + x_2}};\ x_1, x_2 = 0,1,2,\ldots$$

where $r, p_1, p_2 \geq 0$. The marginal mass function is given by

$$f(x_i) = \binom{r + x_i - 1}{x_i}\left(\frac{p_i}{r + p_i}\right)^{x_i}\left(\frac{r}{r + p_i}\right)^{r}$$

with correlation coefficient

$$corr(x_1, x_2) = \sqrt{\frac{p_1 p_2}{p_1 p_2 + r(1 + p_1 + p_2)}}$$

and the characteristic function given by $\varphi_{X_1,X_2}(t,s) = \left(\frac{r}{r+p_1+p_2-p_1e^{it}-p_2e^{is}}\right)^{r}$

2.3 Dependent Bivariate Negative Binomial

Dependent two-variate negative binomial distribution with $\rho = \frac{1-p_1-p_2}{\sqrt{p_1 p_2}}$ correlation. Given is the probability density function.

$$f(x_1, x_2) = \binom{x_1 + x_2 + r - 1}{x_1, x_2}(1 - p_1 - p_2)^r {p_1}^{x_1}{p_2}^{x_2};\ \ x_1, x_2 = 0,1, \ldots$$

where $r > 0, 0 \leq p_1, p_2 \leq 1,\ p_1 + p_2 < 1.$

with the characteristic function presented by$\varphi_{X_1,X_2}(t,s) = \left(\frac{1-p_1-p_2}{1-p_1e^{it}-p_2e^{is}}\right)^{r}$

2.4 Arbous and Sichel Bivariate Negative Binomial

A symmetric bivariate negative binomial distribution with a probability mass function was first introduced by Arbous and Sichel (1954)

$$f_{X_1,X_2}(x_1, x_2) = \frac{(x_1 + x_2 + r - 1)!}{x_1!\, x_2!\, (r-1)!}\left(\frac{r}{r + 2\theta}\right)^{r}\left(\frac{\theta}{r + 2\theta}\right)^{x_1+x_2}; x_1, x_2 = 0,1, \ldots$$

where $r, \theta > 0.$

The characteristic function defined by $\varphi_{X_1,X_2}(t,s) = \left(\frac{r}{r+2\theta-\theta e^{it}-\theta e^{is}}\right)^{r}$

and the marginal probability mass function of X_i

$$f_{X_i}(x_i) = \binom{x_i + r - 1}{x_i}\left(\frac{r}{r + \theta}\right)^{r}\left(\frac{\theta}{r + \theta}\right)^{x_i}; x_i = 0,1, \ldots$$

2.5 Lundberg's Bivariate Negative Binomial

The bivariate negative binomial distributions created by Arbous and Sichel (1954) are a special case of those created by Lundberg (1940), where $\rho = \frac{r}{r+\theta},$ represents for the bivariate negative binomial distributions with the probability mass function

$$f_{X_1,X_2}(x_1, x_2) = \frac{(x_1 + x_2 + r - 1)!}{x_1!\, x_2!\, (r-1)!}\left(\frac{1-\rho}{1+\rho}\right)^{r}\left(\frac{\rho}{1+\rho}\right)^{x_1+x_2}; x_1, x_2 = 0,1, \ldots$$

where $r > 0, 0 < \rho < 1.$

and the characteristic function $\varphi_{X_1,X_2}(t,s) = \left(\frac{1-\rho}{1+\rho-\rho e^{it}-\rho e^{is}}\right)^{r}.$

2.6 Rao Bivariate Negative Binomial

Rao et al. (1973) gave a bivariate negative binomial distribution with probability mass function $f_{X_1,X_2}(x_1,x_2) = \binom{x_1+x_2+r-1}{x_1+x_2}\binom{x_1+x_2}{x_1} w^r \left(\frac{1-w}{2}\right)^{x_1+x_2}$; $x_1, x_2 = 0,1,\ldots$

where $r > 0, 0 < w < 1$.

With the characteristic function given by $\varphi_{X_1,X_2}(t,s) = \left(\frac{2w}{2-(1-w)e^{it}-(1-w)e^{is}}\right)^r$

2.7 Bivariate Negative Binomial by Redaction Method

Suppose that $Y_1 = X_0 + X_1$ and $Y_2 = X_0 + X_2$ have a negative binomial distribution, where $X_i \sim NB(r_i, p)$, and $X_i, i = 0,1,2$ are independent. The joint probability mass function is given by

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{(r_0+r_1+y_1-1)!\,(r_0+r_2+y_2-1)!}{y_1!\,y_2!\,(r_0+r_1-1)!\,(r_0+r_2-1)!}(1-p)^{2r_0+r_1+r_2}p^{y_1+y_2}$$

for $y_1, y_2 = 0,1,\ldots$, where $r_0, r_1, r_2 > 0$ and $0 < p < 1$.

with the characteristic function given by $\varphi_{Y_1,Y_2}(t,s) = \left(\frac{1-p}{1-pe^{it}}\right)^{r_0+r_1}\left(\frac{1-p}{1-pe^{is}}\right)^{r_0+r_2}$

Conclusion 1. When we compare the characteristic function, we find that there are only differences between double negative binomial distributions and dependent bivariate negative binomial distributions. We can find the other bivariate distributions by reparametrized double negative binomial distributions or dependent bivariate negative binomial distributions.

