QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

GENERALIZING THE POINT BISERIAL TO MEASURE THE ASSOCIATION BETWEEN A SET OF DICHOTOMOUS VARIABLES AND A CONTINUOUS VARIABLE

BY

MASHAEL MOHAMMED R A ALDOSARI

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Applied Statistics

January 2022

© 2022 Mashael Mohammed Aldosari. All Rights Reserved.

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Mashael Mohammed Aldosari defended on 02/12/2021.

Dr. Abdel-Salam Gomaa Abdel-Salam
Thesis/Dissertation Supervisor

Dr. Ayman Baklizi
Committee Member

Approved:

Ahmed Elzatahry, Dean, College of Arts and Sciences

ABSTRACT

ALDOSARI, MASHAEL, ALDOSARI, Masters : January : 2022, Applied Statistics

Title: Generalizing the Point Biserial to Measure the Association Between a set of dichotomous Variables and a Continuous Variable

Supervisor of Thesis: Abdel-Salam Gomaa Abdel-Salam.

Exploring the statistical association between more than two variables requires utilizing a proper technique/test along with meeting its required assumptions. Measures of correlation are used to explain such associations by intervals ranging [0–1] or [-1–1], where values near one imply a strong positive relationship and vice versa. Numerous measures of association exist for variables with similar characteristics, such as nominal vs. nominal or ordinal vs. ordinal. However, only a handful of measures exploring the relationship between quantitative and qualitative variables are available. To the best of my knowledge, there is no available measure for measuring the association between a set of dichotomous variables and a continuous variable.

Therefore, the present study aims to propose measures of association to evaluate the strength of the relationship between a set of dichotomous variables and a continuous variable, namely, mixed data. The proposed measures generalized the Point Biserial Correlation Coefficient for dichotomous variables with an identical or non-identical probability.

The study utilized the Mean Square Error (MSE) and Bias as criteria for comparing the performance of the aforementioned measures of the association through extensive simulations and real data analyses.

This study contributed by introducing association measures that can be applied

to data from any field that depends in most cases on dichotomous variables and a continuous variable to study their association. However, it is a common phenomenon in the education sector. Therefore, the proposed measures applied to real datasets derived from the Education sector in Qatar. Education is an essential human virtue, a necessity of society, the basis of a good life, and a sign of freedom. Education is important for the integration of separate entities.

Simulation study and real-data applications were carried out to compare the performances of the η_2^* , and the proposed measures based on MSE and bias considering different probabilities, sample sizes, correlation coefficients, and a different number of dichotomous variables. The research demonstrates that the two proposed measures had the best performances when the sample size and the number of dichotomous increased compared to η_2^* .

DEDICATION

I dedicate my work to *my wonderful family, Father, Mother, aunt, Brothers, and sisters*

Particularly to *my small family, Husband and my little boys Mohammed, Mubarak and Ahmed*

Finally to my special and unique friend Sara

ACKNOWLEDGMENTS

I would like to acknowledge everyone who contributed to my academic accomplishments. Special gratitude and thanks to my supervisor Dr. Abdel-Salam Gomaa who provided advice, guidance, and unfailing support throughout the research process. Adept gratitude is also owed to my committee member Dr. Ayman Baklizi for his endless support. In addition, I am highly indebted to my doctors for imparting their knowledge and constant supervision during my master's journey.

TABLE OF CONTENTS

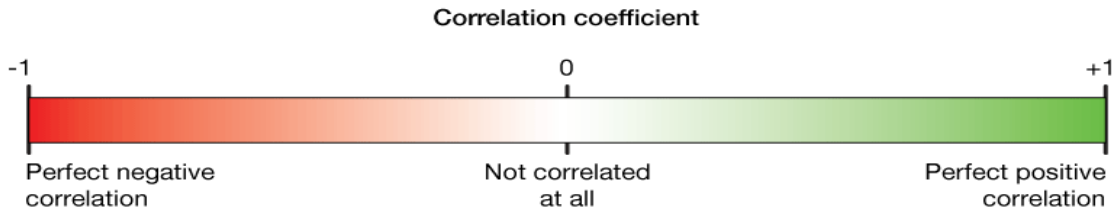
LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: BACKGROUND AND INTRODUCTION

1.1 Introduction

The size of the association between two variables can be summarized using measures of association. Several association measures were designed to have a range of merely 0 to 1, whereas others have a range of -1 to +1. The latter allows you to see if the two variables have a positive or negative relationship with one another. For example, if the correlation coefficient, R, is positive, then an increase in X would increase Y. However, if the measure of association were negative, an increase in X would result in a decrease in Y. Larger measure of association, such as 0.8, would suggest a stronger relationship between the variables, while figures like 0.3 would suggest weaker ones as shown:



The objective of this thesis is to introduce and utilize two generalized forms for the Point Biserial. The first form for measuring the association between a continuous variable and a set of independent dichotomous variables with the same trial probabilities. The second form where the independent binaries with different probabilities of each trial. Furthermore, compared the performance for the proposed measures and η_2^* measure of association for a set of dichotomous and continuous

variables. Lastly, apply the considered measures in the study to real educational data from Qatar.

1.2 Importance of measure of association

- Associations show how variables are related to each other and quantify the relationship between these variables.
- Used to determine the direction and strength of each relationship.
- Another benefit of correlational research is that it opens up a great deal of further research to other scholars when researchers begin investigating a phenomenon or relationship for the first time.

1.3 Definition of terms

This section of the study contains definitions for key terms that used throughout the research:

1.3.1 Direction: the measure's sign indicates whether the relationship is positive or negative. When one variable in a positive relationship is high, so is the other. When one variable is high in a negative relationship, the other is low.

1.3.2 Measures of association: a single number that summarizes the strength of the relationship. This statistic depicts the magnitude and/or direction of a variable-to-variable relationship.

1.3.3 Magnitude: the closer the association is to the absolute value of one, the stronger it is. There is no relationship between the two variables if the measure equals 0.

1.3.4 Generalization: Generalization is the method of developing a general

mathematical formula that can be valid for general and specific cases.

1.4 Level of measurement

Before performing statistical analysis, it is imperative to determine the measurement scale. The measurement scales usually differ with regards to their meaning and numbers. In most instances, the scales are grouped into four measurement levels that fall into two broad types of variables:

1. **Qualitative or Categorical** – under this level, the variables are grouped into two major types. These are nominal and ordinal variables. Researchers need to understand the definition and nature of these variables to know how they are handled during a research process.
    i. **Nominal Variables:** These are variables that differ because they have different names. Therefore, it is not possible to place them in any kind of order when used in a study. Some of the common nominal variables may include ethnicity, race, neighborhood, hair color, and gender.
    ii. **Ordinal Variables:** The second type is the ordinal variables that are unique and can be ranked into different categories when used in a research project or study. The common ordinal variables include class level (junior, senior, freshman, and sophomore) and level of education (college degree, High School Diploma, and Less than High School Diploma).

2. **Quantitative or Continuous/Scale** –In this group of variables, the researcher can use data or information that falls along a given spectrum characterized by

standard intervals. The common ones that can be used in a study are intervals and ratios.

i. **Interval:** Interval refers to variables that do not have absolute zero value. In this case, absolute zero refers to the absence of a value or something. When the interval variable is used, zero is just considered another kind of data point along the scale being used. Therefore, it should not be perceived to be an indication that there is no value. For instance, when talking about the temperature scale, the data of zero degrees is not an indication that there is no heat or some form of temperature. Instead, it is included in the scale just like the other values that show whether it is cold or hot.

ii. **Ratio:** A ratio is considered a research variable containing absolute zero, which is meaningful. It implies that when the ratio variable is given the number 0, it implies that there is nothing that exists. For example, zero oranges mean that there is no orange.

In the actual practices, six variable combinations may be used. These combinations have unique features that the researchers need to understand when using them in a study or project. The common combinations include:

1. Continuous vs. continuous
2. Continuous vs. ordinal
3. Continuous vs. nominal
4. Ordinal vs. ordinal
5. Ordinal vs. nominal
6. Nominal vs. nominal

This study focuses on the combination of multinomial and continuous measures of association.

1.5 Point biserial correlation

product-moment correlation in which one variable is continuous (must be ratio scale or interval scale) and the other one is a discrete random variable (dichotomous), which takes the values 0 and 1, based on a random sample (X_i, Y_i), i = 1, 2…. , n.

Point biserial correlation is defined by

$$\boldsymbol{r_{pb} = \left(\frac{\overline{Y}_1 - \overline{Y}_2}{s_Y}\right)\sqrt{\frac{np_0(1-p_0)}{n-1}}}$$

The formula for the point biserial correlation coefficient is:

- $\boldsymbol{\overline{Y}_1}$ = the mean value (for the entire test) on the continuous variable Y for all data points in of the group that received the positive dichotomous variable (i.e., the "1").
- $\boldsymbol{\overline{Y}_2}$ = the mean value (for the entire test) on the continuous variable Y for all data points in of the group that received the negative dichotomous variable (i.e., the "0").
- $\boldsymbol{s_Y}$ = standard deviation for the entire test.
- $\boldsymbol{p_0}$= Proportion of cases in the "0" group.
- $\boldsymbol{p_1}$ = Proportion of cases in the "1" group.

Where

$$s_Y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{n-1}}$$

$$\bar{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}$$

$$p_1 = \frac{\sum_{k=1}^{n} x_k}{n}$$

$$p_0 = 1 - p_1$$

1.6 Multivariate measure of association

In statistics, different variables can be used to refer to distributions and make sense of statistical data sets. For instance, the researchers can decide to make inferences about single distribution through the use of univariate statistics. This particular model forms the basis of other kinds of statistics. However, it does not show the relationship in the data set that is being examined. For the researcher to explore relationships among variables, it is imperative to use bivariate statistics, which show the association between two variables (Garson, G. D.,2012).

In other cases, the researcher frequently wishes to move beyond this to multivariate statistics, where the relationships among several variables are examined simultaneously. The present study focuses on examining the relationship between X's, which is a multinomial, and Y, which is a continuous variable. The association will be investigated using the multivariate distribution method.

1.7 Theoretical background

Several reports have studied the generalized point biserial correlation coefficient using different ways. Gupta (1960) investigated the point multi-serial correlation coefficient between a continuous variable and a recoded nominal variable based on the corresponding means on the continuous variable. Furthermore, Olsson, Drasgow, and Dorans (1982) introduced the point polyserial correlation coefficient between polychotomous-ordinal categories and continuous variables. Also, the multivariate extension investigated by Olkin and Tate (1961) for a multinomial distribution, where Y conditional distribution for fixed X is multivariate normal. Therefore, the first approach in this study to utilize generalized the point biserial in two forms. First, by measuring the association between a continuous variable and a set of independent dichotomous variables with the same trial probabilities. The second form is a special generalized of the former one, where the independent binaries have different trial probabilities.

To the best knowledge of the researcher and based on the research that has been explored, the contribution of this study is to fill in the gap in the literature that no one has studied the association between a set of dichotomous variables and a continuous variable. Furthermore, the regular Point Biserial correlation coefficient considers each dichotomous outcome separately in a univariate framework. However, this strategy is less efficient than the proposed approach incense of avoiding the increase of the probability of type I error.

Therefore, this study extends the research by Lev (1949), which measures the association between a set of independent dichotomous variables and a continuous

variable by generalizing the Point Biserial correlation coefficient. In the other method, η_2^* will be using that investigated by Taha and Hadi (2016), where will transform nominal variables into a set of dichotomous variables to measure the association between nominal and continuous variables. In addition, it utilizes η_2^* suggested by Taha and Hadi (2016) after transforming a set of dichotomous variables to a nominal variable and then transform that nominal into dependent dichotomous variables.

1.8 Research Questions

Proceeding from the research problem, and under the scope of the literature that was reviewed, the research's questions can be:

(i) How is the correlation between a set of dichotomous variables and a continuous variable can be measured?

(ii) What is the performance of the proposed and other techniques in measuring the association?

(iii) How can researchers compute and interpret those measures of association in a real-life application?

CHAPTER 2: LITERATURE REVIEW

2.1 Level of measurements

It is well-established in the field of statistics that there are four types of measurement levels (nominal, ordinal, ratio, and interval), each requiring a specific statistical analysis. Such variables can be quantitative or qualitative in nature. Quantitative variables (ratios and intervals) involve the use of numbers, while qualitative ones (nominal and ordinal) are solely based on labels.

Nominal qualitative factors code data by assigning limited numbers to categories within ≥ two sets and lacking certain orders or ranking. On the other hand, ordered or implied factors can only be represented as ordinal variables. Further, the nominal type can also be referred to as a dichotomous variable as it contains two dichotomous data. In other words, only two possible outcomes are involved (i.e., gender: male vs. female), with an artificial dichotomy occurring when researchers create a variable via recoding quantitative variables using cutoff values. For example, a researcher can create two age groups in which zero and one are defined as (age >40) and (age <40), respectively (Ulrich & Wirtz, 2004).

As for quantitative variables, interval scales contain data with equal distances between values (i.e., the same distance between 4 and 5 is found between 14 and 15). The most common example of an interval scale is the temperature data using either Celsius or Fahrenheit, where zero ≠ nil. Thus, even though ratio and interval have similar properties, the former has a meaningful or true zero value (Berry, Johnston, & Mielke Jr, 2018; Boslaugh, 2012).

2.2 Measures of association

Researchers should be aware of levels of measurements in order to pose the ability to choose the correct measure of association properly, and therefore, correctly assess the relationship between two or more variables (Islam & Rizwan, 2020; Khamis, 2008).

Commonly, bivariate measures of association with the same characteristic are carried out through Pearson product-moment, Spearman rank-order, Phi, Tetrachoric, and Gamma correlation coefficients (Perinetti, 2019).

- Pearson product-moment is the most common bivariate measure of association introduced by Pearson (1909). It provides the magnitude as well as the direction of the association, in which both scales are intervals or ratios and are normally distributed. The correlation uses these scales to draw a line of best fit and explore the extent of variation found within variable points throughout the line. However, researchers have developed several analyses using the Pearson correlation.

- The Spearman rank-order proposed by Spearman (1906) is used for ordinal scales, in which one or both variables' distribution is unknown since it is a non-parametric measure. Studies such as Hotelling and Pabst (1936); Maurice G Kendall (1948); Maurice George Kendall (1948); Kendall, Kendall, and Smith (1939); Kendall and Smith (1939) are among others emphasizing the use of Spearman rank correlation coefficient in such circumstances.

- Phi correlation coefficient introduced by Yule (1912) is used in cross-tabulated data, where both scales are dichotomous in nature. Interpreting phi correlation coefficients is similar to Pearson product-moment correlation coefficient as the estimated product-moment for two naturally dichotomous variables will return the Phi correlation coefficient. Moreover, the Phi correlation mainly depends on a two-dimensional contingency table containing frequencies by category.

- The tetrachoric correlation coefficient, Pearson (1900), is utilized to estimate the correlation between two normally distributed variables, describing a linear relationship between two continuous variables that have each been measured on an artificially dichotomous scale (Bonett & Price, 2005).

- Gamma correlation coefficient recognized by Goodman and Kruskal (1979) shows the strength of association when two ordinal variables. The Gamma technique measures the association by translating the scale numbers into ordinal "rank," contrary to the Pearson product-moment correlation that assesses the relationship between two continuous variables.

On the contrary, there are measures of association studying the relationship between different types of variables (Barbiero & Hitaj, 2020). For instance,

- As mentioned earlier, the Pearson correlation explores the relationship between continuous variables. Nevertheless, if one of the variables of interest is continuous that is transformed to form dichotomous categories, the Biserial correlation coefficient investigated by Robert Fleming Tate (1950) is more meaningful. However, the original continuous variable must be transformed into another one with dichotomous outcomes.

- Articles by Lev (1949) and Robert F Tate (1954) reviewed a special case of the Pearson's product-moment correlation that fulfills the same assumptions and named Point Biserial correlation coefficient denoted as r_{pb}. A correlation of a relationship strength between a continuous level variable (interval or ratio) and a dichotomous variable is obtained from the Point Biserial Correlation Coefficient.

  Dichotomous variables are nominal scale variables with only two possible values, and researchers often refer to them as dummy or dichotomous variables when performing regression analysis. Dichotomous variables are widely used to denote the presence or membership of one category of specimens that exist, such as male or female. They may also be generated by recoding variables or grouping cases for the analysis–if necessary–where Biserial correlation rather than Point Biserial is used to measure the association (Tare, 1949). However, the Point Biserial correlation coefficient assumes that both dependent and independent variables are random variables and Y

follows normal distribution N(μ_1, σ^2) when X=1 while, N(μ_0, σ^2) when X=0. Where, μ_1 is the mean value for the continuous when X=1; therefore: $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ for all the data points in group 1 with size n_1 and $\mu_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i$. The combined sample size is given by $n = n_0 + n_1$. Yet, authors estimated r_{pb} using the maximum likelihood estimation for bivariate normal distribution by standardized X using Bernoulli distribution, since X has dichotomous outcomes.

- Taha and Hadi (2016) presented two measures for the power of linking between two definite categorical variables, namely, η_1 and η_2. Additionally, more extensions of their measures are provided and broken down so that they can be used to quantify the power of connecting mixed variables, which is where some variables can be qualitative and others being quantitative or categorical. For mixed data, authors transformed nominal variables to be multi- dichotomous variables in which each dichotomous variable depends on the others. The basic idea for those measures is to find a singular value. For two categorical variables, the performance of both measures is better compared with other five different measures of associations. Moreover, they introduced η_1^* and η_2^* for mixed data (categorical vs. continuous) and showed that η_2^* is better than η_1^* based on bias and RMSE.

2.3 Measures of association extended of point biserial correlation coefficient

- A study by Gupta (1960) introduced the point multi-serial correlation

coefficient and examined some of its properties. It is noted that the product-moment correlation coefficient can measure linear relation between qualitative characters when the number of categories is more than two, and the set scores that are assigned to these classes are well-known (Wherry & Taylor, 1946). The moment correlation coefficient is also termed the point multi-serial correlation coefficient (PMS). By definition, Y is a discrete random variable with values y_i (i = 1, 2, … l) and probability Pi and X are continuous random variables to the extent that when Y=y_i, then the conditional variance and mean of X are $\sigma_i^2$and m_i, respectively. Some of the main properties of PMS include $m_1 = m_2 = \cdots = m_l$, which is an adequate condition to take PMS to be equivalent to 0 even though it is not an essential condition. Additionally, when l = 2, PMS becomes invariant for linear transformation of y_1 and y_2. Therefore, 1 and 0 can replace y_2 and When l = 3 and $y_1 + y_2 + y_3$ = 0, where y_1 is greater than y_2, which is also greater than y_3, then p becomes unchanged if the three y values are replaced 1, 0, and -1, respectively.

- A frequently arising model, originally from psychology experiments, contains both continuous and discrete variables. The model is a discrete variable X that takes 0 or 1 and a continuous variable y. X represents the absence or presence of an attribute. The frequency and orientation of the connection between one continuous variable and one discrete variable are calculated via a Point Biserial correlation (LeBlanc & Cox, 2017). Researchers can, for example, use a dot Biserial link to investigate if

wages in the U.S. are associated with gender. “Salary” will be the continuous variable in this example, while “gender” will be the discrete variable as it has two categories (male and female). The multivariate extension investigated by Olkin and Tate (1961) has X as a binomial distribution, and the condition distribution of y for fixed x is normal in this model. Therefore $X = (x_0, x_1, \dots, x_n)$ has a multinomial distribution, and y conditional distribution is $Y = (y_0, y_1, \dots, y_n)$ for fixed X is multivariate normal.

- Olsson et al. (1982) considered Point Polyserial and Polyserial correlations as Point Biserial and Biserial correlations generalizations to derive the association between Polyserial and Point Polyserial correlation. The authors considered the case when one variable has polychotomous-ordinal categories while the other is continuous. Polychotomous-ordinal categories or ordered categorical variables that have more than two categories and some kind of order such as *“1-* if you earn up to 10,000QR”, “2- if you earn 10,000QR- 20,000QR” and “3- if you earn over 20,000QR”.

Worthy mentions that most of the association measures are related to the Pearson product-moment, such as the Point Biserial correlation coefficient, Biserial correlation coefficient, Tetrachoric correlation, Spearman rank-order, etc. In addition, several studies create different ways to transform multi-dichotomous variables to nominal variables and vice versa. Therefore, from the above studies, the suggested approach is a special case of Pearson product-moment by generalizing Point Biserial

correlation while the other approach is derived from η_2^*.

CHAPTER 3: CONTRIBUTIONS IN THIS THESIS

Two proposed measures of association were generalizing point biserial association will discuss and clarify in this section. In addition to the mathematical formulas for η_2^* measure of association and goodness-of-fit criteria.

3.1 Extension of Point Biserial correlation

Lev (1949) proved that the Point Biserial correlation coefficient is a Pearson correlation between a continuous and a dichotomous variable. The proof began with the joint probability distribution function of the two variables. It ended up with a modified version of the typical Pearson formula by replacing the expectation and the variance of one of the two continuous variables with the expectation and the variance of the dichotomous variable. Therefore, it can be concluded that the Point Biserial correlation is a particular case of bivariate Pearson correlation. It proceeds from that the Point Biserial formula and can be extended to involve more than one dichotomous variable to measure the correlation between a set of binaries and a continuous variable. The idea of the extension in this study is developed on the basis that adding up the dichotomous variables together to give a new variable follows Binomial distribution under the assumption of independent and identical dichotomous variables. Considering that the identical assumption is not fulfilled, then the new variable follows the Poisson-Binomial distribution. Nevertheless, that changes nothing in the correlation formula because Binomial distribution may be seen as a special case of the Poisson-Binomial distribution.

3.1.1 Correlation between a set of iid dichotomous and a continuous:

- **Definition:**

A measure of association quantifies a relationship between a continuous variable and a set of dichotomous variables. The measure is a generalization of point Biserial correlation which was introduced by Lev (1949) when k is more than 1, and p is constant. Given Y, being normally distributed a continuous random variable having mean=μ and with variance= σ^2 and given $x_1, x_2, \dots\dots, x_k$ are dichotomous variables, where $k > 1$ represents the number of trials. p is the probability for each of the dichotomous variables, where $0 \leq p \leq 1$. Since x's are Bernoulli trials, so $X \sim bin(k, p)$ is a Binomial distribution. Hence, the summation of x's denoted by X, where X can take values from zero to k.

- **Properties:**

1- The continuous variable should be normally distributed.

2- The dichotomous variables are independent and identical Bernoulli random variables.

3- The data should not contain outlier points.

4- The 1's categories on dichotomous variables must correspond to the higher mean on the continuous variable, and the 0's categories must correspond to the lowest mean on the continuous variable or vice versa.

- **Derivation:**

Let $Y \sim N(\mu, \sigma^2)$, and

Let x_j are identical independent distributions (iid) Bernoulli trial, j=1,…, k,

Note that;

$$n_0 + n_1 = n \qquad \text{for each } x_j, \text{j} = 1,2,3,\dots, \qquad \text{k}$$

Where, n_0 is the number of fails trials for each of the k trials, and n_1 is the number of successes trials for each of the k trials.

Thus,

$$p = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n} \qquad \text{for each } x_j, \text{j} = 1,2,3,\dots, \text{k}$$

Now; Let

$$\text{X} = \sum_{j=1}^{k} x_j = x_1 + x_2 + \cdots \dots + x_k \qquad (1)$$

Therefore,

$$X \sim bin(k, p)$$

The standard form of the correlation coefficient is:

$$cor(X, Y) = \rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \qquad (2)$$

Where,

$$cov(X, Y) = E[XY] - E[X]E[Y] \qquad (3)$$

The mean (μ_X, sometimes denoted as the expected value) as well as the standard deviation (σ_X) of a binomially distributed variable are derived using equations (4) and (5). X is the sum of Bernoulli trials, so simply the mean and variance will be the summation of different probabilities of successes of the Bernoulli distributions:

$$E[X] = \mu_X = kp = \frac{kn_1}{n} \qquad (4)$$

$$\sigma_X = \sqrt{kp(1-p)} = \sqrt{\frac{kn_1}{n}\left(1 - \frac{n_1}{n}\right)} \qquad (5)$$

While the mean (μ_Y, which is also denoted as the **expected value**) and standard deviation (σ_Y) of a normally distributed variable are derived using equations (6) and (7)

$$E[Y] = \mu_Y = \frac{\sum_{i=1}^{n} Y_i}{n} \quad (6)$$

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} \quad (7)$$

Therefore,

$$cov(X,Y) = E[XY] - E[X]E[Y] = \left[\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{kn_1}{n^2}\sum_{i=1}^{n} Y_i\right]$$

Multiply n and divided by n to $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$

Thus,

$$cov(X,Y) = \frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i - kn_1 \sum_{i=1}^{n} Y_i\right] \quad (8)$$

where j =1,.., k

By substituting in equation (2), then

$$cor(X,Y) = \rho = \frac{\frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i - kn_1 \sum_{i=1}^{n} Y_i\right]}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}\sqrt{\frac{kn_1}{n}\left(1 - \frac{n_1}{n}\right)}}$$

Therefore, the proposed measure of association can be obtained using the following formula:

$$\rho_B = \frac{\frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i - kn_1 \sum_{i=1}^{n} Y_i\right]}{\sigma_Y\sqrt{\frac{kn_1}{n}\left(1 - \frac{n_1}{n}\right)}}$$

or

$$\rho_B = \frac{\frac{1}{n}[\sum_{i=1}^{n} X_i Y_i - kp \sum_{i=1}^{n} Y_i]}{\sigma_Y \sqrt{kp(1-p)}} \quad (9)$$

Where,

ρ_B= proposed measure of association for a Binomial experiment,

n = the number of observations,

k = the number of trials,

p = the probability of successes for each of the k trials,

X_i = value of the summation of the independent Bernoulli trials (for ith observation),

Y_i = value of y (for ith observation),

n_0= the number of fails trials for each of the k trials, and

n_1= the number of successful trials for each of the k trials.

Algorithm (a1):

An algorithm may be used to derive the association between a set of dichotomous variables and one continuous variable for the proposed measure ρ_B.

Input:

A mixed dataset **B** consists of a set of dichotomous variables and one continuous variable.

Algorithm:

Step 1. Recode the dichotomous variables to have 1's categories correspond to the highest mean on Y.

Step 2. Compute X, which is the summation of each row in the set of dichotomous variables.

Step 3. Compute p which is the probability of successes

Step 4. Compute μ_X and μ_y using (4) & (6).

Step 5. Compute σ_x and σ_y using (5) & (7).

Step 6. Compute the covariance of X and Y using (8).

Step 7. Compute ρ_B using (9)

Output: the proposed measure of association ρ_B

3.1.2 Correlation between a set of independent but not id dichotomous variables and a continuous:

- **Definition "Poisson-Binomial" :**

The second case, given Y, being normally distributed a continuous random variable having mean=μ and with variance= σ^2 and given $x_1, x_2, \dots\dots, x_k$ are dichotomous variables, where $k > 1$ represents a number of trials. The probability of each dichotomous variable is $p_j = pr(I_j = 1) = 1 - pr(I_j = 0)$, where $0 \leq p_j \leq 1, where\ j = 1, \dots, k$. Since x's are Bernoulli trials, so $X \sim Bernoulli(p_j)$ which is a Poisson-binomial distribution. Hence, the summation of x's denoted by X, where X can take values from zero to k. Hence, this measure could be considered a generalization of the Point Biserial correlation when k is more than one and unequal p (Chen & Liu, 1997; Hong, 2013; Neammanee, 2005; Samuels, 1965).

- **Proprieties:**

The assumptions will be as the assumption in 6.1.1; the only change is that the dichotomous variables are not identically Bernoulli random variables.

- **Derivation:**

Let $Y \sim N(\mu, \sigma^2)$, and x_j independent non-identical distributed Bernoulli trial, j=1,…, k,

Note that;

$$n_{0j} + n_{1j} = n \qquad \text{for each } x_j, \text{j} = 1,2,3, \ldots \ldots, \quad \text{k}$$

$$\frac{n_{1j}}{n_{0j} + n_{1j}} + \frac{n_{0i}}{n_{0i} + n_{1j}} = 1 \text{ , similar } \frac{n_{1j}}{n} + \frac{n_{0j}}{n} = 1 \quad \text{for each } x_j, \text{j} = 1,2,3, \ldots \ldots,$$

k

Thus,

$$p_j = \frac{n_{1j}}{n_{0j} + n_{1j}} = \frac{n_{1j}}{n} \qquad \text{for each } x_j, \text{j} = 1,2,3, \ldots \ldots, \quad \text{k}$$

Now; Let

$$\text{X} = \sum_{j=1}^{k} x_j = x_1 + x_2 + \cdots \ldots + x_k \quad (10)$$

Therefore, $X \sim Bernoulli(p_j)$

The mean (μ_X, sometimes denoted as the expected value) as well as the standard deviation (σ_X) of a Poisson binomial distributed variable are derived using equations (11) and (12). X is the sum of Bernoulli trials, so simply the mean and variance will be the summation of different probabilities of successes of the Bernoulli distributions:

$$E[X] = \mu_X = \sum_{j=1}^{k} p_j = \sum_{j=1}^{k} \frac{n_{1j}}{n_{0j} + n_{1j}} \quad (11)$$

$$\sigma_X = \sqrt{\sum_{j=1}^{k}(1-p_j)p_j} = \sqrt{\sum_{j=1}^{k}\left[\frac{n_{1j}}{n}(1-\frac{n_{1j}}{n})\right]} = \sqrt{\sum_{j=1}^{k}\left[\frac{n_{1j}}{n}(\frac{n_{0j}}{n})\right]}$$

$$= \sqrt{\sum_{j=1}^{k}\frac{n_{0j}n_{1j}}{n^2}} \quad (12)$$

While the mean (μ_Y, which is also denoted as the expected value) and standard deviation (σ_Y) of a normally distributed variable are derived using equations (6) and (7)

Therefore,

$$cov(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{n}\left[\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{k} p_i \sum_{i=1}^{n} Y_i\right]$$

$$= \left[\frac{\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i}{n} - \sum_{j=1}^{k}\frac{n_{1j}}{n}\cdot\frac{\sum_{i=1}^{n} Y_i}{n}\right]$$

$$= \left[\frac{\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i}{n} - \frac{1}{n^2}\sum_{j=1}^{k} n_{1j}\sum_{i=1}^{n} Y_i\right]$$

Multiply n and divided by n to $\frac{\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})Y_i}{n}$

Thus,

$$cov(X,Y) = = \frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k} x_{ij})\, Y_i - \sum_{j=1}^{k} n_{1j}\sum_{i=1}^{n} Y_i\right] \quad (13)$$

By substituting in equation (2)

$$cor(x,y) = \rho = \frac{\frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k}x_{ij})\,Y_i - \sum_{j=1}^{k}n_{1j}\sum_{i=1}^{n}Y_i\right]}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}\sqrt{\sum_{j=1}^{k}\frac{n_{0j}n_{1j}}{n^2}}}$$

$$= \frac{\frac{1}{n^2}\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k}x_{ij})\,Y_i - \sum_{j=1}^{k}n_{1j}\sum_{i=1}^{n}Y_i\right]}{\frac{1}{n}\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}\sqrt{\sum_{j=1}^{k}n_{0j}n_{1j}}}$$

$$= \frac{\frac{1}{n}\left[\sum_{i=1}^{n}Y_i\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k}x_{ij}) - \sum_{j=1}^{k}n_{1j}\right]\right]}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}\sqrt{\sum_{j=1}^{k}n_{0j}n_{1j}}}$$

Therefore, the proposed measure of association can be obtained using the following formula:

$$\rho_{PB} = \frac{\frac{1}{n}\left[\sum_{i=1}^{n}Y_i\left[n\sum_{i=1}^{n}(\sum_{j=1}^{k}x_{ij}) - \sum_{j=1}^{k}n_{1j}\right]\right]}{\sigma_Y\sqrt{\sum_{j=1}^{k}n_{0j}n_{1j}}}$$

or

$$\rho_{PB} = \frac{\frac{1}{n}\left[\sum_{i=1}^{n}X_iY_i - \sum_{j=1}^{k}p_j\sum_{i=1}^{n}Y_i\right]}{\sigma_Y\sqrt{\sum_{j=1}^{k}(1-p_j)p_j}} \quad (14)$$

Where,

ρ_{PB}= proposed measure of association,

n = the number of observations,

k = the number of trials,

p_i = the probability of successes for each of the k trials,

X_i = value of the summation of the independent Bernoulli trials (for ith

observation),

Y_i = value of y (for ith observation),

n_{0j}= The number of fails trials for each of the k trials,

n_{1j}= The number of successful trials for each of the k trials,

Algorithm (a2):

An algorithm may be used to derive the association between a set of dichotomous variables and one continuous variable ρ_{PB}.

Input:

A mixed dataset matrix (**B)** consists of a set of dichotomous variables and one continuous variable.

Algorithm:

Step 1. Recode the dichotomous variables to have 1's categories correspond to the highest mean on Y.

Step 2. Compute X, which is the summation of each row in the set of dichotomous variables.

Step 3. Compute p_i which is the different probabilities of successes

Step 4. From step3 compute the summation of p_i.

Step 5. Compute μ_X and μ_y using (6) & (11).

Step 6. Compute σ_x and σ_y using (7) & (12).

Step 7. Compute the covariance of X and Y using (13).

Step 8. Compute ρ_{PB} using (14)

Output: The proposed measure for the association (ρ_{PB})

3.2 η_2^* correlation coefficient

- **Definition:**

The measure was introduced by Taha and Hadi (2016), where the nominal variable will be transformed into a set of dependent dichotomous variables. For this measure, the set of independent dichotomous variables will be transformed into a nominal variable, and then the nominal one will be transformed into a set of dependent dichotomous variables.

- **Properties:**

One quantitative and one categorical variable.

- **Derivation:**

First, **B** is a set of dichotomous variables, say $x_1, x_2, \dots\dots, x_k$ transforming to the categorical variable with 2^k categories, say **X,** then transformed to dichotomous data $\boldsymbol{B}^*$ consisting of 2^k dichotomous variables, where 2^k is the number of categories in **X**. Then $\boldsymbol{B}^*$ is augmented to the quantitative variable *Y* and obtain the augmented *n* $\times(2^k + 1)$ matrix **W** = (**Y** : $\boldsymbol{B}^*$). Then we compute the $2^k + 1$ singular values of the matrix **W** . These are denoted by $\delta_1 \geq \cdots \geq \delta_{2^k+1}$. Note that the set of dichotomous variables, $\boldsymbol{B}^*$ representing the categorical variable is linearly dependent, then at least one (the smallest) singular value of **W** is zero.

Therefore, the formula is

$$\boldsymbol{\eta_2^*} = \frac{\delta_1 - \delta_{2^k}}{\delta_1 + \delta_{2^k}} \quad (18)$$

Where,

δ_1= the highest singular value,

δ_{2^k}= the second smallest singular value,

k= number of independent dichotomous variables,

Algorithm (c):

An algorithm to derive the association between a set of dichotomous variables and one continuous variable η_2^*.

Input:

A mixed dataset matrix **B** consists of a set of dichotomous variables and one continuous variable.

Algorithm:

Step 1. Obtain the categorical variable by transforming a set of binaries variable to a nominal variable with 2^kgroups.

Step 2. Compute the binary matrix B^* corresponding to the categorical variable X.

Step 3. Compute the ordered singular values of **W,** $\delta = \{\delta_1 \geq \cdots \geq \delta_{2^k+1}\}$.

Step 4. Compute the η_2^* Correlation coefficient using equation (18).

Output: The Correlation coefficient $\boldsymbol{\eta_2^*}$

3.3 Criteria for evaluation

In this study, there are two criteria: bias and MSE will be utilized to assess which measures produce better predictions. To estimate the true correlation, ρ, assume that θ represents one of the three association measures in this study; to generate the synthetic data, we will use that. Assume that θ_i denotes the association measure for the *i*th dataset, $i = 1,\ldots, N$ for each configuration (Ratner, 2009; Taha & Hadi, 2016).

The formula of the two criteria is defined as:

(1) The bias of θ where shows the better performance of the measures of association when the values of bias close to zero is,

$$B(\theta) = \frac{1}{N}\sum_{i=1}^{N}\theta_i - \rho.$$

(2) The MSE of θ for determining the better measures where the measure have the smallest MSE values means that it is the better measure and more precise the predictions is,

$$MSE(\theta) = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2 + \left(\frac{1}{N}\sum_{i=1}^{N}\theta_i - \rho\right)^2$$

Where

$$var(\theta) = \frac{1}{N-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2.$$

Where *N* is the number of simulation runs for each configuration and $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i$.

Therefore, the MSE defined as:

$$MSE(\theta) = var(\theta) + B^2(\theta).$$

CHAPTER 4: EMPIRICAL STUDY

This chapter introduced the simulation study as an empirical study to test measures mentioned in Chapter 3. The performances of the proposed methods are compared using MSE and Bias (given in section 3.3) criteria to decide which of the measures has the best performance. In addition, applied those measures on real data focused on the Education dataset.

4.1 Monte Carlo simulation

In the 1940's the Monte Carlo analysis was developed to obtain a probabilistic approximation to the solution of a mathematical equation or model through a computer-based analysis method. Monte Carlo analysis is a numerical analysis technique that utilizes random sampling to simulate real-world phenomena. Simulation in the context of Monte Carlo analysis is the process of approximating a model's output through repetitive random application of a model's algorithm (Raychaudhuri, 2008) .

Monte Carlo simulation is a type of simulation that computes simulation results using repeated random sampling and statistical analysis. Simulation experiments are conducted to compare η_2^* and proposed measures and to make specific recommendations for practitioners.

In this study, the Monte Carlo simulation technique is considered to assess the performance of two correlation cases: (1) Correlation between a set of independent identical distributed (iid) dichotomous variables (Binomially distributed variables) and a normally distributed variable, (2) Correlation between a set of independent non-identical distributed dichotomous variables (Poisson Binomial distributed variables) and a normal distributed variable. After that, random variables were generated from these distributions based on specific parameters under different scenarios, as will be

presented in the following pages. The simulation process was repeated 10,000 times for each scenario to calculate the bias and mean square error (MSE) of these estimates mentioned in Chapter 3, where all the computations are made using R-Software.

4.1.1 Measure based on a set of iid dichotomous and a continuous

This section evaluated the performance of the generalized point biserial correlation coefficient when we have a set of iid dichotomous. The design parameters that govern the generation of the simulated data are:

- K: the number of dichotomous variables k = 2, 3, 5, 7
- n: the sample size. We chose four settings for n = 30, 100, 250 and 500
- ρ: the correlation coefficient between the two variables. Five settings are chosen for ρ = 0.25, 0.50, 0.70 and 0.95.
- p: the probability for each dichotomous variable. We chose four settings for p = 0.25, 0.50, 0.65 and 0.80.

Regarding the above parameters, the generating process was conducted based on a multiple linear regression that considered one dependent variable and k independent variables, where all variables were generated based on a multivariate normal distribution. Then, the independent variables were recoded to dichotomous variables based on the probabilities that were considered as cut points. After that, the proposed measures and η_2^* measure were applied on the simulated data, followed by the computation of goodness-of-fit criteria, Bias and MSE.

4.1.1.1 Results and comparison

The following tables and charts present the results of the simulation study on how ρ_B measure performs to detect the association between a set of dichotomous

variables and a continuous variable compared to η_2^* measure considering different scenarios of sample sizes and the number of dichotomous variables, as was explained before.

Table 1 contains the results of MSE and Bias that were calculated to assess the performance of ρ_B measure when the number of dichotomous variables is two. In contrast, Table 2 contains the results of the same MSE and Bias to assess the performance of η_2^* measure under the same criteria. Comparing Table 1 to Table 2 shows that the Bias and MSE do not indicate satisfactory performance when the sample size is 30. However, the results suggest that as the sample size increases, the performance of the considered association measures increases by giving smaller MSE and Bias. For instance, Tables 1 and 2 show that when P is 0.25 and the correlation coefficient equals 0.25, the MSE and bias of ρ_B for a sample size of 30 were greater than MSE and bias of η_2^* for a sample size of 30.

Table 1: The Bias and MSE's of ρ_B measure of association for two iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.0308	0.0147	0.0097	0.0104	0.1081	0.0721	0.0581	0.0590
	0.50	0.0297	0.0125	0.0097	0.0116	0.1086	0.0679	0.0590	0.0635
	0.65	0.0291	0.0126	0.0099	0.0113	0.1096	0.0681	0.0587	0.0614
	0.80	0.0369	0.0136	0.0119	0.0106	0.1253	0.0753	0.0648	0.0588
100	0.25	0.0044	0.0026	0.0032	0.0035	0.0415	0.0311	0.0380	0.0404
	0.50	0.0038	0.0027	0.0032	0.0044	0.0365	0.0324	0.0372	0.0458

	0.65	0.0038	0.0027	0.0029	0.0043	0.0394	0.0326	0.0344	0.0444
	0.80	0.0041	0.0027	0.0027	0.0038	0.0406	0.0329	0.0340	0.0421
250	0.25	0.0009	0.0011	0.0015	0.0023	0.0193	0.0247	0.0306	0.0388
	0.50	0.0009	0.0011	0.0018	0.0026	0.0190	0.0253	0.0333	0.0416
	0.65	0.0010	0.0012	0.0016	0.0024	0.0206	0.0250	0.0318	0.0402
	0.80	0.0012	0.0010	0.0015	0.0020	0.0217	0.0230	0.0296	0.0361
500	0.25	0.0004	0.0007	0.0010	0.0018	0.0149	0.0217	0.0278	0.0376
	0.50	0.0005	0.0008	0.0012	0.0018	0.0160	0.0232	0.0303	0.0379
	0.65	0.0004	0.0008	0.0011	0.0019	0.0148	0.0231	0.0294	0.0392
	0.80	0.0005	0.0007	0.0009	0.0016	0.0150	0.0206	0.0258	0.0351

Table 2: The Bias and MSE's of η_2^* measure of association for two iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.0246	0.0109	0.0054	0.0011	-0.1388	-0.0904	-0.0632	-0.0312
	0.50	0.0123	0.0051	0.0023	0.0004	-0.0887	-0.0541	-0.0357	-0.0166
	0.65	0.0173	0.0069	0.0025	0.0006	-0.1098	-0.0690	-0.0423	-0.0213
	0.80	0.0279	0.0148	0.0070	0.0018	-0.1505	-0.1076	-0.0749	-0.0403
100	0.25	0.0220	0.0078	0.0030	0.0008	-0.1423	-0.0844	-0.0535	-0.0289
	0.50	0.0051	0.0012	0.0004	0.0000	-0.0640	-0.0316	-0.0189	-0.0086
	0.65	0.0117	0.0032	0.0012	0.0002	-0.1015	-0.0546	-0.0343	-0.0163
	0.80	0.0295	0.0109	0.0047	0.0014	-0.1656	-0.1008	-0.0672	-0.0367

250	0.25	0.0202	0.0064	0.0026	0.0007	-0.1390	-0.0789	-0.0509	-0.0269
	0.50	0.0031	0.0006	0.0002	0.0000	-0.0529	-0.0251	-0.0147	-0.0068
	0.65	0.0099	0.0026	0.0009	0.0002	-0.0964	-0.0501	-0.0303	-0.0146
	0.80	0.0265	0.0095	0.0042	0.0012	-0.1600	-0.0964	-0.0644	-0.0357
500	0.25	0.0187	0.0062	0.0025	0.0007	-0.1354	-0.0784	-0.0507	-0.0268
	0.50	0.0025	0.0005	0.0001	0.0000	-0.0492	-0.0230	-0.0137	-0.0062
	0.65	0.0093	0.0023	0.0008	0.0002	-0.0951	-0.0484	-0.0295	-0.0143
	0.80	0.0251	0.0090	0.0039	0.0012	-0.1570	-0.0945	-0.0628	-0.0354

For more insight in the performance of ρ_B and η_2^* measure, Figure 1 and Figure 2 were developed to reflect a clear vision about the behaviors of the considered measures regarding the sample size, in addition to a visual comparison between the two measures, where Figure 1 depicts the behavior of MSE of the two measures versus sample size, while Figure 2 depicts the behavior of the absolute Bias of the two measures versus sample size.