Proof. The characteristic function for each bivariate is given by:

Chou bivariate negative binomial distribution: $\varphi_{X_1,X_2}(t,s) = \left(\frac{r}{r+p_{12}+p_{22}-p_{12}e^{it}-p_{22}e^{is}}\right)^r$

Dependent bivariate negative binomial distribution: $\varphi_{X_1,X_2}(t,s) = \left(\frac{1-p_{13}-p_{23}}{1-p_{13}e^{it}-p_{23}e^{is}}\right)^r$

Arbous and Sichel bivariate negative binomial distribution: $\varphi_{X_1,X_2}(t,s) = \left(\frac{r}{r+2\theta-\theta e^{it}-\theta e^{is}}\right)^r$

Lundberg bivariate negative binomial distribution: $\varphi_{X_1,X_2}(t,s) = \left(\frac{1-\rho}{1+\rho-\rho e^{it}-\rho e^{is}}\right)^r$

Rao bivariate negative binomial distribution: $\varphi_{X_1,X_2}(t,s) = \left(\frac{2w}{2-(1-w)e^{it}-(1-w)e^{is}}\right)^r$

By comparing characteristic functions, we find that:

If $p_{12} = \frac{rp_{13}}{1-p_{13}-p_{23}}$ and $p_{22} = \frac{rp_{23}}{1-p_{13}-p_{23}}$, then Chou $\Leftrightarrow$ dependent

If $p_{12} = p_{22} = \theta$, then Chou $\Leftrightarrow$ Arbous and Sichel

If $p_{12} = p_{22} = \frac{r\rho}{(1-\rho)}$, then Chou $\Leftrightarrow$ Lundberg's

If $p_{12} = p_{22} = \frac{r(1-w)}{2w}$, then Chou $\Leftrightarrow$ *Rao*

Thus, the joint distributions according to dependent, Chou, Arbous and Sichel, Lundberg and Rao bivariate negative binomial distributions are corresponding distributions.

The characteristic function for the independent bivariate negative binomial distribution and the bivariate one using the redaction method given by

$$\varphi_{X_1,X_2}(t,s) = \left(\frac{1-p_1}{1-p_1e^{it}}\right)^{r_{11}}\left(\frac{1-p_2}{1-p_2e^{is}}\right)^{r_{21}}$$

$$\varphi_{X_1,X_2}(t,s) = \left(\frac{1-p}{1-pe^{it}}\right)^{r_{07}+r_{17}} \left(\frac{1-p}{1-pe^{is}}\right)^{r_{07}+r_{27}}$$

we find that, if $p_1 = p_2 = p$, $r_{11} = r_{07} + r_{17}$ and $r_{21} = r_{07} + r_{27}$, then the independent bivariate negative binomial distribution and the bivariate with the redaction method are corresponding distributions.

Then we only need to define two different distributions for the negative binomial difference distribution.

3. Negative Binomial Difference Distributions

3.1 Independent Negative Binomial Difference

If X_1 and X_2 are jointly distributed by double negative binomial distribution, then the random variable $Z = X_1 - X_2$ has the negative binomial difference distribution. The probability distribution is given by Ong, et. Al (2008):

$$f_z(z) = \binom{r_1 + z - 1}{z} (1-p_1)^{r_1}(1-p_2)^{r_2}\, p_1^z \;\; {}_2F_1(r_2, r_1 + z; 1 + z; p_1p_2)\;;\; z = 0,1,2,\ldots$$

and $f(z; r_1, p_1, r_2, p_2) = f(-z; r_2, p_2, r_1, p_1)$

or

$$f_{X_1-X_2}(z) = \begin{cases} \binom{r_1 + z - 1}{z} (1-p_1)^{r_1}(1-p_2)^{r_2}\, p_1^z \;\; {}_2F_1(r_2, r_1 + z; z + 1; p_1p_2) & ;\; z = 0,1,2,\ldots \\ \binom{r_2 - z - 1}{-z} (1-p_1)^{r_1}(1-p_2)^{r_2} p_2^{-z} \;\; {}_2F_1(r_1, r_2 - z; 1 - z; p_1p_2)\;;\; z = -1,-2,\ldots \end{cases}$$

The characteristic function is given by $\varphi_Z(t) = \left(\frac{1-p_1}{1-p_1e^{it}}\right)^{r_1} \left(\frac{1-p_2}{1-p_2e^{-it}}\right)^{r_2}$.

The expected value is $E(Z) = \frac{r_1p_1}{1-p_1} - \frac{r_2p_2}{1-p_2}$, while the variance is $V(Z) = \frac{r_1p_1}{(1-p_1)^2} + \frac{r_2p_2}{(1-p_2)^2}$.