As it is well known about MSE and absolute bias as the goodness-of-fit criteria, the smallest MSE, and absolute bias, the better the measure's performance. Clearly, it can be noticed that the performance of η_2^* is better than ρ_B at sample size 30 where η_2^* produced less values of MSE and absolute bias than what ρ_B produced. However, when the sample size is greater than 30, the performance of ρ_B substantially is better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_B are significantly less than the corresponding values of η_2^*. Moreover, the absolute Bias and MSE converge reasonably well to zero when the sample size increases.

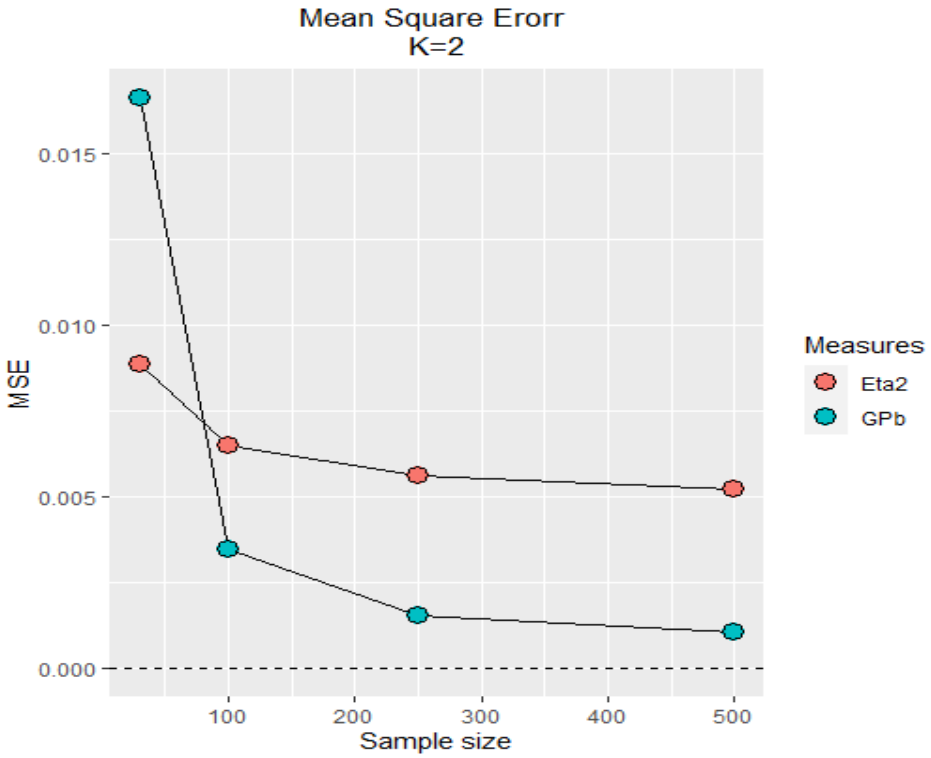


Figure 1. MSE of ρ_B versus MSE of η_2^* for two iid dichotomous variables

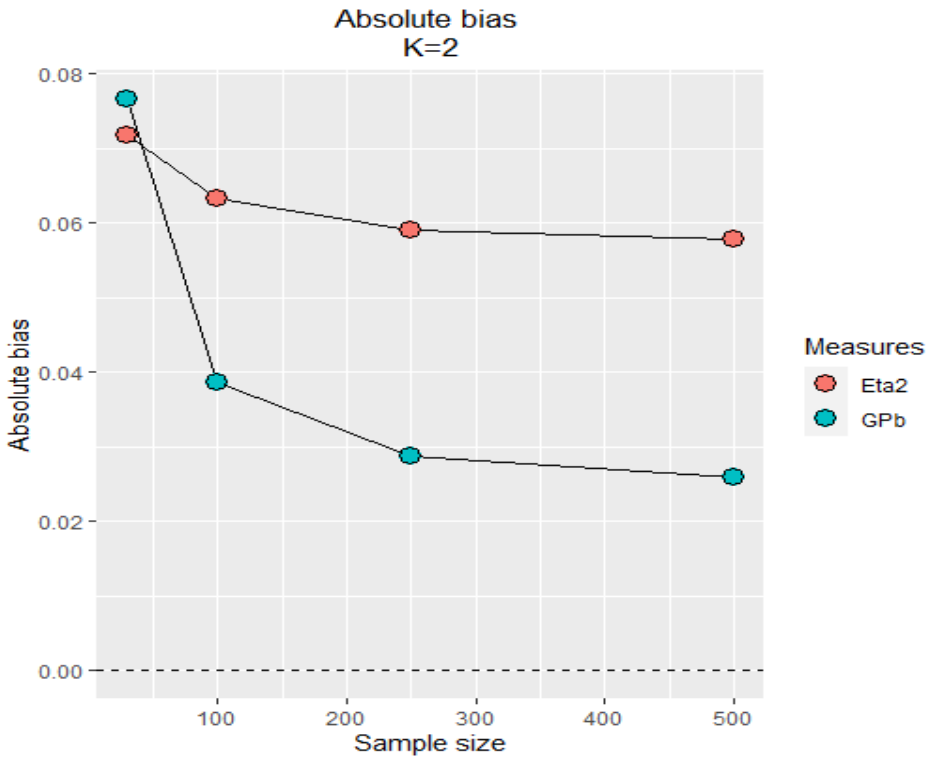


Figure 2. Absoulte bias of ρ_B versus Absoulte bias of η_2^* for two iid dichotomous variables

Also, Table 3 contains the results of MSE and Bias that were calculated to assess the performance of ρ_B measure when the number of dichotomous variables increased to three. On the other hand, Table 4 contains the results of the same MSE and Bias to assess the performance of η_2^* measure under the same criteria. Comparing Table 3 to Table 4 shows almost similar results when the number of dichotomous variables was two that the Bias and MSE do not indicate satisfactory performance when the sample size is less than 100. For instance, Tables 3 and 4 show that when P is 0.25 and the correlation coefficient equals 0.25, the MSE and bias of ρ_B for a sample size of 30 were greater than MSE and bias of η_2^* for a sample size of 30.

Table 3: The Bias and MSE's of ρ_B measure of association for three iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.0488	0.0240	0.0203	0.0172	0.1644	0.1083	0.0993	0.0839
	0.50	0.0515	0.0237	0.0184	0.0177	0.1686	0.1100	0.0957	0.0876
	0.65	0.0479	0.0222	0.0187	0.0177	0.1664	0.1073	0.0933	0.0874
	0.80	0.0552	0.0251	0.0180	0.0173	0.1786	0.1183	0.0921	0.0914
100	0.25	0.0078	0.0055	0.0053	0.0066	0.0639	0.0536	0.0553	0.0607
	0.50	0.0081	0.0048	0.0058	0.0069	0.0650	0.0511	0.0583	0.0635
	0.65	0.0080	0.0053	0.0054	0.0067	0.0669	0.0539	0.0546	0.0626
	0.80	0.0080	0.0050	0.0051	0.0061	0.0663	0.0505	0.0538	0.0592
250	0.25	0.0022	0.0021	0.0029	0.0042	0.0356	0.0369	0.0452	0.0560
	0.50	0.0022	0.0024	0.0031	0.0049	0.0344	0.0403	0.0473	0.0613
	0.65	0.0021	0.0024	0.0030	0.0048	0.0344	0.0399	0.0456	0.0604
	0.80	0.0022	0.0020	0.0027	0.0038	0.0354	0.0356	0.0431	0.0535
500	0.25	0.0011	0.0013	0.0020	0.0034	0.0256	0.0310	0.0408	0.0539
	0.50	0.0010	0.0017	0.0025	0.0041	0.0246	0.0359	0.0451	0.0588
	0.65	0.0010	0.0014	0.0022	0.0038	0.0252	0.0327	0.0430	0.0570
	0.80	0.0009	0.0013	0.0019	0.0031	0.0236	0.0315	0.0393	0.0505

Table 4: The Bias and MSE's of η_2^* measure of association for three iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.0300	0.0190	0.0090	0.0024	-0.1494	-0.1162	-0.0794	-0.0420
	0.50	0.0344	0.0205	0.0097	0.0026	-0.1589	-0.1189	-0.0811	-0.0414
	0.65	0.0338	0.0191	0.0095	0.0024	-0.1576	-0.1171	-0.0806	-0.0399
	0.80	0.0286	0.0173	0.0093	0.0028	-0.1487	-0.1115	-0.0836	-0.0463
100	0.25	0.0183	0.0070	0.0028	0.0007	-0.1263	-0.0784	-0.0500	-0.0263
	0.50	0.0084	0.0029	0.0010	0.0002	-0.0788	-0.0462	-0.0275	-0.0128
	0.65	0.0105	0.0037	0.0015	0.0003	-0.0924	-0.0549	-0.0350	-0.0163
	0.80	0.0246	0.0102	0.0044	0.0012	-0.1491	-0.0957	-0.0634	-0.0342
250	0.25	0.0127	0.0038	0.0015	0.0004	-0.1089	-0.0602	-0.0386	-0.0197
	0.50	0.0023	0.0005	0.0002	0.0000	-0.0424	-0.0208	-0.0127	-0.0058
	0.65	0.0046	0.0012	0.0004	0.0001	-0.0640	-0.0333	-0.0199	-0.0094
	0.80	0.0204	0.0069	0.0029	0.0008	-0.1392	-0.0815	-0.0532	-0.0289
500	0.25	0.0108	0.0030	0.0011	0.0003	-0.1023	-0.0546	-0.0342	-0.0177
	0.50	0.0009	0.0002	0.0000	0.0000	-0.0280	-0.0127	-0.0078	-0.0035
	0.65	0.0030	0.0007	0.0002	0.0000	-0.0534	-0.0262	-0.0155	-0.0073
	0.80	0.0173	0.0057	0.0024	0.0007	-0.1300	-0.0752	-0.0490	-0.0270

For more clarify in the performance of ρ_B and η_2^* measure, Figures 3 and 4 were created to reflect a clear vision of the behaviors of the considered measures in terms of sample size and a visual comparison between the two measures. Where Figure 3 depicts the behavior of MSE of the two measures versus sample size, while Figure 4 depicts the behavior of the absolute Bias of the two measures versus sample size.

As it is well known about MSE and absolute bias as the goodness-of-fit criteria, the smallest MSE, and absolute bias, the better the performance of the measure. Clearly, it can be noticed that the performance of η_2^* is better than ρ_B when the sample size was less than 100 where η_2^* produced smaller values of MSE and absolute bias than what ρ_B produced. However, when the sample size is larger than 30, the performance of ρ_B slightly better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_B were significantly less than the corresponding values of η_2^*. Moreover, Figure 3 shows that performance of the ρ_B in terms of MSE versus the sample size n, almost give same performance compared to η_2^* at sample size 100, while Figure 4 depicts the behavior of the absolute Bias of the two measures versus absolute bias clarify that the ρ_B is outperform compared to η_2^*.

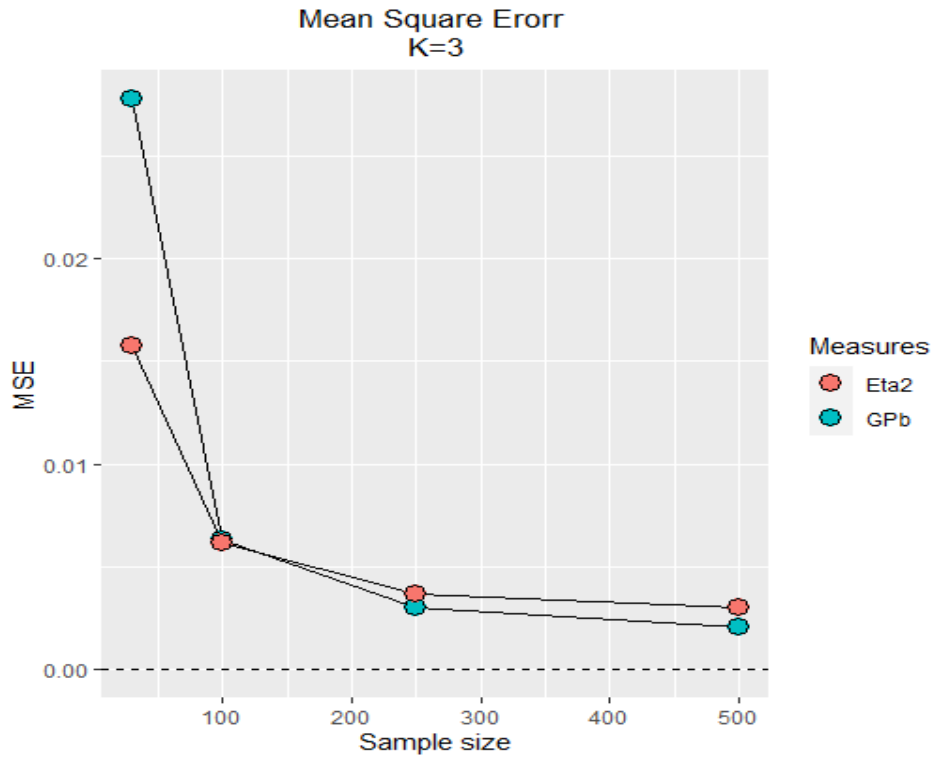


Figure 3. MSE of ρ_B versus MSE of η_2^* for three iid dichotomous variables

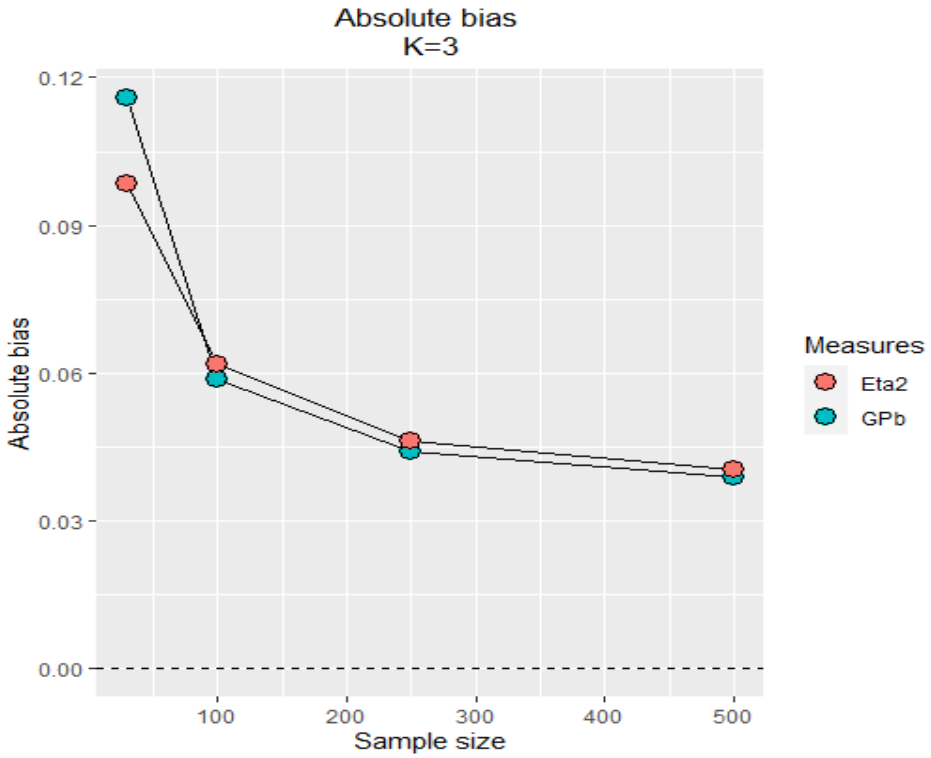


Figure 4. Absolute bias of ρ_B versus Absolute bias of η_2^* for three iid dichotomous variables

Table 5 presents the results of ρ_B and η_2^* to show the performance when the number of dichotomous variables increased to five variables.

In addition, Table 6 shows the results of the same MSE and Bias to evaluate the performance of η_2^* measure under the same goodness-of-fit criteria. However, when Table 5 and Table 6 are compared, Bias and MSE indicate the ρ_B are performing satisfactorily for all considered sample sizes, even for the small sample size.

Table 5: The Bias and MSE's of ρ_B measure of association for five iid dichotomous variables

	MSE	Bias

n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.1085	0.0519	0.0363	0.0282	0.2809	0.1853	0.1530	0.1255
	0.50	0.0933	0.0487	0.0359	0.0268	0.2577	0.1799	0.1504	0.1177
	0.65	0.0949	0.0496	0.0369	0.0280	0.2584	0.1838	0.1506	0.1229
	0.80	0.1089	0.0548	0.0406	0.0287	0.2790	0.1913	0.1611	0.1243
100	0.25	0.0172	0.0099	0.0095	0.0108	0.1079	0.0818	0.0797	0.0864
	0.50	0.0165	0.0095	0.0093	0.0108	0.1068	0.0809	0.0784	0.0855
	0.65	0.0169	0.0096	0.0098	0.0105	0.1073	0.0803	0.0821	0.0837
	0.80	0.0168	0.0092	0.0092	0.0100	0.1089	0.0796	0.0787	0.0824
250	0.25	0.0046	0.0039	0.0047	0.0066	0.0571	0.0547	0.0605	0.0729
	0.50	0.0043	0.0043	0.0050	0.0076	0.0543	0.0574	0.0626	0.0786
	0.65	0.0047	0.0040	0.0052	0.0074	0.0574	0.0549	0.0638	0.0778
	0.80	0.0048	0.0040	0.0046	0.0063	0.0573	0.0547	0.0598	0.0708
500	0.25	0.0018	0.0025	0.0037	0.0053	0.0364	0.0447	0.0557	0.0684
	0.50	0.0020	0.0027	0.0040	0.0064	0.0380	0.0472	0.0592	0.0752
	0.65	0.0017	0.0025	0.0038	0.0060	0.0359	0.0459	0.0575	0.0732
	0.80	0.0020	0.0024	0.0033	0.0053	0.0387	0.0448	0.0531	0.0685

Table 6: The Bias and MSE's of η_2^* measure of association for five iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95

30	0.25	0.0910	0.0630	0.0396	0.0121	-0.2796	-0.2304	-0.1779	-0.0992
	0.50	0.1711	0.1159	0.0686	0.0201	-0.3945	-0.3220	-0.2453	-0.1316
	0.65	0.1440	0.0945	0.0545	0.0189	-0.3597	-0.2887	-0.2149	-0.1248
	0.80	0.0582	0.0393	0.0256	0.0093	-0.2178	-0.1756	-0.1390	-0.0839
100	0.25	0.0557	0.0259	0.0134	0.0036	-0.2262	-0.1537	-0.1100	-0.0569
	0.50	0.0893	0.0432	0.0199	0.0049	-0.2898	-0.2002	-0.1355	-0.0675
	0.65	0.0733	0.0360	0.0168	0.0044	-0.2608	-0.1827	-0.1239	-0.0636
	0.80	0.0416	0.0226	0.0116	0.0035	-0.1939	-0.1433	-0.1017	-0.0561
250	0.25	0.0252	0.0097	0.0042	0.0010	-0.1536	-0.0950	-0.0626	-0.0315
	0.50	0.0274	0.0097	0.0040	0.0009	-0.1600	-0.0955	-0.0609	-0.0290
	0.65	0.0277	0.0099	0.0041	0.0009	-0.1610	-0.0964	-0.0619	-0.0299
	0.80	0.0252	0.0105	0.0046	0.0013	-0.1540	-0.0991	-0.0661	-0.0352
500	0.25	0.0121	0.0039	0.0015	0.0004	-0.1070	-0.0611	-0.0385	-0.0194
	0.50	0.0096	0.0027	0.0010	0.0002	-0.0945	-0.0509	-0.0311	-0.0146
	0.65	0.0104	0.0031	0.0011	0.0002	-0.0989	-0.0542	-0.0329	-0.0155
	0.80	0.0159	0.0056	0.0023	0.0006	-0.1240	-0.0734	-0.0473	-0.0247

Figure 5 indicates that the MSE of ρ_B for all cases is outperforms compared to η_2^*. Furthermore, Figure 6 depicts the behavior of the absolute Bias of the two measures versus sample size. Clearly, it can be noticed that even absolute Bias of ρ_B is outperforms than the corresponding values of η_2^*. However, it can be seen that at sample size 500, results for the goodness-of-fits criteria shows that ρ_B and η_2^* has almost similar results, but on average, still ρ_B has better performance than η_2^*.