If $r_1 = r_2 = r$

$$f(z) = \binom{r + |z| - 1}{|z|} [(1-p_1)(1-p_2)]^r \;\; {}_2F_1(r, r + |z|; 1 + |z|; p_1p_2) \begin{cases} p_1^z; z = 0,1,2,\ldots \\ p_2^{|z|}; z = -1,-2,\ldots \end{cases}$$

$$r_1, r_2 \geq 0, 0 \leq p_1, p_2 \leq 1$$

$$\varphi_Z(t) = \left(\frac{(1-p_1)(1-p_2)}{(1-p_1e^{it})(1-p_2e^{-it})}\right)^r$$

3.2 Dependent Negative Binomial Difference

Let X_1 and X_2 be jointly distributed dependent bivariate negative binomial distribution, then the probability distribution for the difference $z = x_1 - x_2$ random variable be given by

$$f_Z(z) = \binom{r + |z| - 1}{|z|} (1-p_1-p_2)^r * \; {}_2F_1\left(\frac{r + |z|}{2}, \frac{r + |z| + 1}{2}; 1 + |z|; 4p_1p_2\right) \begin{cases} p_1^z; z = 0,1,2,\ldots \\ p_2^{|z|}; z = -1,-2,\ldots \end{cases}$$

The characteristic function is given by $\varphi_Z(t) = \left(\frac{1-p_1-p_2}{1-p_1e^{it}-p_2e^{-it}}\right)^r$. The expected value is $E(Z) = \frac{r(p_1-p_2)}{1-p_1-p_2}$, and the variance is $V(Z) = \frac{r(p_1+p_2-2p_1p_2)}{(1-p_1-p_2)^2}$.

Conclusion 2. The negative binomial difference between X_1 and X_2 is the same for any bivariate negative binomial distribution.

Proof. The characteristic function from both negative binomial differences is compared, and we discover that, for every $0 \leq p_1, p_2 \leq 1$, $p_1 + p_2 < 1$, or $0 \leq q_i \leq 1$, there are $p_i = \frac{q_i}{1+q_1q_2}$, then, $\varphi_Z(t) = \left(\frac{1-p_1-p_2}{1-p_1e^{it}-p_2e^{-it}}\right)^r \Leftrightarrow \varphi_Z(t) = \left(\frac{(1-q_1)(1-q_2)}{(1-q_1e^{it})(1-q_2e^{-it})}\right)^r$, where $q_i = \frac{1-\sqrt{1-4p_1p_2}}{2p_j}, i \neq j$.

4. Bivariate Geometric Distributions

4.1 Independent Bivariate Geometric

Let X and Y be independent, bivariate geometric distributions, and

$$f(x,y) = (1-p_1)(1-p_2){p_1}^x{p_2}^y; x,y = 0,1,2,\dots$$

be their probability density function.

The $\varphi_{X,Y}(t,s) = \frac{(1-p_1)(1-p_2)}{(1-p_1e^{it})(1-p_2e^{is})}$ provided characteristic function.

4.2 Dependent Bivariate Geometric

Let X and Y be dependent bivariate geometric distributions, where

$$f(x_1,x_2) = \binom{x_1+x_2}{x_1}(1-p_1-p_2){p_1}^x{p_2}^y;\ x,y = 0,1,\dots$$

where $0 \le p_1,p_2 \le 1,\ p_1+p_2 < 1.$

denotes the probability density function and $\varphi_{X,Y}(t,s) = \frac{1-p_1-p_2}{1-p_1e^{it}-p_2e^{is}}$ denotes the characteristic function.

4.3 Omey and Minkova Bivariate Geometric

A bivariate geometric distribution with a probability density function supplied by

$$f(x,y) = \begin{cases} p_1p_2q^{x-1}(1-p_2)^{y-x-1}; y > x \ge 1 \\ 0 \qquad\qquad\quad ; x = y \\ p_1p_2q^{y-1}(1-p_1)^{x-y-1}; x > y \ge 1 \end{cases}$$

where $p_1,p_2,q \ge 0,\ p_1+p_2+q = 1$

and a characteristic function defined by $\varphi_{X,Y}(t,s) = \frac{p_1p_2e^{it+is}}{1-qe^{it+is}}\left(\frac{e^{it}}{1-(1-p_1)e^{it}} + \frac{e^{is}}{1-(1-p_2)e^{is}}\right)$ was proposed by Omey and Minkova (2013).

4.4 Bao Bivariate Geometric

Bao (2011) suggested a bivariate geometric distribution with the characteristic function denoted by $\varphi_{X,Y}(t,s) = \frac{e^{it+is}}{(1-qe^{it+is})}\left(p_{12} + \frac{p_2(p_2+p_{12})e^{it}}{1-(1-p_2-p_{12})e^{it}} + \frac{p_1(p_1+p_{12})e^{is}}{1-(1-p_1-p_{12})e^{is}}\right)$ and a probability density function denoted by

$$f(x,y) = \begin{cases} p_1(p_1+p_{12})q^{x-1}(1-p_1-p_{12})^{y-x-1}; y > x, x,y = 1,2,\dots \\ p_{12}q^{x-1} \qquad ; x = y, x,y = 1,2,\dots \\ p_2(p_2+p_{12})q^{y-1}(1-p_2-p_{12})^{x-y-1}; x > yx, y = 1,2,\dots \end{cases}$$