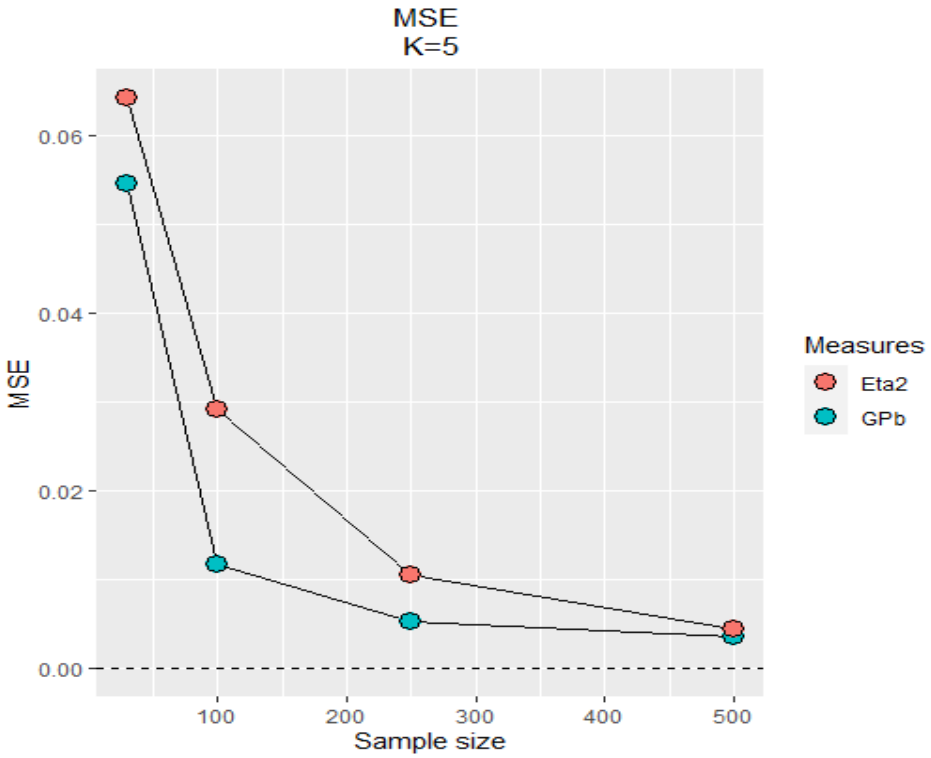


Figure 5. MSE of ρ_B versus MSE of η_2^* for five iid dichotomous variables

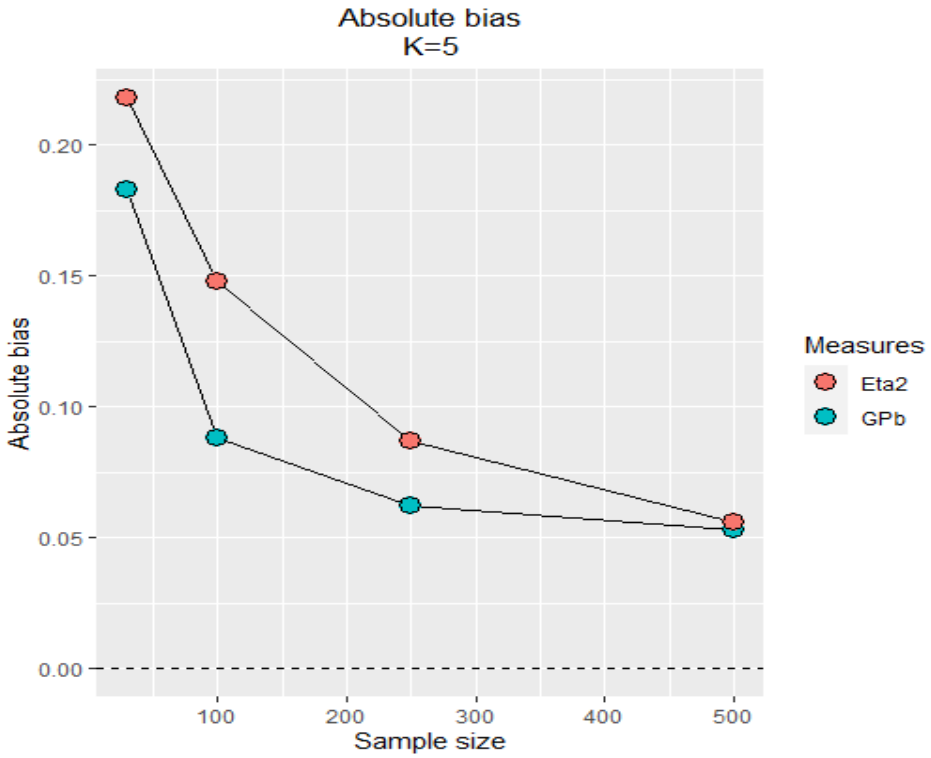


Figure 6. Absolute bias of ρ_B versus Absolute bias of η_2^* for five iid dichotomous variables

Table 7 presents the best performance of the ρ_B measure for all considered sample sizes when the number of dichotomous variables is seven compared to η_2^* performance in Table 8. So it is clear that when the number of dichotomous variables increases, the performance of the ρ_B measure improves in contrast to η_2^*. However, the results provide that as the sample size and dichotomous variables increases, the performance of the association measure ρ_B increases by giving smaller MSE and Bias. Further, the below figures will clarify that the performance of measures.

Table 7: The Bias and MSE's of ρ_B measure of association for seven iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.1560	0.0811	0.0552	0.0407	0.3482	0.2485	0.1979	0.1590
	0.50	0.1453	0.0754	0.0530	0.0377	0.3367	0.2343	0.1909	0.1514
	0.65	0.1458	0.0780	0.0525	0.0392	0.3382	0.2392	0.1915	0.1556
	0.80	0.1561	0.0874	0.0588	0.0446	0.3514	0.2554	0.2063	0.1707
100	0.25	0.0284	0.0152	0.0128	0.0137	0.1464	0.1072	0.0980	0.1004
	0.50	0.0256	0.0148	0.0129	0.0143	0.1405	0.1052	0.0970	0.1015
	0.65	0.0279	0.0150	0.0132	0.0138	0.1448	0.1056	0.0987	0.0997
	0.80	0.0290	0.0149	0.0132	0.0129	0.1499	0.1056	0.0984	0.0966
250	0.25	0.0080	0.0059	0.0061	0.0085	0.0781	0.0682	0.0705	0.0846
	0.50	0.0071	0.0054	0.0064	0.0095	0.0735	0.0663	0.0729	0.0891
	0.65	0.0069	0.0056	0.0065	0.0085	0.0730	0.0668	0.0732	0.0840

	0.80	0.0078	0.0057	0.0061	0.0074	0.0771	0.0671	0.0701	0.0778
500	0.25	0.0031	0.0033	0.0046	0.0069	0.0501	0.0534	0.0637	0.0790
	0.50	0.0028	0.0035	0.0050	0.0079	0.0475	0.0543	0.0667	0.0847
	0.65	0.0030	0.0035	0.0050	0.0073	0.0483	0.0548	0.0667	0.0814
	0.80	0.0031	0.0031	0.0044	0.0064	0.0493	0.0515	0.0620	0.0757

Table 8: The Bias and MSE's of η_2^* measure of association for seven iid dichotomous variables

		MSE				Bias			
n	p	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.1433	0.1023	0.0687	0.0269	-0.3607	-0.3027	-0.2447	-0.1521
	0.50	0.2120	0.1510	0.0962	0.0339	-0.4470	-0.3719	-0.2954	-0.1737
	0.65	0.1942	0.1417	0.0900	0.0335	-0.4251	-0.3608	-0.2850	-0.1715
	0.80	0.0904	0.0675	0.0459	0.0206	-0.2815	-0.2400	-0.1963	-0.1303
100	0.25	0.1415	0.0843	0.0466	0.0158	-0.3690	-0.2835	-0.2101	-0.1216
	0.50	0.2891	0.1701	0.0929	0.0266	-0.5317	-0.4057	-0.2993	-0.1601
	0.65	0.2267	0.1315	0.0717	0.0223	-0.4696	-0.3563	-0.2624	-0.1457
	0.80	0.1034	0.0601	0.0342	0.0125	-0.3134	-0.2377	-0.1785	-0.1077
250	0.25	0.0948	0.0463	0.0227	0.0066	-0.3040	-0.2117	-0.1481	-0.0801
	0.50	0.1858	0.0949	0.0459	0.0122	-0.4278	-0.3051	-0.2118	-0.1092
	0.65	0.1458	0.0735	0.0343	0.0094	-0.3784	-0.2679	-0.1828	-0.0957
	0.80	0.0698	0.0343	0.0169	0.0053	-0.2601	-0.1819	-0.1277	-0.0715
500	0.25	0.0552	0.0231	0.0102	0.0027	-0.2324	-0.1502	-0.1001	-0.0513

0.50	0.0908	0.0390	0.0171	0.0041	-0.2993	-0.1959	-0.1296	-0.0640
0.65	0.0786	0.0329	0.0146	0.0035	-0.2779	-0.1799	-0.1196	-0.0591
0.80	0.0427	0.0182	0.0083	0.0023	-0.2042	-0.1329	-0.0902	-0.0479

Previous Figures show that the performance of the ρ_B and η_2^* measure for several numbers of variables k, while Figure 7 and Figure 8 clarify that ρ_B outperforms η_2^* for $30 \leq n \leq 500$ when the number of dichotomous variables increased to seven. However, when the sample size increases, the performance of ρ_B is better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_B are significantly less than the corresponding values of η_2^*. Furthermore, it can be noticed that increasing the number of dichotomous variables provide that ρ_B measure is more qualified for all cases compared to η_2^*.

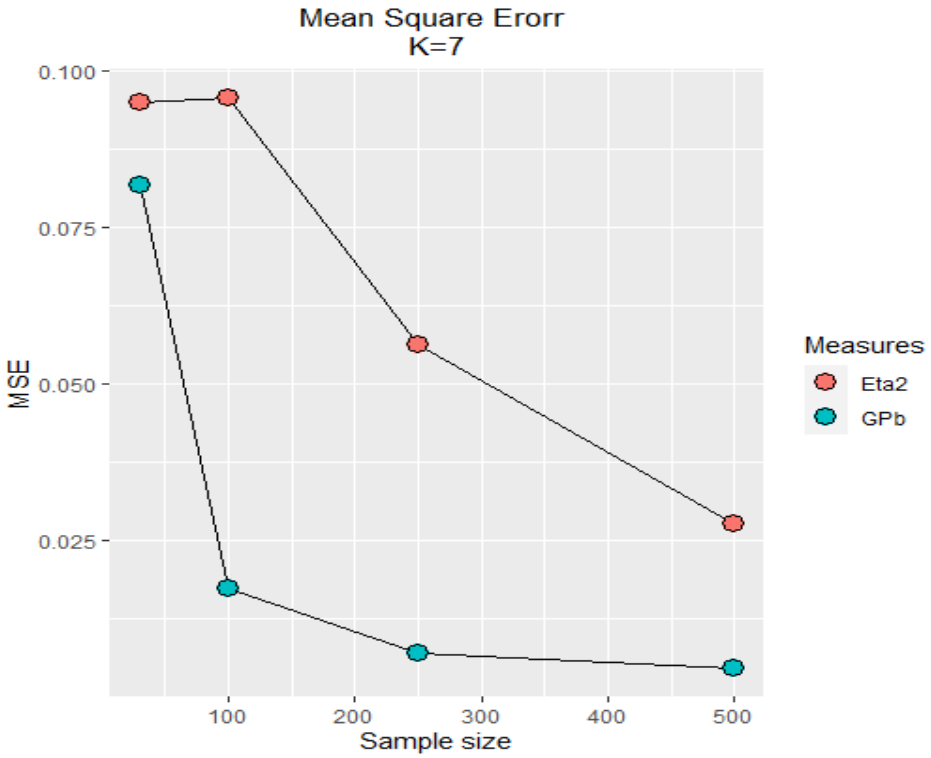


Figure 7. MSE of ρ_B versus MSE of η_2^* for seven iid dichotomous variables

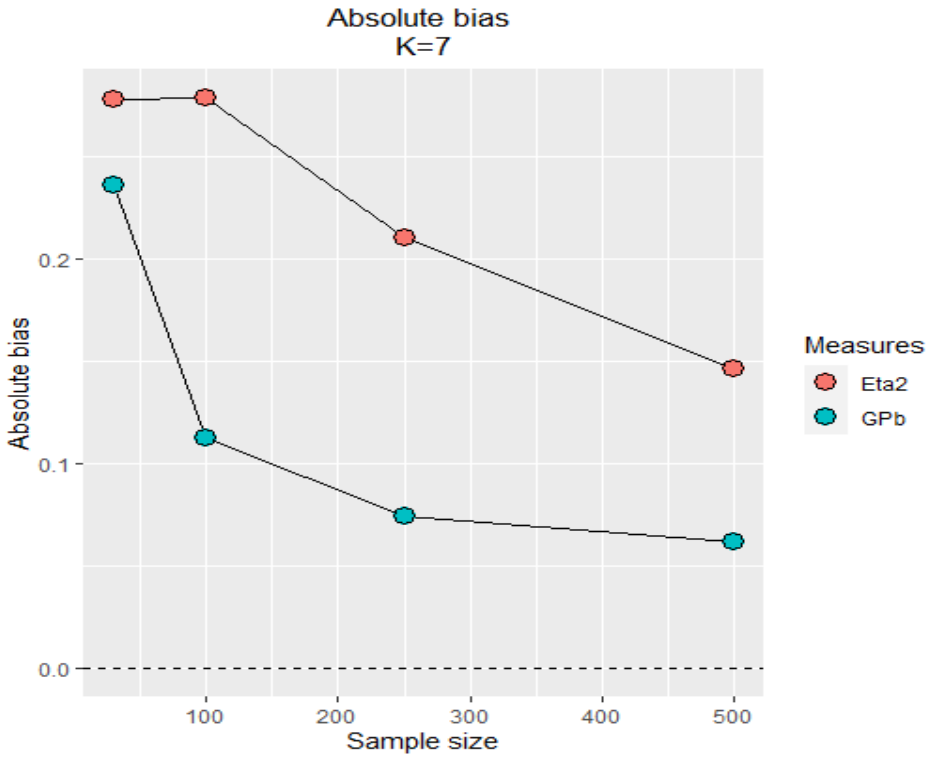


Figure 8. Absolute bias of ρ_B versus Absolute bias of η_2^* for seven iid dichotomous variables

4.1.2 Measure based on a set of independent but non-id dichotomous variables and a continuous

This section evaluates the performance of the generalized point biserial correlation coefficient when we have a set of non-id dichotomous. The design parameters that govern the generation of the simulated data are:

- K: the number of dichotomous variables k = 2, 3, 5, and 7.
- *n:* the sample size. We chose four settings for n = 30, 100, 250 and 500.
- ρ: the correlation coefficient between the two variables. Five settings are chosen for ρ = 0.25, 0.50, 0.70 and 0.95.

- p: the probability for each dichotomous variable is Four settings for each variable depends on the number of variables as follows:

- For k=2, the probabilities are p_1 = 0.25, 0.50, 0.65 and 0.80 and p_2 = 0.65, 0.70, 0.40 and 0.50.
- For k=3, the probabilities are p_1 = 0.25, 0.50, 0.65 and 0.80, p_2 = 0.65, 0.70, 0.40 and 0.50 and p_3 = 0.70, 0.40, 0.80 and 0.35.
- For k=5, the probabilities are p_1 = 0.25, 0.50, 0.65 and 0.80, p_2 = 0.65, 0.70, 0.40 and 0.50, p_3 = 0.70, 0.40, 0.80 and 0.35, p_4 = 0.35, 0.80, 0.20 and 0.90 and p_5 = 0.80, 0.85, 0.60 and 0.25.
- For k=7, the probabilities are p_1 = 0.25, 0.50, 0.65 and 0.80, p_2 = 0.65, 0.70, 0.40 and 0.50, p_3 = 0.70, 0.40, 0.80 and 0.35, p_4 = 0.35, 0.80, 0.20 and 0.90, p_5 = 0.80, 0.85, 0.60 and 0.25, p_6 = 0.30, 0.20, 0.90 and 0.70 and p_7 = 0.90, 0.75, 0.35 and 0.20.

Regarding the above parameters, the generating process was conducted based on a multiple linear regression that considered one dependent variable and k independent variables, where all variables were generated based on a multivariate normal distribution. Then, the independent variables were recoded to dichotomous variables based on the different probabilities for each independent variable that were considered as cut points. After that, the proposed measures and η_2^* measure were applied on the simulated data, followed by the computation of goodness-of-fit criteria, Bias, and MSE.

4.1.2.1 Results and comparison

The following tables and charts present the results of the simulation study on how ρ_{PB} measure performs to detect the association between a set of dichotomous

variables and a continuous variable compared to η_2^* measure considering different scenarios of sample sizes and the number of dichotomous variables as was explained before.

Table 9 and Table 10 contains the results of MSE and Bias that were calculated to assess the performance of ρ_{PB} measure when the number of dichotomous variables is two. In contrast, Table 11 and Table 12 contains the results of the same MSE and Bias to assess the performance of η_2^* measure under the same criteria. Comparing Table 9 and Table 10 to Table 11 and Table 12 shows that the Bias and MSE do not indicate satisfactory performance when the sample size is 30. However, the results suggest that as the sample size increases, the performance of the considered association measures increases by giving smaller MSE and Bias. For instance, Table 9 and Table 10 show that when P is 0.25 and the correlation coefficient equals 0.25, the MSE and bias of ρ_{PB} for a sample size of 30 were greater than MSE and bias of η_2^* for a sample size of 30.