where $q = 1-p_1-p_2-p_{12}, 0 \le p_1,p_2,p_{12} \le 1.$

4.5 Basu and Dhar Bivariate Geometric

A bivariate geometric model (BGD (B&D)) similar to Marshall and Olkin's (1967) bivariate distribution with the pmf

$$f(x,y) = \begin{cases} q_1(1-p_2p_{12}){p_1}^{x-1}(p_2p_{12})^{y-1}; y > x \\ (p_1p_2p_{12})^{x-1}(1-p_1p_{12}-p_2p_{12}+p_1p_2p_{12}); x = y \\ q_2(1-p_1p_{12}){p_2}^{y-1}(p_1p_{12})^{x-1}; x > y \end{cases}$$

where $1 \le x,y \in Z^+, q_i = 1-p_i; i = 1,2, 0 \le p_1,p_2 \le 1$

and characteristic function given by

$\varphi_{X,Y}(t,s) = \frac{e^{it+is}}{(1-p_1p_2p_{12}e^{it+is})}\left((1-p_1p_{12}-p_2p_{12}+p_1p_2p_{12}) + \frac{p_1p_{12}q_2(1-p_1p_{12})e^{it}}{1-p_1p_{12}e^{it}} + \frac{p_2p_{12}q_1(1-p_2p_{12})e^{is}}{1-p_2p_{12}e^{is}}\right)$ was proposed by Basu and Dhar (1995).

Conclusion 3. There are only two different bivariate geometric distributions, and by reparametrizing these two, we can identify the remaining bivariate distributions.

Proof. According to separately, Omey and Minkova, Bao and Basu, and Dhar, the joint distributions are the equivalent distributions.

By displaying the distinctive properties of each distribution, which include

$$\varphi_{X,Y}(t,s)=\frac{(1-p_1)(1-p_2)}{(1-p_1e^{it})(1-p_2e^{is})}=\frac{(1-p_1)(1-p_2)}{1-p_1p_2e^{it+is}}\left(1+\frac{e^{it}p_1}{1-p_1e^{it}}+\frac{e^{is}p_2}{1-p_2e^{is}}\right)$$

at $X,Y\neq 0$, then

$$\varphi_{X,Y}(t,s)=\frac{(1-p_1)(1-p_2)e^{it+is}}{1-p_1p_2e^{it+is}}\left(1+\frac{e^{it}p_1}{1-p_1e^{it}}+\frac{e^{is}p_2}{1-p_2e^{is}}\right)$$

$$\varphi_{X,Y}(t,s)=\frac{p_{13}p_{23}e^{it+is}}{1-q_3e^{it+is}}\left(\frac{e^{it}}{1-q_{13}e^{it}}+\frac{e^{is}}{1-q_{23}e^{is}}\right)$$

$$=\frac{(1-(q_{13}+q_{23})e^{it+is})p_{13}p_{23}e^{it+is}}{(1-q_3e^{it+is})(1-q_{13}e^{it})(1-q_{23}e^{is})}-\frac{p_{13}p_{23}e^{it+is}(1-e^{it}-e^{is})}{(1-q_3e^{it+is})(1-q_{13}e^{it})(1-q_{23}e^{is})}$$

Bao: $\varphi_{X,Y}(t,s)=\frac{e^{it+is}}{(1-q_4e^{it+is})}\left(p_{124}+\frac{p_{24}(p_{24}+p_{124})e^{it}}{1-(1-p_{24}-p_{124})e^{it}}+\frac{p_{14}(p_{14}+p_{124})e^{is}}{1-(1-p_{14}-p_{124})e^{is}}\right)$

If $p_{124}=(1-p_1)(1-p_2)$, $p_{14}=p_2(1-p_1)$, $p_{24}=p_1(1-p_2)$, and $q_4=p_1p_2$, then $independent$ at $X,Y\neq 0\Leftrightarrow$ Bao.

Basu and Dhar Bivariate Geometric: $\varphi_{X,Y}(t,s)=\frac{e^{it+is}}{(1-p_{15}p_{25}p_{125}e^{it+is})}\left((1-p_{15}p_{125}-p_{25}p_{125}+p_{15}p_{25}p_{125})+\right.$

$\left.\frac{p_{15}p_{125}q_{25}(1-p_{15}p_{125})e^{it}}{1-p_{15}p_{125}e^{it}}+\frac{p_{25}p_{125}q_{15}(1-p_{25}p_{125})e^{is}}{1-p_{25}p_{125}e^{is}}\right)$

If $p_{125}=1$, then $independent$ at $X,Y\neq 0\Leftrightarrow$ Basu and Dhar.

5. Geometric Difference Distributions

5.1 Independent Geometric Difference

The probability distribution for the difference $Z=X-Y$ is a random variable if X and Y are simultaneously distributed according to an independent bivariate geometric distribution.

$$f(z)=\frac{(1-p_1)(1-p_2)}{1-p_1p_2}\begin{cases}{p_2}^{-z};z\leq 0\\ {p_1}^{z};z>0\end{cases}$$

The characteristic function is $\varphi_Z(t)=\left(\frac{(1-p_1)(1-p_2)}{(1-p_1e^{it})(1-p_2e^{-it})}\right)$. The expected value is $E(Z)=\frac{p_1}{1-p_1}-\frac{p_2}{1-p_2}$, while the variance is $V(Z)=\frac{p_1}{(1-p_1)^2}+\frac{p_2}{(1-p_2)^2}$.

which corresponds to the Laplace distribution.