Table 9: The MSE's of ρ_{PB} measure of association for two non-id dichotomous variables

			MSE			
n	P1	P2	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.0317	0.0123	0.0091	0.0098
	0.50	0.70	0.0290	0.0136	0.0116	0.0117
	0.65	0.40	0.0330	0.0108	0.0099	0.0119
	0.80	0.50	0.0265	0.0118	0.0083	0.0084
100	0.25	0.65	0.0036	0.0022	0.0026	0.0036

	0.50	0.70	0.0038	0.0030	0.0036	0.0045
	0.65	0.40	0.0039	0.0026	0.0030	0.0040
	0.80	0.50	0.0033	0.0019	0.0023	0.0031
250	0.25	0.65	0.0009	0.0010	0.0013	0.0020
	0.50	0.70	0.0013	0.0015	0.0018	0.0030
	0.65	0.40	0.0008	0.0011	0.0014	0.0024
	0.80	0.50	0.0007	0.0007	0.0010	0.0014
500	0.25	0.65	0.0004	0.0006	0.0010	0.0014
	0.50	0.70	0.0005	0.0009	0.0014	0.0023
	0.65	0.40	0.0004	0.0007	0.0011	0.0017
	0.80	0.50	0.0003	0.0004	0.0006	0.0009

Table 10: The Bias of ρ_{PB} measure of association for two non-id dichotomous variables

			Bias			
n	P1	P2	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.1148	0.0639	0.0536	0.0557
	0.50	0.70	0.1077	0.0712	0.0627	0.0659
	0.65	0.40	0.1129	0.0623	0.0580	0.0649
	0.80	0.50	0.1019	0.0640	0.0500	0.0463
100	0.25	0.65	0.0367	0.0289	0.0329	0.0379
	0.50	0.70	0.0386	0.0368	0.0418	0.0480
	0.65	0.40	0.0379	0.0330	0.0361	0.0431
	0.80	0.50	0.0347	0.0257	0.0296	0.0339

250	0.25	0.65	0.0188	0.0225	0.0271	0.0356
	0.50	0.70	0.0229	0.0287	0.0346	0.0465
	0.65	0.40	0.0184	0.0243	0.0280	0.0398
	0.80	0.50	0.0175	0.0181	0.0218	0.0276
500	0.25	0.65	0.0134	0.0189	0.0262	0.0329
	0.50	0.70	0.0156	0.0255	0.0331	0.0440
	0.65	0.40	0.0145	0.0223	0.0287	0.0368
	0.80	0.50	0.0109	0.0154	0.0201	0.0255

Table 11: The MSE's of η_2^* measure of association for two non-id dichotomous variables

			MSE			
n	P1	P2	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.0196	0.0086	0.0034	0.0008
	0.50	0.70	0.0166	0.0080	0.0028	0.0005
	0.65	0.40	0.0154	0.0067	0.0026	0.0005
	0.80	0.50	0.0253	0.0121	0.0046	0.0010
100	0.25	0.65	0.0168	0.0043	0.0014	0.0003
	0.50	0.70	0.0110	0.0023	0.0006	0.0001
	0.65	0.40	0.0092	0.0019	0.0006	0.0001
	0.80	0.50	0.0190	0.0062	0.0024	0.0005
250	0.25	0.65	0.0142	0.0033	0.0010	0.0002
	0.50	0.70	0.0089	0.0014	0.0004	0.0000

	0.65	0.40	0.0067	0.0012	0.0003	0.0000
	0.80	0.50	0.0173	0.0051	0.0020	0.0004
500	0.25	0.65	0.0135	0.0030	0.0010	0.0002
	0.50	0.70	0.0081	0.0012	0.0003	0.0000
	0.65	0.40	0.0062	0.0010	0.0003	0.0000
	0.80	0.50	0.0167	0.0049	0.0019	0.0004

Table 12: The Bias of η_2^* measure of association for two non-id dichotomous variables

			Bias			
n	P1	P2	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	-0.1206	-0.0757	-0.0474	-0.0234
	0.50	0.70	-0.1085	-0.0692	-0.0401	-0.0179
	0.65	0.40	-0.1024	-0.0640	-0.0392	-0.0178
	0.80	0.50	-0.1406	-0.0947	-0.0587	-0.0289
100	0.25	0.65	-0.1218	-0.0615	-0.0357	-0.0169
	0.50	0.70	-0.0959	-0.0438	-0.0236	-0.0110
	0.65	0.40	-0.0865	-0.0397	-0.0231	-0.0106
	0.80	0.50	-0.1314	-0.0759	-0.0477	-0.0234
250	0.25	0.65	-0.1158	-0.0561	-0.0321	-0.0150
	0.50	0.70	-0.0900	-0.0360	-0.0203	-0.0094
	0.65	0.40	-0.0780	-0.0334	-0.0189	-0.0087
	0.80	0.50	-0.1289	-0.0707	-0.0446	-0.0215
500	0.25	0.65	-0.1144	-0.0544	-0.0313	-0.0146

0.50	0.70	-0.0880	-0.0345	-0.0191	-0.0089
0.65	0.40	-0.0766	-0.0316	-0.0178	-0.0081
0.80	0.50	-0.1279	-0.0697	-0.0435	-0.0211

For more insight into the performance of ρ_{PB} and η_2^* measures, Figure 9 and Figure 10 were developed to reflect a clear vision about the behaviors of the considered measures regarding the sample size, in addition to a visual comparison between the two measures, where Figure 9 depicts the behavior of MSE of the two measures versus sample size, while Figure 10 depicts the behavior of the absolute Bias of the two measures versus sample size.

As it is well known about MSE and absolute bias as the goodness-of-fit criteria, the smallest MSE, and absolute bias the better the measure's performance. Clearly, it can be noticed that the performance of η_2^* is better than ρ_{PB} at sample size 30 where η_2^* produced less values of MSE and absolute bias than what ρ_{PB} produced. However, when the sample size is larger than 30, the performance of ρ_{PB} substantially is better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_{PB} were significantly less than the corresponding values of η_2^*. Moreover, the absolute Bias and MSE converge reasonably well to zero when the sample size increases.

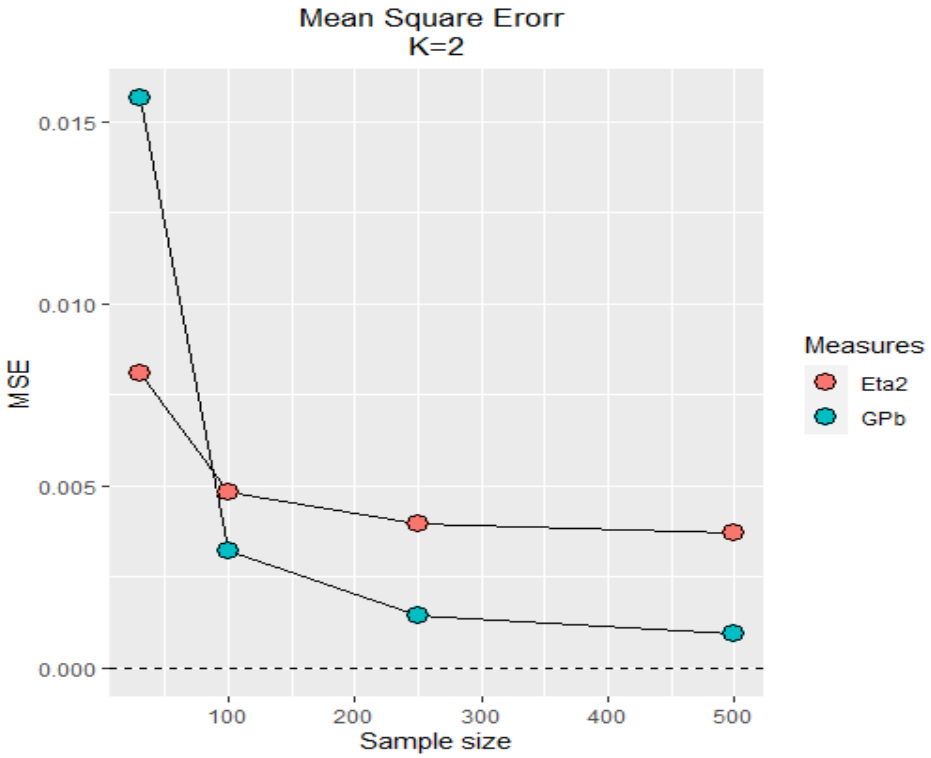


Figure 9. MSE of ρ_{PB} versus Absoulte bias of η_2^* for two non-id dichotomous variables

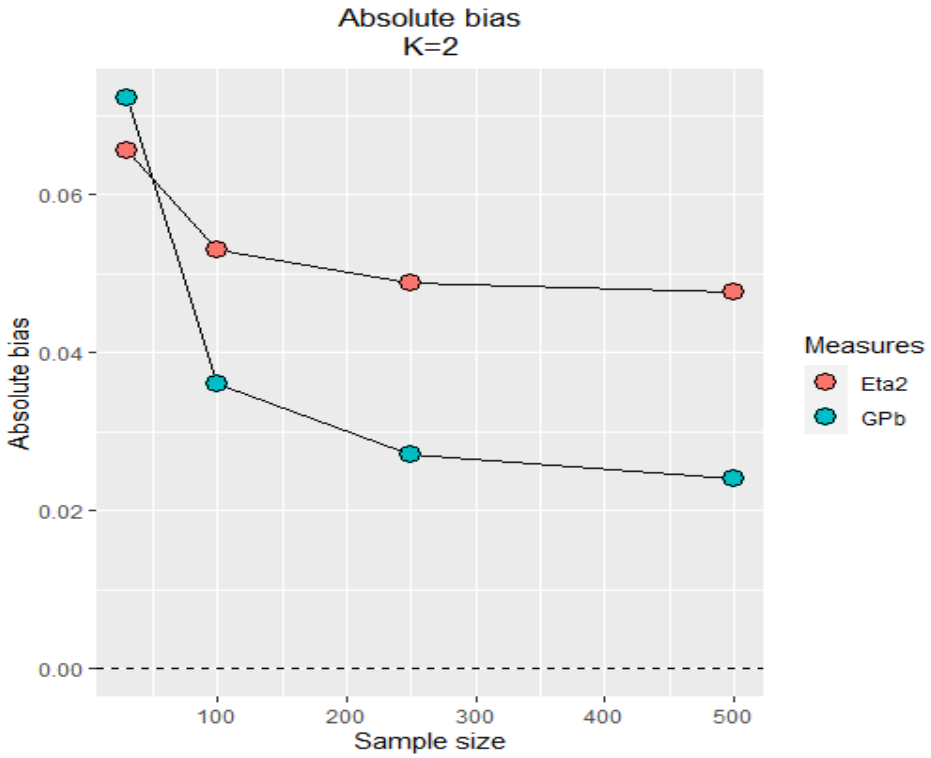


Figure 10. Absolute bias of ρ_{PB} versus Absolute bias of η_2^* for two non-id dichotomous variables

In addition, Table 13 and Table 14 contains the results of MSE and Bias that were calculated to assess the performance of ρ_{PB} measure when the number of dichotomous variables increased to three. On the other hand, Table 15 and Table 16 contains the results of the same MSE and Bias to assess the performance of η_2^* measure under the same criteria. Comparing the four tables below shows that ρ_{PB} and η_2^* measures had almost similar results when the number of dichotomous variables was two that the Bias and MSE do not indicate satisfactory performance when the sample size is less than 100. For instance, Tables 13 and 15 show that when P is 0.25 and the correlation coefficient equals 0.25, the MSE and bias of ρ_{PB} for a sample size of 30 were greater than MSE and bias of η_2^* for a sample size of 30.

Table 13: The MSE's of ρ_{PB} measure of association for three non-id dichotomous variables

				MSE			
n	P1	P2	P3	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.0432	0.0200	0.0126	0.0113
	0.50	0.70	0.40	0.0494	0.0192	0.0123	0.0117
	0.65	0.40	0.80	0.0520	0.0187	0.0146	0.0126
	0.80	0.50	0.35	0.0485	0.0168	0.0117	0.0101
100	0.25	0.65	0.70	0.00679	0.0028	0.0029	0.0033
	0.50	0.70	0.40	0.00621	0.0029	0.0028	0.0038

	0.65	0.40	0.80	0.00630	0.0033	0.0030	0.0037
	0.80	0.50	0.35	0.00659	0.0027	0.0024	0.0032
250	0.25	0.65	0.70	0.00143	0.0010	0.0011	0.0016
	0.50	0.70	0.40	0.00157	0.0011	0.0013	0.0018
	0.65	0.40	0.80	0.00158	0.0011	0.0014	0.0020
	0.80	0.50	0.35	0.00134	0.0008	0.0009	0.0014
500	0.25	0.65	0.70	0.00051	0.0005	0.0007	0.0010
	0.50	0.70	0.40	0.00059	0.0006	0.0008	0.0014
	0.65	0.40	0.80	0.00065	0.0006	0.0009	0.0014
	0.80	0.50	0.35	0.00047	0.0004	0.0005	0.0008

Table 14: The Bias of ρ_{PB} measure of association for three non-id dichotomous variables

				Bias			
n	P1	P2	P3	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.1568	0.0971	0.0705	0.0578
	0.50	0.70	0.40	0.1644	0.0959	0.0679	0.0627
	0.65	0.40	0.80	0.1691	0.0990	0.0795	0.0624
	0.80	0.50	0.35	0.1617	0.0886	0.0675	0.0506
100	0.25	0.65	0.70	0.0585	0.0359	0.0339	0.0323
	0.50	0.70	0.40	0.0569	0.0365	0.0330	0.0386
	0.65	0.40	0.80	0.0587	0.0403	0.0366	0.0387

	0.80	0.50	0.35	0.0571	0.0347	0.0295	0.0304
250	0.25	0.65	0.70	0.0274	0.0225	0.0219	0.0275
	0.50	0.70	0.40	0.0286	0.0243	0.0263	0.0296
	0.65	0.40	0.80	0.0279	0.0252	0.0272	0.0334
	0.80	0.50	0.35	0.0255	0.0195	0.0198	0.0230
500	0.25	0.65	0.70	0.0160	0.0170	0.0198	0.0243
	0.50	0.70	0.40	0.0175	0.0198	0.0221	0.0297
	0.65	0.40	0.80	0.0184	0.0202	0.0248	0.0321
	0.80	0.50	0.35	0.0156	0.0147	0.0171	0.0212

Table 15: The MSE's of η_2^* measure of association for three non-id dichotomous variables

				MSE			
n	P1	P2	P3	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.0317	0.0202	0.0084	0.0024
	0.50	0.70	0.40	0.0384	0.0196	0.0098	0.0024
	0.65	0.40	0.80	0.03146	0.01879	0.00910	0.00233
	0.80	0.50	0.35	0.0295	0.0183	0.0088	0.0021
100	0.25	0.65	0.70	0.0132	0.0046	0.0017	0.0003
	0.50	0.70	0.40	0.0098	0.0030	0.0011	0.0002
	0.65	0.40	0.80	0.0130	0.0037	0.0015	0.0003
	0.80	0.50	0.35	0.0113	0.0039	0.0014	0.0003
250	0.25	0.65	0.70	0.0076	0.0016	0.0006	0.0001

	0.50	0.70	0.40	0.0031	0.0007	0.0002	0.0000
	0.65	0.40	0.80	0.0059	0.0010	0.0003	0.0000
	0.80	0.50	0.35	0.0053	0.0012	0.0004	0.0001
500	0.25	0.65	0.70	0.0055	0.0011	0.0003	0.0000
	0.50	0.70	0.40	0.0015	0.0002	0.00010	0.0000
	0.65	0.40	0.80	0.0040	0.0005	0.0001	0.0000
	0.80	0.50	0.35	0.0035	0.0007	0.0002	0.0000

Table 16: The Bias of η_2^* measure of association for three non-id dichotomous variables

				Bias			
n	P1	P2	P3	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	-0.1522	-0.1177	-0.0753	-0.0396
	0.50	0.70	0.40	-0.1667	-0.1153	-0.0801	-0.0403
	0.65	0.40	0.80	-0.1514	-0.1133	-0.0776	-0.0386
	0.80	0.50	0.35	-0.1456	-0.1108	-0.0745	-0.0375
100	0.25	0.65	0.70	-0.1045	-0.0603	-0.0371	-0.0179
	0.50	0.70	0.40	-0.0862	-0.0471	-0.0288	-0.0135
	0.65	0.40	0.80	-0.1019	-0.0539	-0.0336	-0.0153
	0.80	0.50	0.35	-0.0957	-0.0553	-0.0337	-0.0166
250	0.25	0.65	0.70	-0.0824	-0.0386	-0.0232	-0.0115
	0.50	0.70	0.40	-0.0499	-0.0246	-0.0141	-0.0067
	0.65	0.40	0.80	-0.0709	-0.0300	-0.0173	-0.0082
	0.80	0.50	0.35	-0.0682	-0.0328	-0.0197	-0.0094

500	0.25	0.65	0.70	-0.0723	-0.0328	-0.0190	-0.0089
	0.50	0.70	0.40	-0.0365	-0.0158	-0.0091	-0.0042
	0.65	0.40	0.80	-0.0603	-0.0212	-0.0123	-0.0059
	0.80	0.50	0.35	-0.0568	-0.0257	-0.0149	-0.0070

For more clarify in the performance of ρ_{PB} and η_2^* measure, Figures 11 and 12 were created to reflect a clear vision of the behaviors of the considered measures in terms of sample size and a visual comparison between the two measures. Where Figure 11 depicts the behavior of MSE of the two measures versus sample size, while Figure 12 depicts the behavior of the absolute Bias of the two measures versus sample size.

As it is well known about MSE and absolute bias as the goodness-of-fit criteria, the smallest MSE, and absolute bias, the better the performance of the measure. Clearly, it can be noticed that the performance of η_2^* is better than ρ_{PB} when the sample size was less than 100 where η_2^* produced smaller values of MSE and absolute bias than what ρ_{PB} produced. However, when the sample size is larger than 30 the performance of ρ_{PB} slightly is better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_{PB}were significantly less than the corresponding values of η_2^*. Moreover, Figure 11 shows that performance of the ρ_{PB} in terms of MSE versus the sample size n, almost give same performance compared to η_2^* at sample size 100, while Figure 12 depicts the behavior of the absolute Bias of the two measures versus absolute bias clarify that the ρ_{PB} is outperform compared to η_2^*.

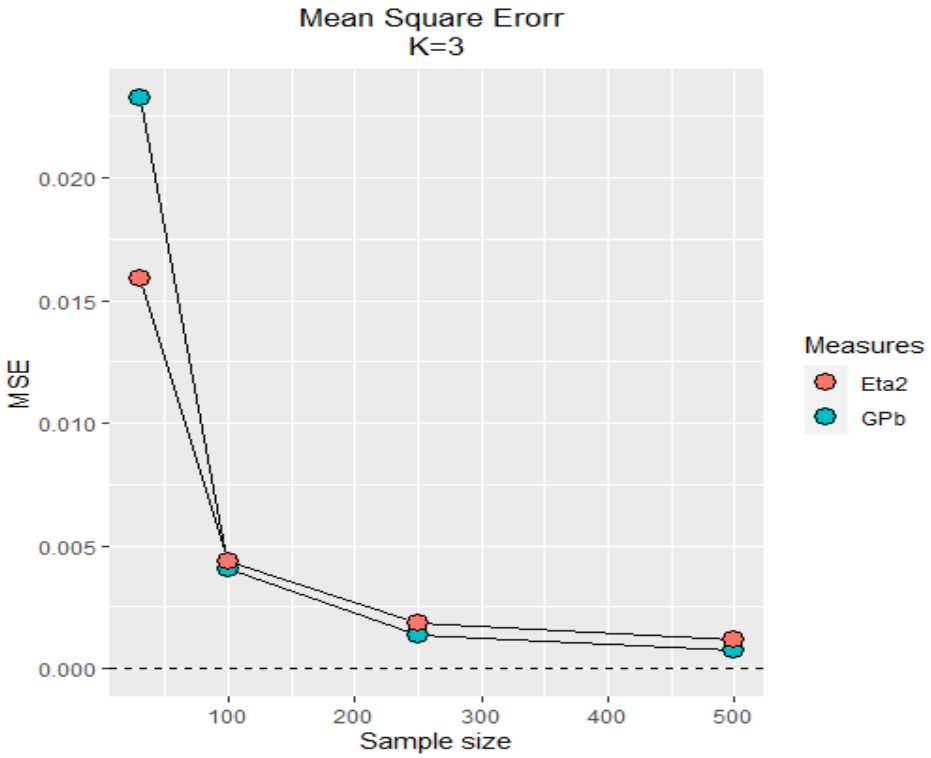


Figure 11. MSE of ρ_{PB} versus MSE of η_2^* for three non-id dichotomous variables

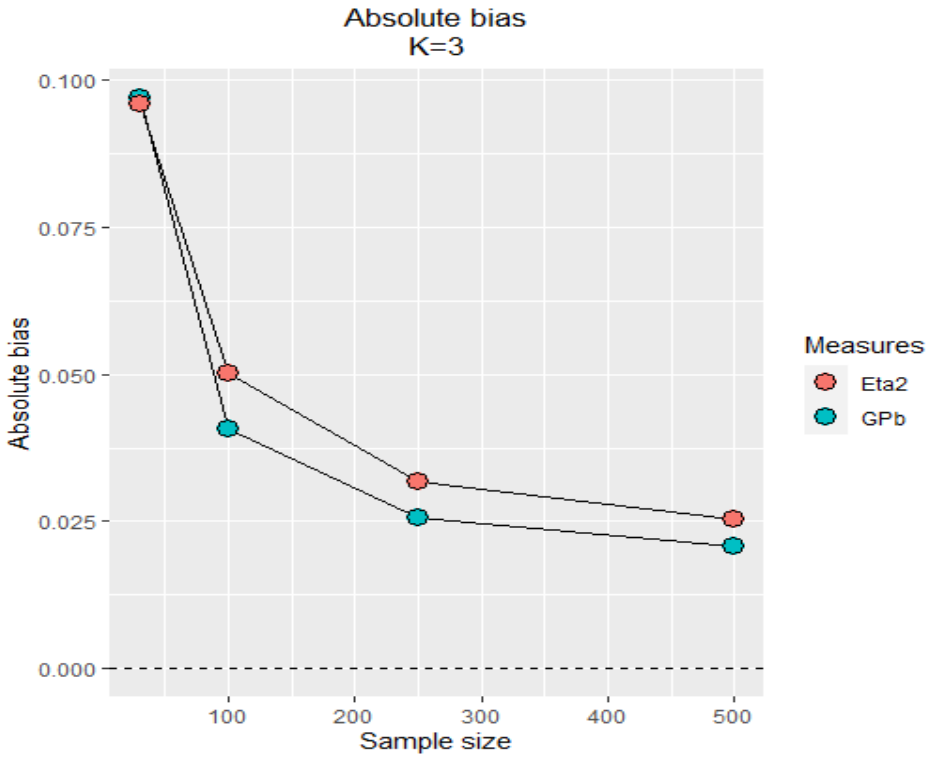


Figure 12. Absolute bias of ρ_{PB} versus Absolute bias of η_2^* for three non-id dichotomous variables

The tables below present the results of ρ_{PB} and η_2^* to show the performance when the number of dichotomous variables increased to five variables.

Table17 and Table 18 contains the results of MSE and Bias that were calculated to assess the performance of ρ_{PB} .In addition, Table 19 and Table 20 shows the results of the same MSE and Bias to evaluate the performance of η_2^* measure under the same goodness-of-fit criteria. However, when all tables below are compared, Bias and MSE indicate the ρ_{PB} performing satisfactorily for all considered sample sizes, even for a small sample size.

Table 17: The MSE's of ρ_{PB} measure of association for five non-id dichotomous variables

						MSE			
n	P1	P2	P3	P4	P5	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.0934	0.0459	0.0304	0.0201
	0.50	0.70	0.40	0.80	0.85	0.0978	0.0504	0.0342	0.0236
	0.65	0.40	0.80	0.20	0.60	0.0914	0.0474	0.0300	0.0228
	0.80	0.50	0.35	0.90	0.25	0.0936	0.0510	0.0315	0.0223
100	0.25	0.65	0.70	0.35	0.80	0.0162	0.0075	0.0061	0.0066
	0.50	0.70	0.40	0.80	0.85	0.0182	0.0084	0.0069	0.0078
	0.65	0.40	0.80	0.20	0.60	0.0154	0.0080	0.0064	0.0067
	0.80	0.50	0.35	0.90	0.25	0.0166	0.0084	0.0063	0.0060
250	0.25	0.65	0.70	0.35	0.80	0.0038	0.0024	0.0027	0.0036
	0.50	0.70	0.40	0.80	0.85	0.0047	0.0030	0.0034	0.0045
	0.65	0.40	0.80	0.20	0.60	0.0039	0.0028	0.0031	0.0038
	0.80	0.50	0.35	0.90	0.25	0.0043	0.0026	0.0028	0.0038
500	0.25	0.65	0.70	0.35	0.80	0.0013	0.0013	0.0017	0.0024
	0.50	0.70	0.40	0.80	0.85	0.0017	0.0018	0.0023	0.0035
	0.65	0.40	0.80	0.20	0.60	0.0015	0.0013	0.0019	0.0029
	0.80	0.50	0.35	0.90	0.25	0.0015	0.0014	0.0019	0.0028

Table 18: The Bias of ρ_{PB} measure of association for five non-id dichotomous variables

						Bias			
n	P1	P2	P3	P4	P5	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.25978	0.1747	0.1339	0.0978

	0.50	0.70	0.40	0.80	0.85	0.2638	0.1833	0.1481	0.1123
	0.65	0.40	0.80	0.20	0.60	0.2505	0.1758	0.1333	0.1095
	0.80	0.50	0.35	0.90	0.25	0.2594	0.1852	0.1408	0.1065
100	0.25	0.65	0.70	0.35	0.80	0.1049	0.0697	0.0606	0.0597
	0.50	0.70	0.40	0.80	0.85	0.1104	0.0739	0.0656	0.0696
	0.65	0.40	0.80	0.20	0.60	0.1038	0.0725	0.0633	0.0626
	0.80	0.50	0.35	0.90	0.25	0.1063	0.0738	0.0642	0.0571
250	0.25	0.65	0.70	0.35	0.80	0.0510	0.0408	0.0430	0.0504
	0.50	0.70	0.40	0.80	0.85	0.0581	0.0458	0.0497	0.0586
	0.65	0.40	0.80	0.20	0.60	0.0523	0.0434	0.0471	0.0512
	0.80	0.50	0.35	0.90	0.25	0.0550	0.0427	0.0440	0.0522
500	0.25	0.65	0.70	0.35	0.80	0.0307	0.0315	0.0364	0.0439
	0.50	0.70	0.40	0.80	0.85	0.0350	0.0369	0.0437	0.0555
	0.65	0.40	0.80	0.20	0.60	0.0323	0.0321	0.0392	0.0489
	0.80	0.50	0.35	0.90	0.25	0.0322	0.0333	0.0384	0.0475

Table 19: The MSE's of η_2^* measure of association for five non-id dichotomous variables

						MSE			
n	P1	P2	P3	P4	P5	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.1000	0.0666	0.0436	0.0154
	0.50	0.70	0.40	0.80	0.85	0.0899	0.0633	0.0396	0.0147
	0.65	0.40	0.80	0.20	0.60	0.0959	0.0684	0.0416	0.0153

	0.80	0.50	0.35	0.90	0.25	0.0711	0.0491	0.0320	0.0112
100	0.25	0.65	0.70	0.35	0.80	0.0596	0.0312	0.0139	0.0039
	0.50	0.70	0.40	0.80	0.85	0.0558	0.0285	0.0146	0.0041
	0.65	0.40	0.80	0.20	0.60	0.0597	0.0292	0.0145	0.0038
	0.80	0.50	0.35	0.90	0.25	0.0439	0.0232	0.0114	0.0033
250	0.25	0.65	0.70	0.35	0.80	0.0248	0.0093	0.0039	0.0009
	0.50	0.70	0.40	0.80	0.85	0.0244	0.0095	0.0040	0.0010
	0.65	0.40	0.80	0.20	0.60	0.0249	0.0094	0.0038	0.0009
	0.80	0.50	0.35	0.90	0.25	0.0215	0.0080	0.0034	0.0008
500	0.25	0.65	0.70	0.35	0.80	0.0106	0.0032	0.0012	0.0002
	0.50	0.70	0.40	0.80	0.85	0.0106	0.0033	0.0013	0.0003
	0.65	0.40	0.80	0.20	0.60	0.0102	0.0030	0.0012	0.0002
	0.80	0.50	0.35	0.90	0.25	0.0099	0.0031	0.0012	0.0003

Table 20: The Bias of η_2^* measure of association for five non-id dichotomous variables

						Bias			
n	P1	P2	P3	P4	P5	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	-0.2941	-0.2379	-0.1897	-0.1115
	0.50	0.70	0.40	0.80	0.85	-0.2778	-0.2296	-0.1800	-0.1088
	0.65	0.40	0.80	0.20	0.60	-0.2888	-0.2414	-0.1854	-0.1115
	0.80	0.50	0.35	0.90	0.25	-0.2448	-0.1991	-0.1609	-0.0938
100	0.25	0.65	0.70	0.35	0.80	-0.2340	-0.1676	-0.1116	-0.0599
	0.50	0.70	0.40	0.80	0.85	-0.2256	-0.1600	-0.1143	-0.0605

	0.65	0.40	0.80	0.20	0.60	-0.2344	-0.1629	-0.1137	-0.0585
	0.80	0.50	0.35	0.90	0.25	-0.1995	-0.1431	-0.0993	-0.0542
250	0.25	0.65	0.70	0.35	0.80	-0.1520	-0.0930	-0.0601	-0.0296
	0.50	0.70	0.40	0.80	0.85	-0.1504	-0.0929	-0.0606	-0.0307
	0.65	0.40	0.80	0.20	0.60	-0.1522	-0.0933	-0.0592	-0.0292
	0.80	0.50	0.35	0.90	0.25	-0.1407	-0.0854	-0.0564	-0.0285
500	0.25	0.65	0.70	0.35	0.80	-0.0994	-0.0550	-0.0337	-0.0161
	0.50	0.70	0.40	0.80	0.85	-0.0996	-0.0557	-0.0347	-0.0169
	0.65	0.40	0.80	0.20	0.60	-0.0978	-0.0532	-0.0336	-0.0161
	0.80	0.50	0.35	0.90	0.25	-0.0961	-0.0543	-0.0343	-0.0167

As mentioned earlier, Figures always clarify the performance of ρ_{PB} and η_2^* measure. Figure 13 indicates that the MSE of ρ_{PB} for all cases is outperforms compared to η_2^*. Furthermore, Figure 14 depicts the behavior of the absolute Bias of the two measures versus sample size. Clearly, it can be noticed that even absolute Bias of ρ_{PB} is outperforms the corresponding values of η_2^*. However, it can be seen that at sample size 500, results for the goodness-of-fits criteria shows that ρ_{PB} and η_2^* has almost similar results, but on average, still ρ_{PB} has better performance than η_2^*.

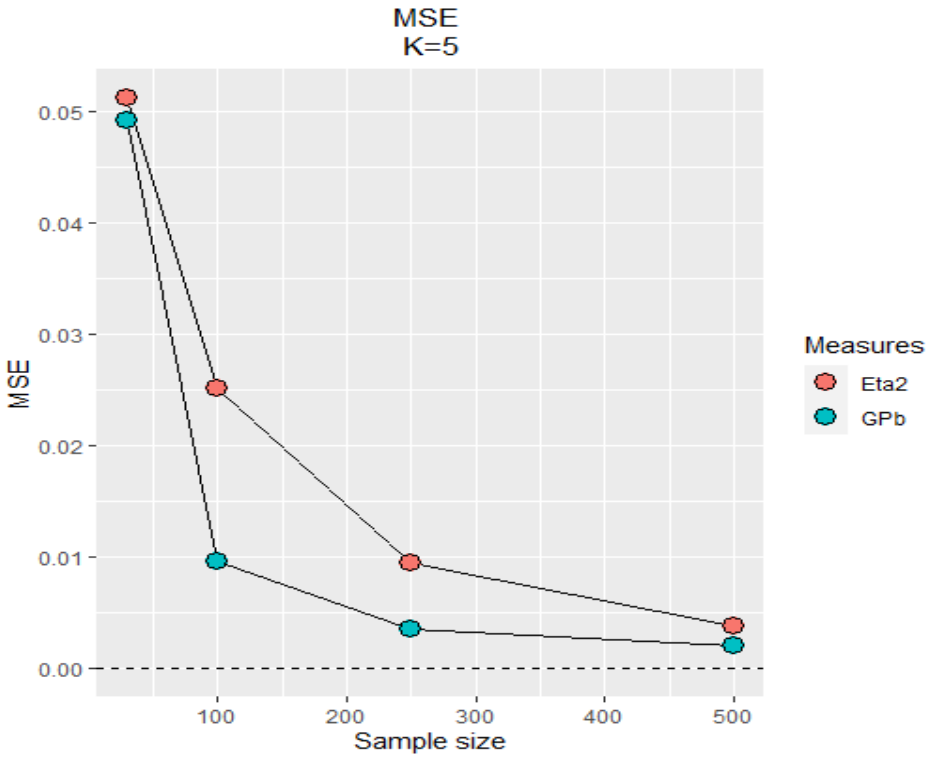


Figure 13. MSE of ρ_{PB} versus MSE of η_2^* for five non-id dichotomous variables

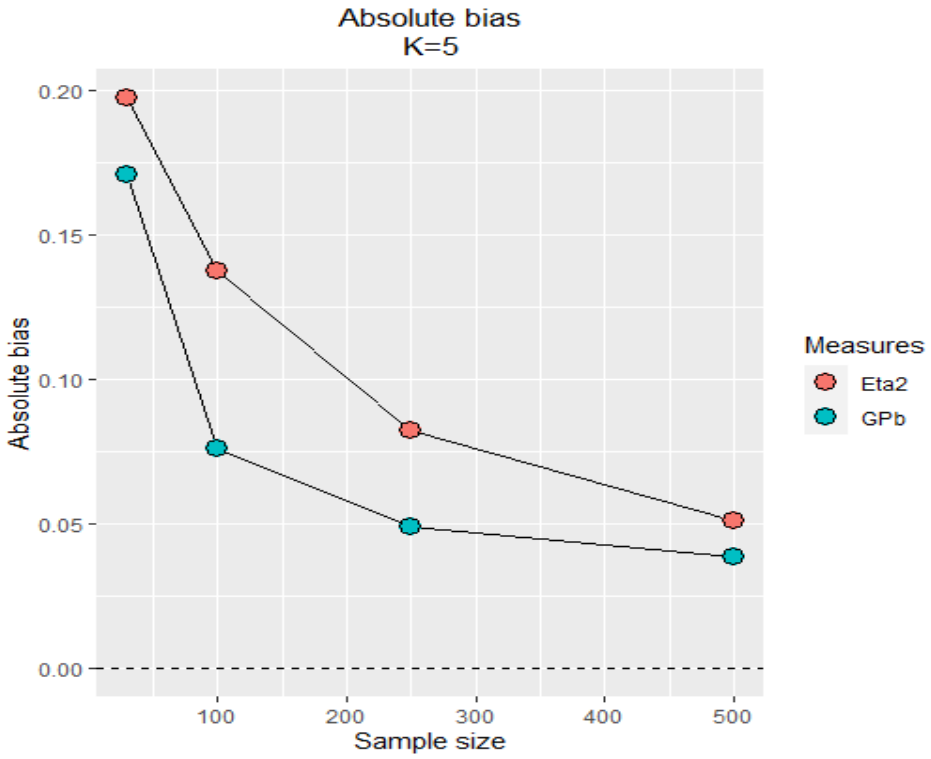


Figure 14. Absolute bias of ρ_{PB} versus Absolute bias of η_2^* for five non-id dichotomous variables

Table 22 presents the best performance of the ρ_{PB} measure for all considered sample sizes when the number of dichotomous variables is seven compared to η_2^* performance in Table 24 regarding to Bias criteria results. In contrast, values of MSE in Table 21 and Table 23 at sample size 30 contains the results of ρ_{PB} indicate less performance compared to η_2^*. However, the results provide that as the sample size and dichotomous variables increases, the performance of the association measure ρ_{PB} increases by giving smaller MSE and Bias. Further, the below figures will clarify that the performance of measures.

Table 21: The MSE's of ρ_{PB} measure of association for seven non-id dichotomous variables

								MSE			
n	P1	P2	P3	P4	P5	P6	P7	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.1445	0.0947	0.0632	0.0465
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.1617	0.0903	0.0638	0.0457
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.1502	0.0878	0.0586	0.0466
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.1559	0.0859	0.0571	0.0417
100	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0276	0.0167	0.0152	0.0139
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0296	0.0181	0.0160	0.0176
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0287	0.0174	0.0145	0.0155
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0300	0.0161	0.0148	0.0145
250	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0084	0.0062	0.0072	0.0093
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0083	0.0071	0.0088	0.0113

	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0086	0.0066	0.0079	0.0104
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0079	0.0062	0.0067	0.0092
500	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0034	0.0038	0.0051	0.0075
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0038	0.0045	0.0063	0.0098
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0036	0.0042	0.0056	0.0088
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0031	0.0039	0.0050	0.0077

Table 22: The Bias of ρ_{PB} measure of association for seven non-id dichotomous variables

								Bias			
n	P1	P2	P3	P4	P5	P6	P7	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.3385	0.2673	0.2161	0.1762
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.3622	0.2649	0.2147	0.1735
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.3469	0.2586	0.2065	0.1763
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.3539	0.2542	0.1996	0.1641
100	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.1448	0.1119	0.1066	0.1006
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.1521	0.1182	0.1123	0.1175
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.1485	0.1151	0.1049	0.1091
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.1513	0.1102	0.1064	0.1042
250	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0815	0.0700	0.0779	0.0893
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0804	0.0762	0.0864	0.0986
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0823	0.0722	0.0811	0.0944
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0781	0.0702	0.0746	0.0882
500	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0519	0.0569	0.0674	0.0830
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0549	0.0621	0.0750	0.0950

	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0529	0.0599	0.0704	0.0896
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0500	0.0573	0.0667	0.0836

Table 23: The MSE's of η_2^* measure of association for seven non-id dichotomous variables

								MSE			
n	P1	P2	P3	P4	P5	P6	P7	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.1355	0.1014	0.0706	0.0348
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.1382	0.1016	0.0712	0.0311
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.1345	0.0975	0.0719	0.0295
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.1266	0.0920	0.0631	0.0270
100	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.1391	0.0859	0.0506	0.0186
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.1469	0.0869	0.0504	0.0174
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.1392	0.0872	0.0490	0.0171
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.1286	0.0766	0.0427	0.0149
250	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0896	0.0454	0.0227	0.0073
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0931	0.0475	0.0234	0.0068
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0896	0.0457	0.0217	0.0066
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0849	0.0401	0.0197	0.0059
500	0.25	0.65	0.70	0.35	0.80	0.30	0.90	0.0525	0.0222	0.0104	0.0030
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	0.0553	0.0230	0.0104	0.0029
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	0.0517	0.0219	0.0097	0.0026
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	0.0487	0.0198	0.0090	0.0024

Table 24: The Bias of η_2^* measure of association for seven non-id dichotomous variables

								Bias			
n	P1	P2	P3	P4	P5	P6	P7	ρ_1=0.25	ρ_2=0.50	ρ_3=0.70	ρ_4=0.95
30	0.25	0.65	0.70	0.35	0.80	0.30	0.90	-0.3504	-0.3006	-0.2496	-0.1726
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	-0.3543	-0.3019	-0.2495	-0.1628
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	-0.3498	-0.2943	-0.2507	-0.1599
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	-0.3378	-0.2850	-0.2329	-0.1518
100	0.25	0.65	0.70	0.35	0.80	0.30	0.90	-0.3657	-0.2859	-0.2186	-0.1322
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	-0.3762	-0.2876	-0.2182	-0.1278
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	-0.3660	-0.2878	-0.2150	-0.1266
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	-0.3513	-0.2693	-0.2005	-0.1180
250	0.25	0.65	0.70	0.35	0.80	0.30	0.90	-0.2953	-0.2093	-0.1482	-0.0838
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	-0.3010	-0.2146	-0.1504	-0.0813
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	-0.2954	-0.2102	-0.1448	-0.0801
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	-0.2868	-0.1967	-0.1376	-0.0754
500	0.25	0.65	0.70	0.35	0.80	0.30	0.90	-0.2264	-0.147	-0.1008	-0.0541
	0.50	0.70	0.40	0.80	0.85	0.20	0.75	-0.2326	-0.1499	-0.1008	-0.0532
	0.65	0.40	0.80	0.20	0.60	0.90	0.35	-0.2249	-0.1462	-0.0971	-0.0509
	0.80	0.50	0.35	0.90	0.25	0.70	0.20	-0.2180	-0.1389	-0.0938	-0.0486

Previous Figures show that the performance of the ρ_{PB} and η_2^* measure for several numbers of variables k, while Figure 15 and Figure 16 clarify that ρ_{PB}outperforms η_2^* for all considered samples size when the number of dichotomous variables increased to seven. However, when the sample size increses the performance

of ρ_{PB} is better than the performance of η_2^*, where the values of MSE and the absolute bias of ρ_{PB} are significantly less than the corresponding values of η_2^*. Furthermore, it can be noticed that Figure 15 and Figure 16 show that the performance of η_2^* is better than ρ_{PB} at sample size 30 where η_2^* gives smaller the MSE. In contrast, values of absolute bias contain the results of ρ_{PB} indicate better performance compared to η_2^*.

Figure 15. MSE of ρ_{PB} versus MSE of η_2^* for seven non-id dichotomous variables

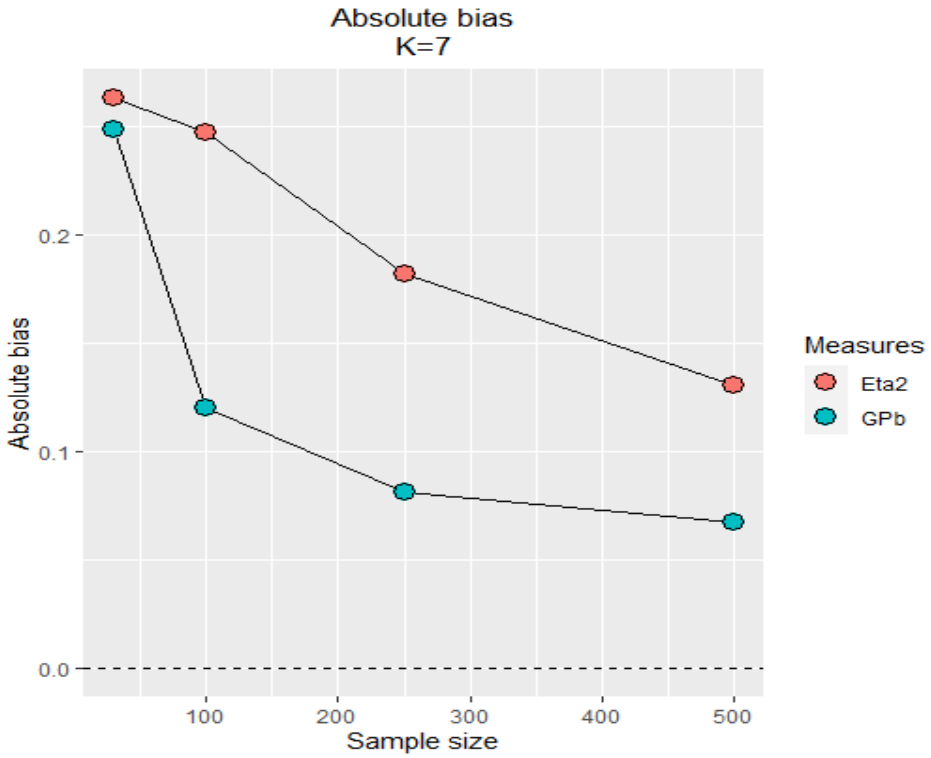


Figure 16. Absolute bias of ρ_{PB} versus Absolute bias of η_2^* for seven non-id dichotomous variables

4.2 Real Application

This section provides data analysis to assess the goodness-of-fit of the proposed measures with respect to Qatar Education data to see how the new measures work in practice. Comparison using real data is not easy because the true measure of association is unknown. Nevertheless, we apply the proposed association measures to a real data set: the teachers' dataset (Social and Economic Survey Research Institute 2021)**.** It consists of 424 teachers where the number of observations became 408 teachers after removing missing values and outliers. The measures of association evaluate the hours per week do teachers spend on activities related to their work according to the following three variables:

1. **FORCE:** Teach a subject out of your specialization.