5.2 Dependent Geometric Difference

If X and Y are jointly distributed according to a dependent bivariate geometric distribution.

The probability distribution for the difference $Z=X-Y$ random variable is given by

$$f_{X_1-X_2}(z)=(1-p_1-p_2)*\ {}_2F_1\left(\frac{1+|z|}{2},\frac{2+|z|}{2};1+|z|;4p_1p_2\right)\begin{cases}p_1^z;z=0,1,2,\ldots\\ p_2^{|z|};z=-1,-2,\ldots\end{cases}$$

The characteristic function is given by $\varphi_Z(t)=\left(\frac{1-p_1-p_2}{1-p_1e^{it}-p_2e^{-it}}\right)$. The expected value is $E(Z)=\frac{p_1-p_2}{1-p_1-p_2}$, while the variance is $V(Z)=\frac{p_1+p_2-2p_1p_2}{(1-p_1-p_2)^2}$.

Conclusion 4. The Laplace distribution is the same geometric difference between X_1 and X_2 if they come from any bivariate geometric distribution.

Proof. For any $0 \le p_1, p_2 \le 1,\ p_1 + p_2 < 1$, or $0 \le q_i \le 1$, there are $p_i = \frac{q_i}{1+q_1 q_2}$,

then, $\varphi_Z(t) = \left(\frac{1-p_1-p_2}{1-p_1 e^{it}-p_2 e^{-it}}\right) \Leftrightarrow \varphi_Z(t) = \left(\frac{(1-q_1)(1-q_2)}{(1-q_1 e^{it})(1-q_2 e^{-it})}\right)$,

where $q_i = \frac{1-\sqrt{1-4p_1p_2}}{2p_j}, i \ne j$.

6. New formula of ${}_2F_1$ Hypergeometric Function

For the hypergeometric function ${}_2F_1(.,.;.;.)$ that is readily obtained from the negative binomial distribution, the following theorem provides additional relations.

For any $a > 0,\ b > 0$ and $0 \le p_1, p_2, p \le 1$.

1. $\sum_{n=0}^{\infty}\left(\frac{(a)_{(n)}(b)_{(n)}}{n!^2}(p_1p_2)^n\right)\left[\ {}_2F_1(n+a,1;n+1;p_1) + \ {}_2F_1(n+b,1;n+1;p_2)\right] = \left[\frac{1}{(1-p_1)^a(1-p_2)^b} + \ {}_2F_1(a,b;1;p_1p_2)\right]$

   i. $\sum_{n=0}^{\infty}\left(\frac{(a)_{(n)}}{n!}p^n\right)^2 \ {}_2F_1(n+a,1;n+1;p) = \frac{1}{2}\left[(1-p)^{-2a} + \ {}_2F_1(a,a;1;p^2)\right]$

   ii. $\sum_{n=0}^{\infty}\left(\frac{(a)_{(n)}}{n!}\right)^2 (p_1p_2)^n\left[\ {}_2F_1(n+a,1;n+1;p_1) + \ {}_2F_1(n+a,1;n+1;p_2)\right] = \left[\frac{1}{[(1-p_1)(1-p_2)]^a} + \ {}_2F_1(a,a;1;p_1p_2)\right]$

   iii. $\sum_{n=-\infty}^{\infty}\int_0^{2\pi}\frac{cos(|n|t)}{(1-2pcos(t)+p^2)^a}dt = 2\pi(1-p)^{-2a}$

2. $\sum_{n=0}^{\infty}\frac{(a)_{(n)}(b)_{(n)}}{n!(n-1)!}(p_1p_2)^n\left[p_1(n+a)\ {}_2F_1(n+a,2;n+2;p_1) - p_2(n+b)\ {}_2F_1(n+b,2;n+2;p_2)\right] = \frac{1}{(1-p_1)^a(1-p_2)^b}\left(\frac{ap_1}{1-p_1} - \frac{bp_2}{1-p_2}\right)$

   or

$$\sum_{n=0}^{\infty}\frac{\Gamma(n+a+1)\Gamma(n+b)}{n!\,(n-1)!}(p_1)^{n+1}(p_2)^n\ {}_2F_1(n+a,2;n+2;p_1) - \frac{\Gamma(a+1)\Gamma(b)p_1}{(1-p_1)^{a+1}(1-p_2)^b}$$

$$= \sum_{n=0}^{\infty}\frac{\Gamma(a+n)\Gamma(b+1+n)}{n!\,\Gamma(n)}(p_1)^n(p_2)^{n+1}\ {}_2F_1(n+b,2;n+2;p_2)$$

$$- \frac{\Gamma(a)\Gamma(b+1)p_2}{(1-p_1)^a(1-p_2)^{b+1}}$$

   i. $\sum_{n=0}^{\infty}\frac{(a)_{(n)}(b)_{(n)}}{n!(n-1)!}p^{2n}\left[(a+n)\ {}_2F_1(a+n,2;n+2;p) - (b+n)\ {}_2F_1(b+n,2;n+2;p)\right] = \frac{(a-b)}{(1-p)^{a+b+1}}$

   ii. $\sum_{n=0}^{\infty}\frac{(a+n)(a)_{(n)}{}^2}{n!\Gamma(n)}(p_1p_2)^n\left[p_1\ {}_2F_1(n+a,2;n+2;p_1) - p_2\ {}_2F_1(n+a,2;n+2;p_2)\right] = \frac{a}{[(1-p_1)(1-p_2)]^a}\left(\frac{p_1}{1-p_1} - \frac{p_2}{1-p_2}\right)$

3. $(1-p_1)^a(1-p_2)^b \sum_{n=0}^{\infty} \frac{n^2}{n!}\left[{p_1}^n(a)_{(n)}\ {}_2F_1(b, a+n; n+1; p_1p_2) + {p_2}^n(b)_{(n)}\ {}_2F_1(a, b+n; n+1; p_1p_2)\right] =$

$\frac{ap_1}{(1-p_1)^2}(1+ap_1) + \frac{bp_2}{(1-p_2)^2}(1+bp_2) - \frac{2abp_1p_2}{(1-p_1)(1-p_2)}$

or

$$(1-p_1)^a(1-p_2)^b \sum_{n=0}^{\infty} \frac{n^2(a)_{(n)}}{\mathrm{n!}} {p_1}^n\ {}_2F_1(b, a+n; n+1; p_1p_2) - \frac{ap_1}{(1-p_1)^2}(1+ap_1) + \frac{abp_1p_2}{(1-p_1)(1-p_2)}$$

$$= -(1-p_1)^a(1-p_2)^b \sum_{n=0}^{\infty} \frac{n^2(b)_{(n)}{p_2}^n}{\mathrm{n!}}\ {}_2F_1(a, b+n; n+1; p_1p_2) + \frac{bp_2}{(1-p_2)^2}(1+bp_2)$$

$$- \frac{abp_1p_2}{(1-p_1)(1-p_2)}$$

i. $(1-p)^{a+b} \sum_{n=0}^{\infty} \frac{n^2p^n}{\mathrm{n!}}\left[(a)_{(n)}\ {}_2F_1(b, a+n; 1+n; p^2) + (b)_{(n)}\ {}_2F_1(a, b+n; n+1; p^2)\right] =$

$\frac{p}{(1-p)^2}[a+b+p(a-b)^2]$

ii. $\sum_{n=0}^{\infty} \frac{n^2p^n(a)_{(n)}}{\mathrm{n!}}\ {}_2F_1(n+a, a; n+1; p^2) = \frac{\mathrm{a}p}{(1-p)^{2(1+a)}}$

iii. $\sum_{n=0}^{\infty} \frac{n^2}{2\pi}\int_0^{2\pi} \frac{cos(nt)}{(1-2pcos(t)+p^2)^a}dt = \frac{\mathrm{a}p}{(1-p)^{2(1+a)}}$

iv. $[(1-p_1)(1-p_2)]^a \sum_{n=0}^{\infty} \frac{n^2(a)_{(n)}}{\mathrm{n!}a}\ {}_2F_1(n+a, a; n+1; p_1p_2)[{p_1}^n + {p_2}^n] = \frac{p_1}{(1-p_1)^2}(1+ap_1) +$

$\frac{p_2}{(1-p_2)^2}(1+ap_2) - \frac{2ap_1p_2}{(1-p_1)(1-p_2)}$

4. $\sum_{n=0}^{\infty} \frac{(a)_{(n)}}{\mathrm{n!}}(p_1e^{it})^n\ {}_2F_1(n+a, b; n+1; p_1p_2) + \sum_{n=0}^{\infty} \frac{(b)_{(n)}}{\mathrm{n!}}(p_2e^{-it})^n\ {}_2F_1(a, n+b; n+1; p_1p_2) -$

${}_2F_1(a, b; 1; p_1p_2) = \frac{1}{(1-p_1e^{it})^a(1-p_2e^{-it})^b}$

i. $\sum_{n=0}^{\infty} \frac{(a)_{(n)}}{\mathrm{n!}}\ {}_2F_1(a, a+n; 1+n; p_1p_2)\left[(p_1e^{it})^n + (p_2e^{-it})^n\right] -\ {}_2F_1(a, a; 1; p_1p_2) =$

$\frac{1}{[(1-p_1e^{it})(1-p_2e^{-it})]^a}$

ii. $\sum_{n=-\infty}^{\infty} e^{itn}\int_0^{2\pi} \frac{cos(|n|t)}{(1-2pcos(t)+p^2)^a}dt = \frac{2\pi}{[(1-pe^{it})(1-pe^{-it})]^a}$

iii. $\sum_{n=0}^{\infty} \frac{(a)_{(n)}}{\mathrm{n!}}\ {}_2F_1(a, a+n; n+1; p^2)\left[(pe^{it})^n + (pe^{-it})^n\right] -\ {}_2F_1(a, a; 1; p^2) = \frac{1}{[(1-pe^{it})(1-pe^{-it})]^a}$

iv. $\sum_{n=-\infty}^{\infty} \frac{(a)_{(|n|)}}{|n|!}p^{|n|}e^{itn}\ {}_2F_1(|n|+a, a; |n|+1; p^2) = \frac{1}{[(1-pe^{it})(1-pe^{-it})]^a}$

v. $\sum_{n=0}^{\infty} \frac{(a)_{(n)}}{\mathrm{n!}}(pe^{it})^n\ {}_2F_1(n+a, b; n+1; p^2) + \sum_{n=0}^{\infty} \frac{(b)_{(n)}}{\mathrm{n!}}(pe^{-it})^n\ {}_2F_1(a, n+b; n+1; p^2) -$