2. **NEW SUBJECT:** Teaching a subject that you have never taught before.
3. **CHOICE**: Teaching your first choice when you joined the education sector.

4.2.1 Study characteristics

Table 25: Qatar Education survey 2018 (Teachers dataset) description

Number of Cases	424		
Data Collection Period	November 2018 – April 2019		
Survey Organization	Social and Economic Survey Research Institute (SESRI), Qatar University		
Interview Method	Face-to-face Computer-assisted personal interviewing (CAPI)		
Data Type	Sample survey data		
Unit of Analysis	408 School Teachers	259	Independent schools teachers
		149	Private schools teachers

The Social and Economic Survey Research in Qatar university prepared and published the Qatar Education Study (QES), which is a series of surveys. Each survey studied various topics on how students, parents, teachers, and administrators view the current education system. The 2018 Qatar Education Study (QES) dataset is the third and last of a series of three datasets(QES 2012, QES 2015, and QES 2018). The four surveys that comprise the 2018 Qatar Education Study include over 3380 participants from 34 preparatory and secondary schools. In addition, the survey of teachers includes 424 respondents who hold positions such as school principal, academic advisor, and subject coordinator were examining the attitudes of all education system members who will support the development of educational plans in Qatar.

4.2.2 Data analysis

Teachers' datasets are used to establish three measures of association where separate datasets into two samples. Alongside independent schools, private sector schools also play an increasingly important role in providing education services in Qatar. Thus, the sample was 259 teachers from independent schools and 149 teachers from private schools. The measures of association examined the teachers' performance under three dichotomous variables mentioned at the beginning regarding the hours per week teachers spend on activities related to their work.

4.2.3 Properties

There are four main properties as follows:

1- The continuous variable should be normally distributed:

Two continuous variables were observed for independent schools and private schools. However, normality tests can be conducted using Shapiro–Wilk test. For both variables, the p-values less than 0.05. Thus, it rejects the hypothesis of normality.

Many methods have been developed over the years to relax this assumption, in cluding generalized linear models, quantile regression, survival models, and so on. One technique that is still used in this context is to "beat the data" into looking normal by applying some kind of normalizing transformation. This could be as simple as a log transformation or as complicated as a Yeo-Johnson transformation. The variables still reject the hypothesis of normality using the simple way, which is a log transformation. Thus, using the Yeo-Johnson transformation was the best solution.

The Yeo-Johnson transformation

The Yeo Johnson transformation (Yeo and Jhonson,2000) attempts to find the value of lambda (in the following equation) that minimizes the Kullback-Leibler distance between the normal distribution and the transformed distribution

$$g(x,\lambda) = \mathbf{1}_{(\lambda\neq 0, x\geq 0)} \frac{(x+1)^{\lambda}-1}{\lambda}$$

$$+\mathbf{1}_{(\lambda=0, x\geq 0)} \log(x+1)$$

$$+\mathbf{1}_{(\lambda\neq 2, x<0)} \frac{(x+1)^{2-\lambda}-1}{\lambda-2}$$

$$+\mathbf{1}_{(\lambda=2, x<0)} -\log(1-x)$$

This method has the advantage of working without having to worry about the domain of x. This λ parameter, like the Box-Cox, can be evaluated using maximum likelihood. However, the two continuous variables were mentioned above transformed to normally distributed variables using Johnson transformation as shown below:

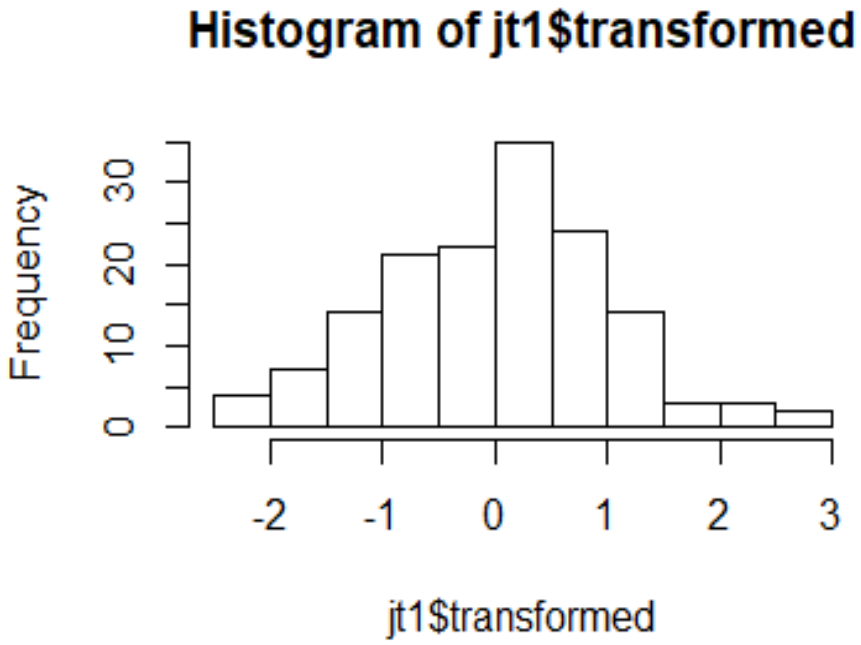


Figure 17: Histogram of the number of hours for private schools

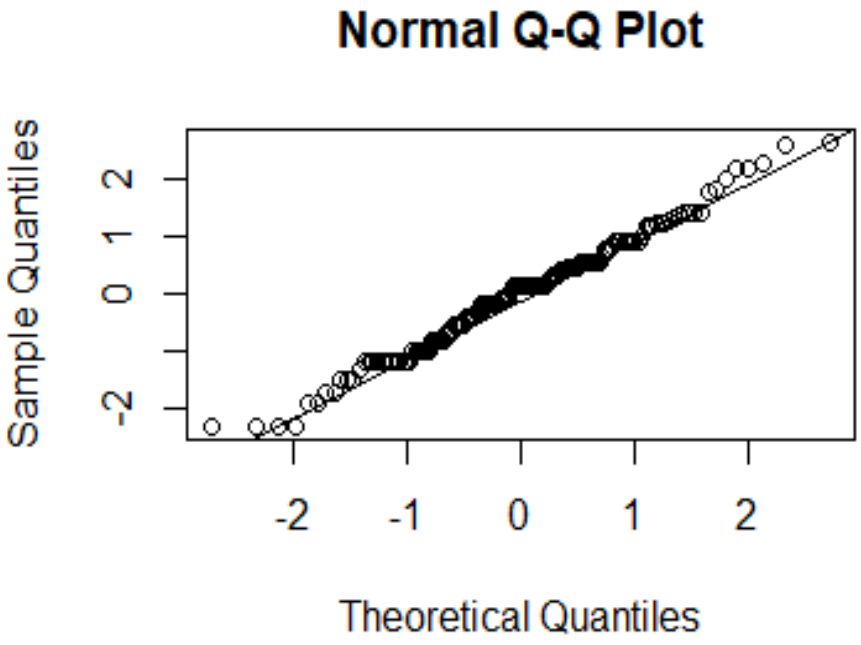


Figure 18: Normal Q–Q Plot of the Number of hours for private schools

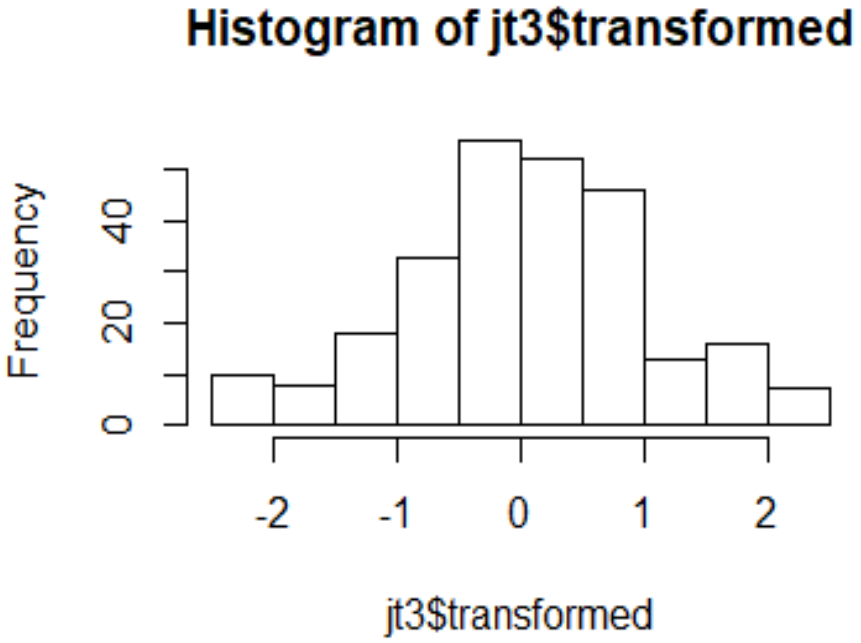


Figure 19: Histogram of the number of hours for independent schools

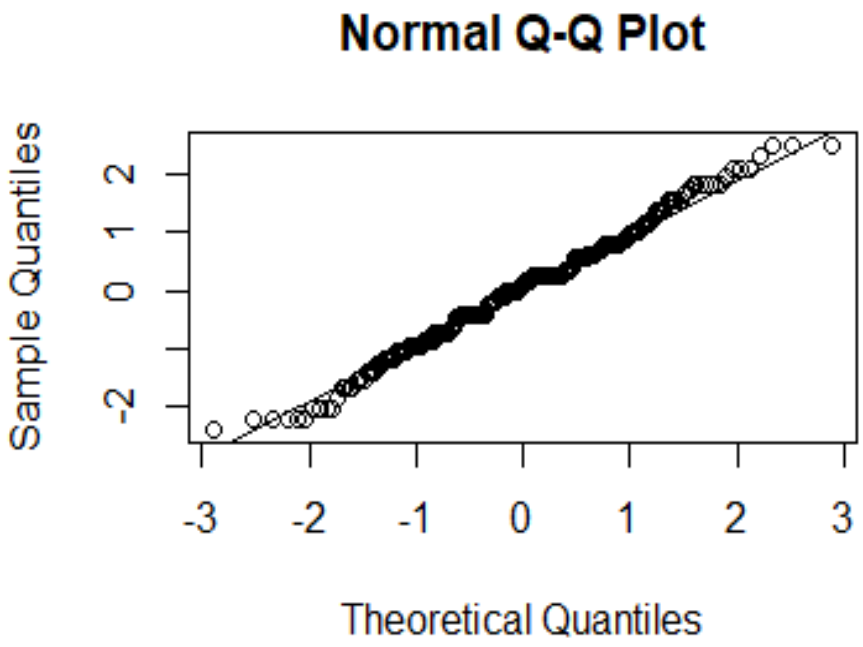


Figure 20: Normal Q–Q Plot of the Number of hours for independent schools

The Q-Q plot shows that the points on the plot for the transformed data closely follow the fitted normal distribution line for both continuous variables. Sometimes the histogram and other visualization techniques are not enough to provide a clear, conclusive answer; statistical inference (Hypothesis Testing) can provide a more objective answer as to whether our variables deviate significantly from a normal. The p-values were more effective to evidence that the normal distribution is a good fit. For independent schools teachers, the p-value was 0.1396927 and 0.2083471 for private schools teachers, so p-values greater than alpha indicate that the data follow the normal distribution.

2- Check if dichotomous variables are independent and identical Bernoulli random variables or non-identical using probability for each variable. For independent schools teachers: FORCE variable has 0.10 for 0's and 0.90 for 1's, NEW SUBJECT variable has 0.05 for 0's and 0.95 for 1's, and CHOICE variable has 0.15 for 0's and 0.85 for 1's. For private schools teachers: FORCE variable has 0.25 for 0's and 0.75 for 1's, NEW SUBJECT variable has 0.05 for 0's and 0.95 for 1's and CHOICE variable has 0.11 for 0's and 0.89 for 1's .Therefore, the second measure of association of non-id and independent dichotomous variables is more relevant.

3- The data should not contain outlier points

Several techniques can be used to get the outliers in data. Mahalanobis distance is one of the popular techniques using for outliers detection. Thus, it was used to detected outliers and removed them from the data.

4- The 1's categories on dichotomous variables correspond to the higher mean on the continuous variable, and the 0's categories correspond to the lowest mean on the continuous variable for two analyses.

4.2.4 Measures of association results

Tables 26 and 27 show the results of the association between the dichotomous variables and the number of hours based on the proposed measure, regression analysis, and $\boldsymbol{\eta_2^*}$. The results show that both $\boldsymbol{\rho_{PB}}$ and regression analysis indicates a weak positive correlation between the dichotomous and the number of hours variable, while $\boldsymbol{\eta_2^*}$ I\indicated a stronger correlation than the previous measures. However, based on

the simulation study that was conducted in chapter 4, $\boldsymbol{\rho_{PB}}$ always perform better than $\boldsymbol{\eta_2^*}$ when the sample size is larger than 30, which is the case in our real data, therefore the $\boldsymbol{\rho_{PB}}$ can be more trusted than $\boldsymbol{\eta_2^*}$ and to support that, regression analysis was conducted to give closer results to $\boldsymbol{\rho_{PB}}$ measure.

Table 26: The different association coefficient between the number of hours for private schools and the dichotomous variables

Methods	**Coefficients**
$\boldsymbol{\rho_{PB}}$	**0.02**
$\boldsymbol{R}$	**0.15**
$\boldsymbol{\eta_2^*}$	**0.33**

Table 27: The different association coefficient between the number of hours for independent schools and the dichotomous variables

Methods	**Coefficients**
$\boldsymbol{\rho_{PB}}$	**0.04**
$\boldsymbol{R}$	**0.10**
$\boldsymbol{\eta_2^*}$	**0.38**

Tables 26 and 27 show the measured association between the number of hours per week do teachers spend on activities related to their work, and their considered dichotomous variables (**FORCE** variable interpret that they teach a subject out of their specialization, **NEW SUBJECT** variable show if teaching was a subject that they have never taught before and **CHOICE** variable show if teaching was their first choice when

they joined education sector). Therefore, and based on $\boldsymbol{\rho_{PB}}$ measured association, a small positive value means that the relationship between the considered variables is positive and weak, which was the situation in both private and independent schools.

CHAPTER 5: CONCLUSION AND SUGGESTIONS FOR FURTHER STUDY

This study focuses on categorical data analysis by investigating two forms of measures the association between multi dichotomous variables and a continuous variable. Instead of using Bernoulli distribution in point biserial correlation, both binomial and Poisson binomial are used by generalizing the point biserial correlation coefficient. Monte Carlo power studies were performed for 10000 replications with various values of sample sizes n, several numbers of dichotomous variables and different probabilities. As sample size increases, MSE and Bias decrease, which shows the accuracy of the proposed methods. The results of Monte Carlo power studies revealed that the proposed methods outperform the $\boldsymbol{\eta}_{\boldsymbol{2}}^{*}$ method in most cases, especially when both the sample size and the number of dichotomous variables increase. Finally, applications on real data sets were applied to demonstrate the measures of association for the proposed methods and $\boldsymbol{\eta}_{\boldsymbol{2}}^{*}$.

Future study

This study focuses on multi-dichotomous variables with only one continuous variable. Thus, future studies may consider both multi dichotomous and multi continuous variables by generalizing point biserial.

Furthermore, proposed measures do not take into account the order of the categories on the dichotomous variables (e.g., in the case of two binaries, the measure treat 0 on X_1 corresponding 1 on X_2 and 1 on X_1 corresponding 0 on X_2 are similarly).

Also, for the second measure where the dichotomous variables are non-identical. Future researchers could use the same technique and build their study by focusing on more different probabilities.

In addition to that, an intense simulation study might be conducted to investigate the impact of the number of dichotomous variables on the performance of the proposed measures.

Researchers in applied fields might use the proposed measures to assess the association between dichotomous and normal distribution variables.

REFERENCES

Barbiero, A., & Hitaj, A. (2020). Goodman and Kruskal's Gamma Coefficient for Ordinalized Bivariate Normal Distributions. *Psychometrika*, 1-21.

Berry, K. J., Johnston, J. E., & Mielke Jr, P. W. (2018). *The measurement of association: a permutation statistical approach*: Springer.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*(2), 213-225.

Boslaugh, S. (2012). *Statistics in a nutshell: A desktop quick reference*: " O'Reilly Media, Inc.".

Chen, S. X., & Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 875-892.

Goodman, L. A., & Kruskal, W. H. (1979). Measures of association for cross classifications. *Measures of association for cross classifications*, 2-34.

Gupta, S. D. (1960). Point biserial correlation coefficient and its generalization. *Psychometrika, 25*(4), 393-408.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis, 59*, 41-51.

Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics, 7*(1), 29-43.

Islam, T. U., & Rizwan, M. (2020). Comparison of correlation measures for nominal data. *Communications in Statistics-Simulation and Computation*, 1-20.

Kendall, M. G. (1948). The advanced theory of statistics. Vols. 1. *The advanced theory of statistics. Vols. 1., 1*(Ed. 4).

Kendall, M. G. (1948). Rank correlation methods.

Kendall, M. G., Kendall, S. F., & Smith, B. B. (1939). The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 251-273.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275-287.

Khamis, H. (2008). Measures of association: how to choose? *Journal of Diagnostic Medical Sonography, 24*(3), 155-162.

LeBlanc, V., & Cox, M. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *Tutorials in Quantitative Methods for Psychology, 13*(1), 46-56.

Lev, J. (1949). The point biserial coefficient of correlation. *Annals of Mathematical Statistics, 20*(1), 125-126.

Louangrath, P. (2014). Correlation coefficient according to data classification. *Available at SSRN 2417910*.

Neammanee, K. (2005). A refinement of Normal approximation to Poisson Binomial. *International Journal of Mathematics and Mathematical Sciences, 2005*(5), 717-728.

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics, 32*(2), 448-465.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*(3), 337-347.

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 195*(262-273), 1-47.

Pearson, K. (1909). On a new method of determining correlation between a measured character a, and a character b, of which only the percentage of cases wherein b exceeds (or falls short of) a given intensity is recorded for each grade of a. *Biometrika, 7*(1/2), 96-105.

Perinetti, G. (2019). StaTips part VI: Bivariate correlation. *South European journal of orthodontics and dentofacial research, 6*(1), 2-5.

Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/− 1, or do they? *Journal of targeting, measurement and analysis for marketing, 17*(2), 139-142.

Samuels, S. M. (1965). On the number of successes in independent trials. *Annals of Mathematical Statistics, 36*(4), 1272-1278.

Spearman, C. (1906). Footrule for measuring correlation. *British Journal of Psychology, 2*(1), 89.

Taha, A., & Hadi, A. S. (2016). Pair-wise association measures for categorical and mixed data. *Information Sciences, 346*, 73-89.

Tare, R. (1949). *The biserial and point biserial correlation coefficient. Univ. North Carolina, Inst*. Retrieved from

Tate, R. F. (1950). *The biserial and point correlation coefficients*. Retrieved from

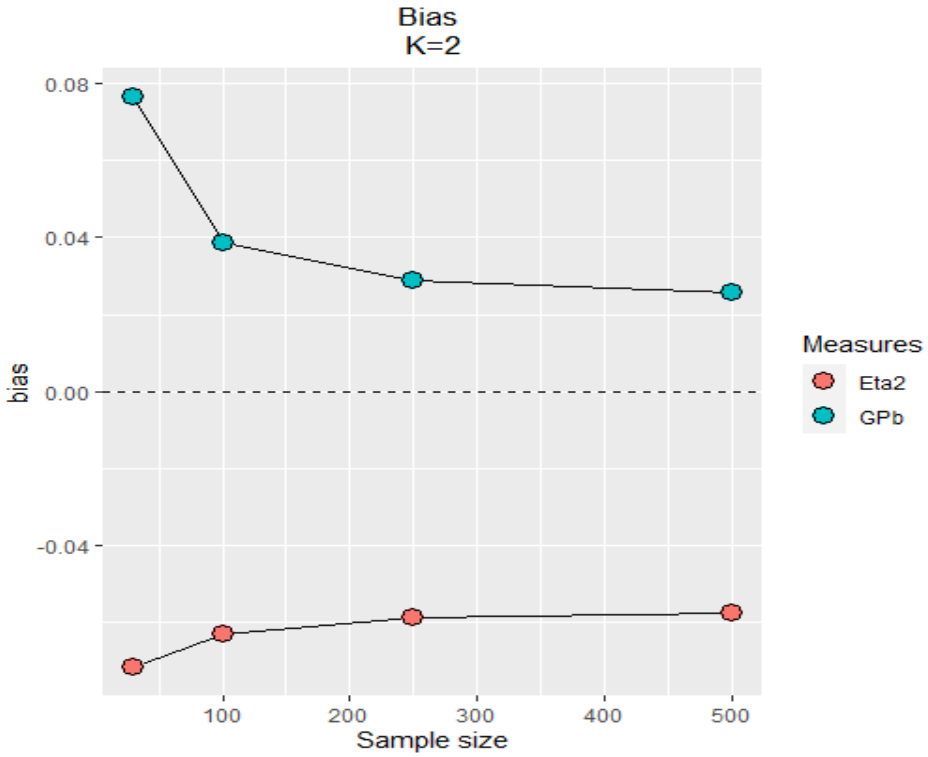
Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics, 25*(3), 603-607.