${}_2F_1(a, b; 1; p^2) = \frac{1}{(1-pe^{it})^a(1-pe^{-it})^b}$

5. $${}_2F_1(a,b;b-c+1;z)=\frac{\Gamma(b+n)\Gamma(c)}{\Gamma(c+n)\Gamma(b)}\left(1-\sqrt{z}\right)^{2n}\ {}_2F_1(a+n,b+n;b-c+1;z)$$

6. $$\sum_{x=-\infty}^{\infty}\frac{(a)_{(|x|)}(b)_{(|z-x|)}}{|z-x|!|x|!}p^{|x|+|z-x|}\ {}_2F_1(|x|+a,a;|x|+1;p^2)\ {}_2F_1(|z-x|+b,b;|z-x|+1;p^2)=$$

$$\frac{(a+b)_{(|z|)}}{|z|!}p^{|z|}\ {}_2F_1(|z|+a+b,a+b;|z|+1;p^2)$$

References

Albert, W., Marshall, & Ingram, O. (1985). A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association, 80*(390), 332–338. https://doi.org/10.1080/01621459.1985.10478116

Arbous, A. G., & Sichel, H. S. (1954). New Techniques for the Analysis of Absenteeism Data. *Biometrika, 41,* 77–90. https://doi.org/10.1093/biomet/41.1-2.77

Bao, N. H. (2011). On the Stability of the Bivariate Geometric Composed Distribution's Characterization. *Stud. Univ. Babes-Bolyai Math, 1,* 135-140.

Basu. A. P., & Dhar, S. (1995). Bivariate Geometric Distribution. *Journal Applied Statistical Science, 2*(1), 33-44.

Dhar, S. K. (1998). Data analysis with discrete analog of freund's model. *Journal of Applied Statistical Science, 7*, 169-183.

Freund, J. E. (1961) A bivariate extension of the exponential distribution. *Journal of the American Statistical Association, 56,* 971–977. https://doi.org/10.1080/01621459.1961.10482138

Lekshmi, S., & Sebastian, S. (2014) A Skewed Generalized Discrete Laplace Distribution. *International Journal of Mathematics and Statistics Invention, 2,* 5-102.

Li, J., & Dhar, S. (2010). Modeling with Bivariate Geometric Distributions. Preprint submitted to New Jersey Institute of Technology.

Lundberg, O. (1940). *On Random Processes and Their Application to Sickness and Accident Statistics.* Uppsala: Almquist and Wicksell.

Nair, K. R., Muraleedharan, & Nair, U. (1988). On characterizing the bivariate exponential and geometric distributions. *Annals of the Institute of Statistical Mathematics, 40*(2), 267–271. https://doi.org/10.1007/BF00052343

Nastic, A. S., Ristic, M. M., & Djordjevic, M. S. (2014). An INAR model with discrete Laplace marginal distributions. Preprint submitted to Brazilian Journal of Probability and Statistics.

Omey, E., & Minkova, L. D. (2013). *Bivariate Geometric Distributions.* Hub Research Papers, 2013/02, Economics & Business Science.

Ong, S. H., Shimizu, K., & Ng, C. M. (2008). A Class of Discrete Distributions Arising from Difference of Two Random Variables. *Computational Statistics & Data Analysis, 52*, 1490-1499. https://doi.org/10.1016/j.csda.2007.04.009

Rao, B. R., Mazumdar, S., Waller, J. M., & Li, C. C. (1973). Correlation Between the Numbers of Two Types of Children in a Family. *Biometrics, 29,* 271-279. https://doi.org/10.2307/2529391

Song, Q., & Smith, A. D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics., 27*, 870-871. https://doi.org/10.1093/bioinformatics/btr030

Sunil, K. D. (1998). Data analysis with discrete analog of freund's model. *Journal of Applied Statistical Science, 7,* 169–183.

Copyrights

International Journal of Statistics and Probability; Vol. 11, No. 6; November 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Reviewer Acknowledgements

International Journal of Statistics and Probability wishes to acknowledge the following individuals for their assistance with peer review of manuscripts for this issue. Their help and contributions in maintaining the quality of the journal is greatly appreciated.

Many authors, regardless of whether *International Journal of Statistics and Probability* publishes their work, appreciate the helpful feedback provided by the reviewers.

Reviewers for Volume 11, Number 6

Besa Shahini, University of Tirana, Albania

Chin-Shang Li, School of Nursing, USA

Emmanuel Akpan, Federal School of Medical Laboratory Technology, Nigeria

Gerardo Febres, Universidad Simón Bolívar, Venezuela

Krishna K. Saha, Central Connecticut State University, USA

Mohieddine Rahmouni, University of Tunis, Tunisia

Philip Westgate, University of Kentucky, USA

Poulami Maitra, NORC at the University of Chicago, India

Renisson Neponuceno de Araujo Filho, Universidade Federal do Tocantins, Brazil

Soukaina Douissi, National School of Applied Sciences (ENSA) Cadi Ayyad University, Morocco

Wendy Smith

On behalf of,

The Editorial Board of *International Journal of Statistics and Probability*

Canadian Center of Science and Education

ACADEMIC AND OPEN ACCESS

Canadian Center of Science and Education

CCSENET.ORG

➢ CALL FOR MANUSCRIPTS

International Journal of Statistics and Probability is a peer-reviewed journal, published by Canadian Center of Science and Education. The journal publishes research papers in all aspects of statistics and probability. The journal is available in electronic form in conjunction with its print edition. All articles and issues are available for free download online.

We are seeking submissions for forthcoming issues. All manuscripts should be written in English. Manuscripts from 3000–8000 words in length are preferred. All manuscripts should be prepared in LaTeX or MS-Word format, and submitted online, or sent to: ijsp@ccsenet.org

Paper Selection and Publishing Process

a) Submission acknowledgement. If you submit manuscript online, you will receive a submission acknowledgement letter sent by the online system automatically. For email submission, the editor or editorial assistant sends an e-mail of confirmation to the submission's author within one to three working days. If you fail to receive this confirmation, please check your bulk email box or contact the editorial assistant.

b) Basic review. The editor or editorial assistant determines whether the manuscript fits the journal's focus and scope. And then check the similarity rate (CrossCheck, powered by iThenticate). Any manuscripts out of the journal's scope or containing plagiarism, including self-plagiarism are rejected.

c) Peer Review. We use a double-blind system for peer review; both reviewers' and authors' identities remain anonymous. The submitted manuscript will be reviewed by at least two experts: one editorial staff member as well as one to three external reviewers. The review process may take four to ten weeks.

d) Make the decision. The decision to accept or reject an article is based on the suggestions of reviewers. If differences of opinion occur between reviewers, the editor-in-chief will weigh all comments and arrive at a balanced decision based on all comments, or a second round of peer review may be initiated.

e) Notification of the result of review. The result of review will be sent to the corresponding author and forwarded to other authors and reviewers.

f) Pay the article processing charge. If the submission is accepted, the authors revise paper and pay the article processing charge (formatting and hosting).

g) E-journal is available. E-journal in PDF is available on the journal's webpage, free of charge for download. If you need the printed journals by post, please order at http://www.ccsenet.org/journal/index.php/ijsp/store/hardCopies.

h) Publication notice. The authors and readers will be notified and invited to visit our website for the newly published articles.

More Information

E-mail: ijsp@ccsenet.org

Website: http://ijsp.ccsenet.org

Paper Submission Guide: http://ijsp-author.ccsenet.org

Recruitment for Reviewers: http://www.ccsenet.org/journal/index.php/ijsp/editor/recruitment

➢ JOURNAL STORE

To order back issues, please contact the journal editor and ask about the availability of journals. You may pay by credit card, PayPal, and bank transfer. If you have any questions regarding payment, please do not hesitate to contact the journal editor or editorial assistant.

Price: $40.00 USD/copy Shipping fee: $20.00 USD/copy

ABOUT CCSE

The Canadian Center of Science and Education (CCSE) is a private for-profit organization delivering support and services to educators and researchers in Canada and around the world.

The Canadian Center of Science and Education was established in 2006. In partnership with research institutions, community organizations, enterprises, and foundations, CCSE provides a variety of programs to support and promote education and research development, including educational programs for students, financial support for researchers, international education projects, and scientific publications.

CCSE publishes scholarly journals in a wide range of academic fields, including the social sciences, the humanities, the natural sciences, the biological and medical sciences, education, economics, and management. These journals deliver original, peer-reviewed research from international scholars to a worldwide audience. All our journals are available in electronic form in conjunction with their print editions. All journals are available for free download online.

Mission

To work for future generations

Values

Scientific integrity and excellence

Respect and equity in the workplace

CONTACT US

1595 Sixteenth Ave, Suite 301,
Richmond Hill, Ontario, L4B 3N9,
Canada
Tel: 1-416-642-2606
E-mail: info@ccsenet.org
Website: www.ccsenet.org

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

Aerospace Database
BASE (Bielefeld Academic Search Engine)
EZB (Elektronische Zeitschriftenbibliothek)
Google Scholar
JournalTOCs
Library and Archives Canada
LOCKSS
MIAR
PKP Open Archives Harvester
SHERPA/RoMEO
Standard Periodical Directory
Ulrich's

International Journal of Statistics and Probability

Bimonthly

Publisher	Canadian Center of Science and Education
Address	1595 Sixteenth Ave, Suite 301, Richmond Hill, Ontario, L4B 3N9, Canada
Telephone	1-416-642-2606
E-mail	ijsp@ccsenet.org
Website	http://ijsp.ccsenet.org

ISSN 1927-7032