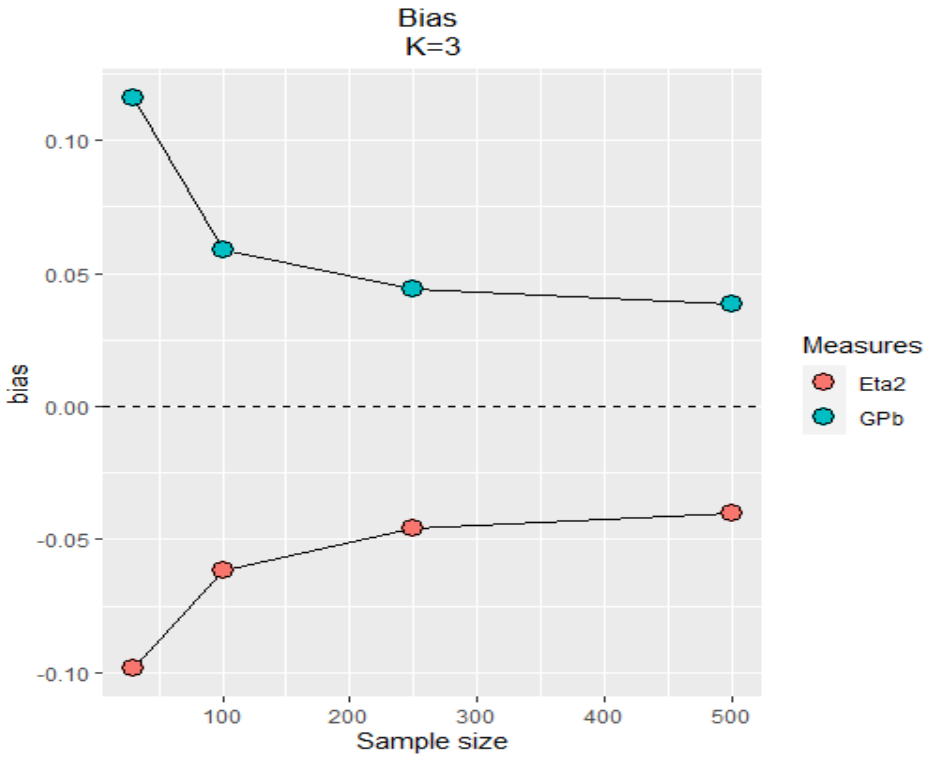
Ulrich, R., & Wirtz, M. (2004). On the correlation of a naturally and an artificially dichotomized variable. *British Journal of Mathematical and Statistical Psychology, 57*(2), 235-251.

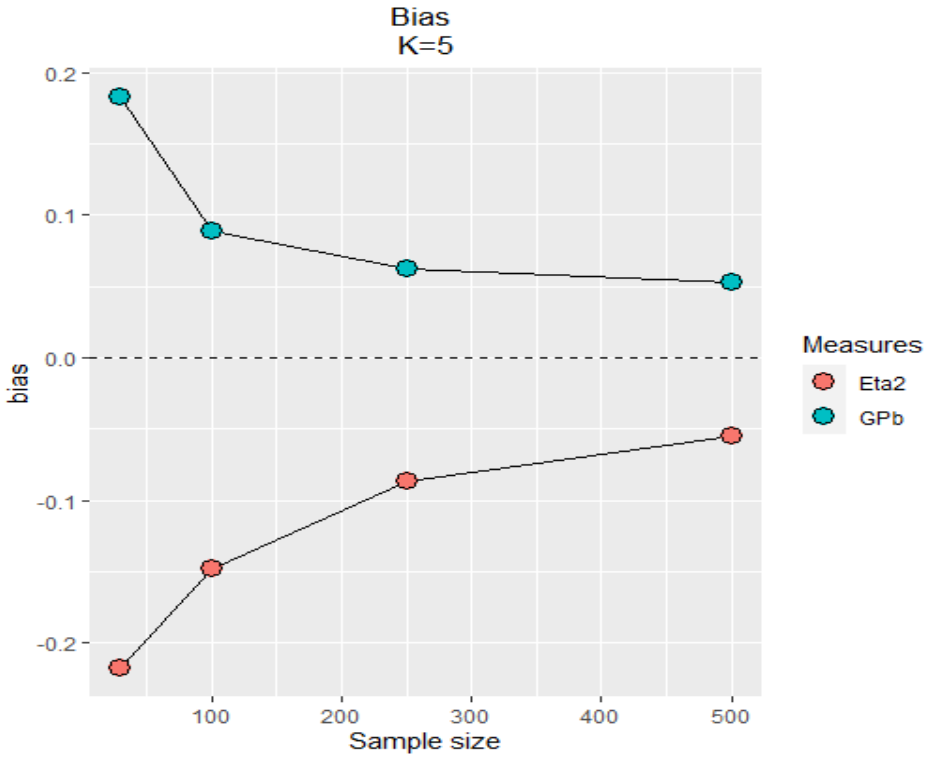
Wherry, R. J., & Taylor, E. K. (1946). The relation of multiserial eta to other measures of correlation. *Psychometrika, 11*(3), 155-161.

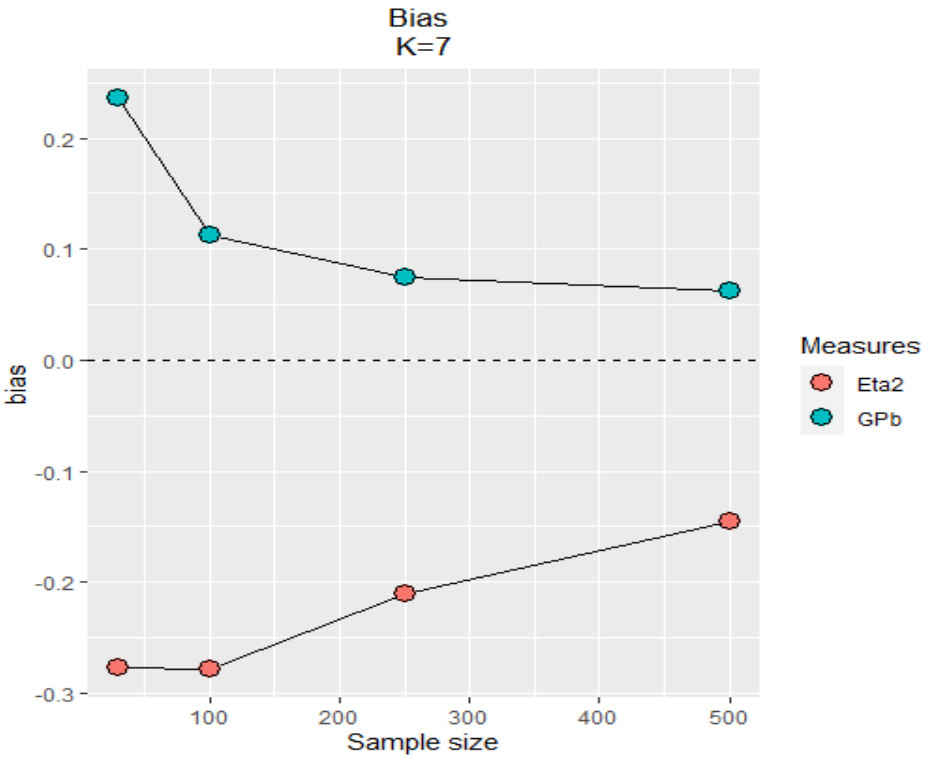
Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*(6), 579-652.

APPENDIX A: BIAS OF ρ_B VERSUS BIAS OF η_2^* FIGURES FOR TWO, THREE, FIVE AND SEVEN IID DICHOTOMOUS VARIABLES

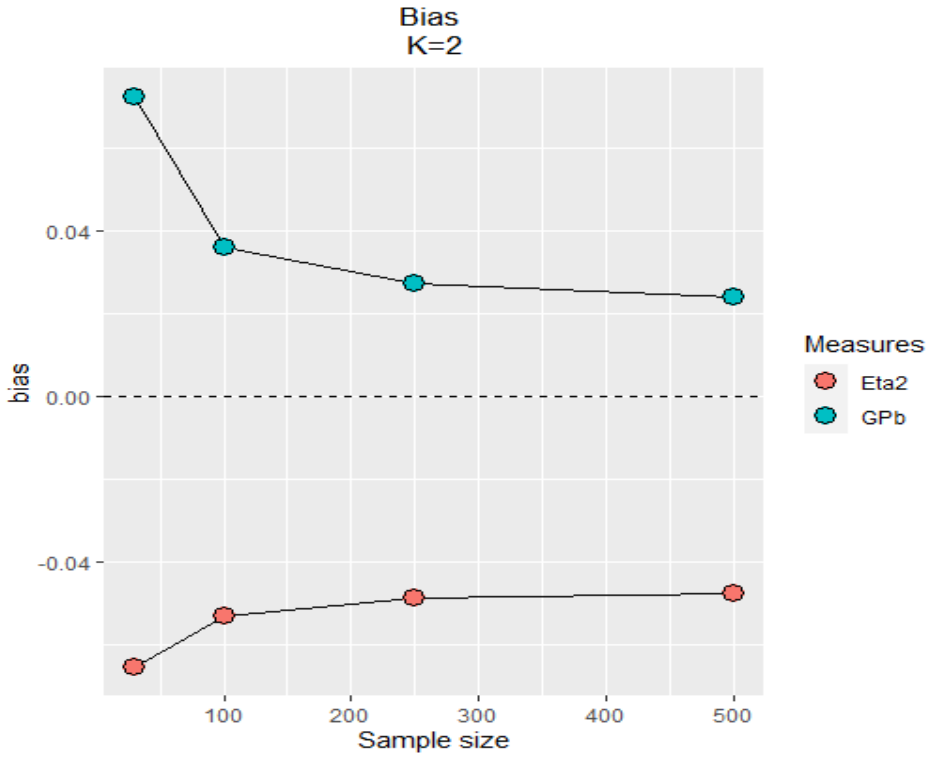


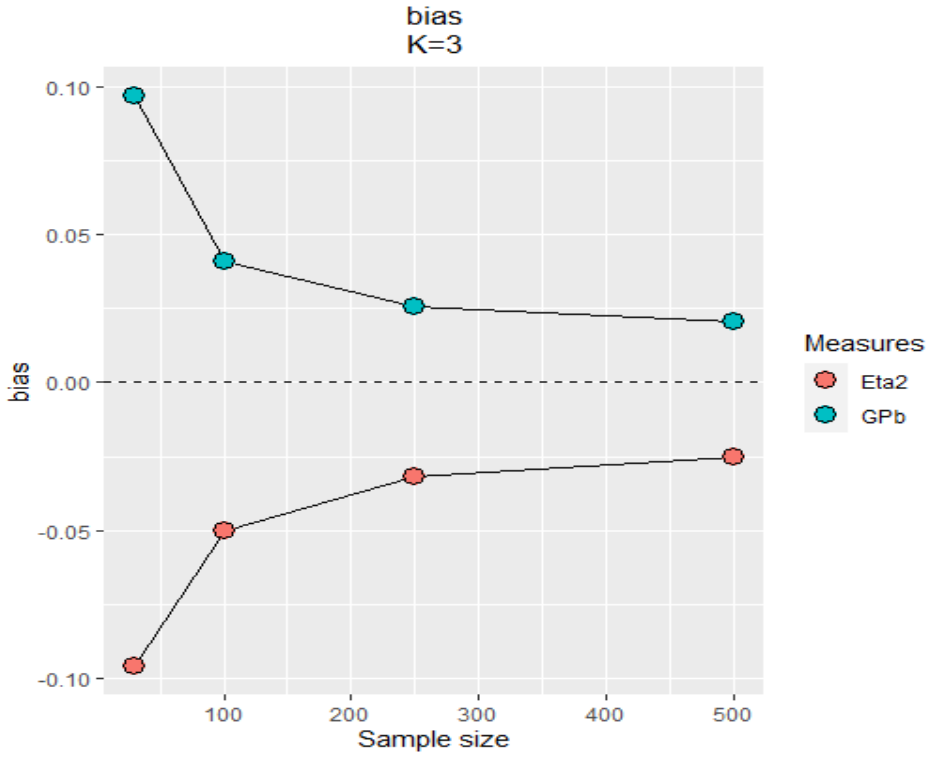

Bias
K=3
0.10
0.05
0.00
-0.05
-0.10
bias
100
200
300
400
500
Sample size
Measures
Eta2
GPb

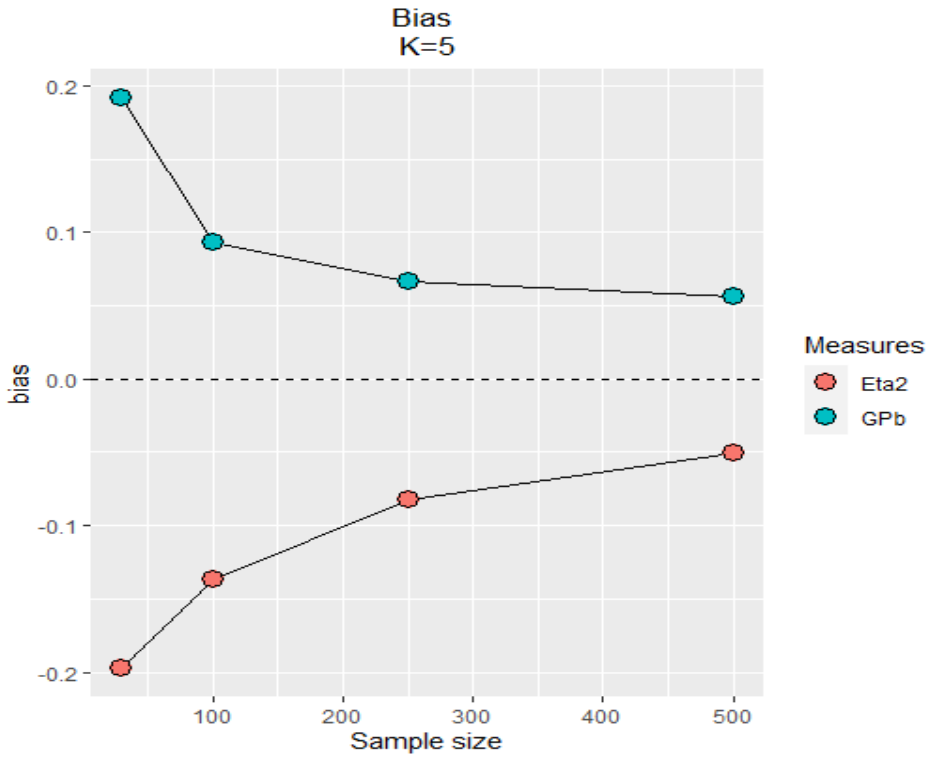

Bias
K=5
0.2
0.1
0.0
-0.1
-0.2
bias
100
200
300
400
500
Sample size
Measures
Eta2
GPb

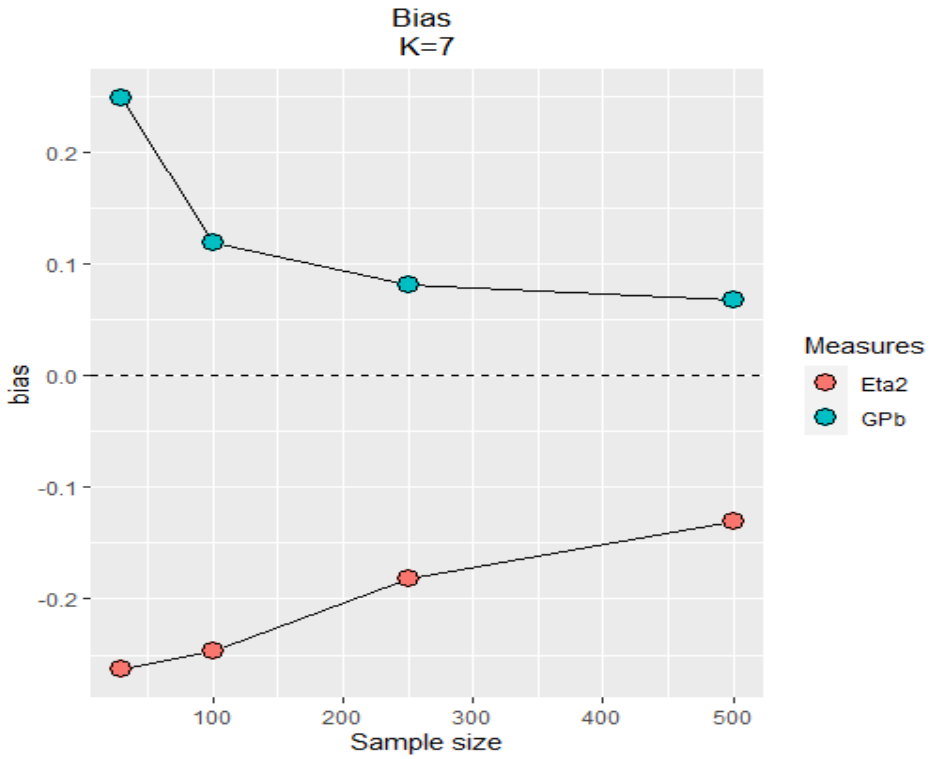

Bias
K=7
0.2
0.1
0.0
-0.1
-0.2
-0.3
bias
100
200
300
400
500
Sample size
Measures
Eta2
GPb

APPENDIX B: BIAS OF ρ_{pB} VERSUS BIAS OF η_2^* FIGURES FOR TWO, THREE, FIVE AND SEVEN NON-ID DICHOTOMOUS VARIABLES

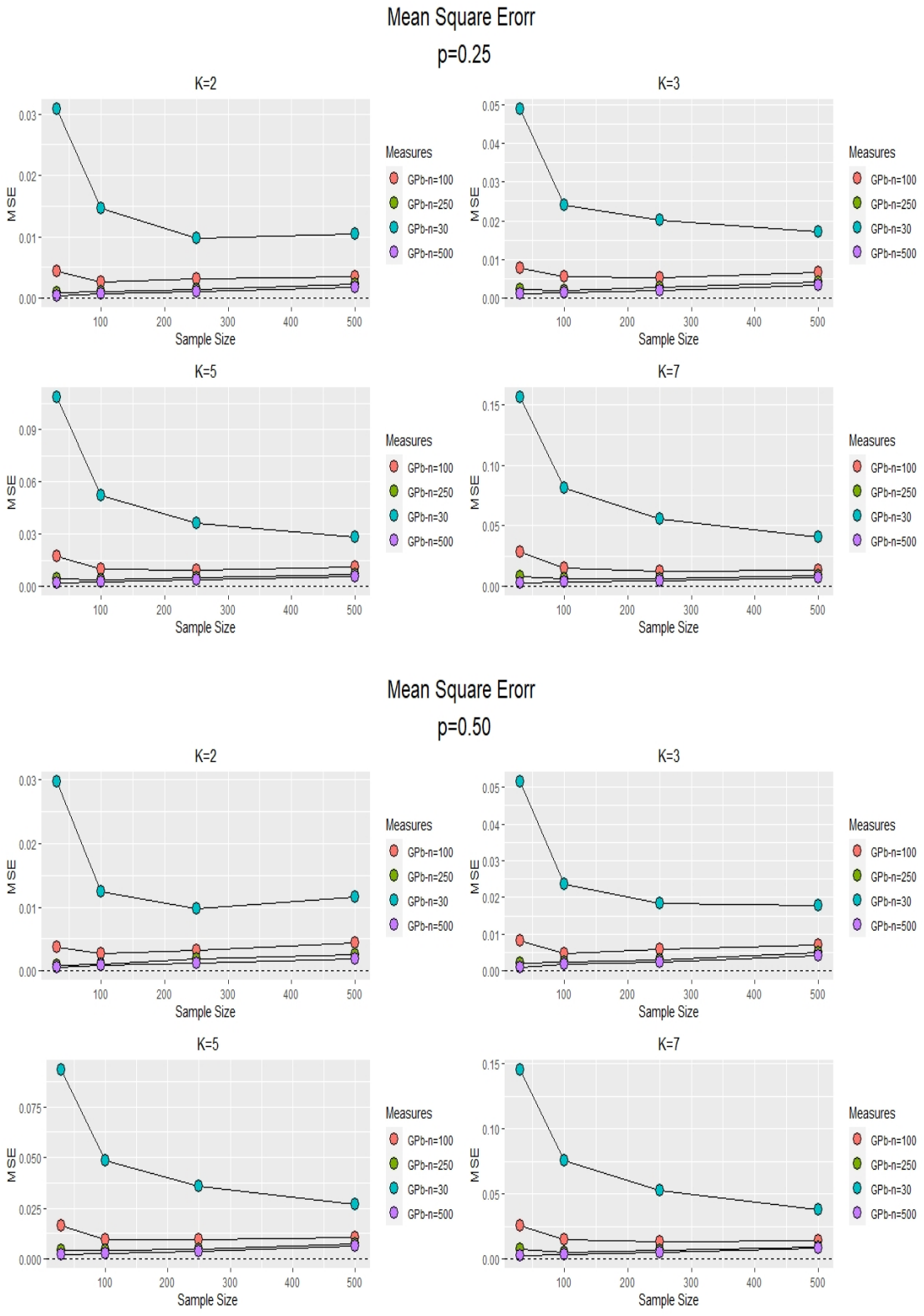


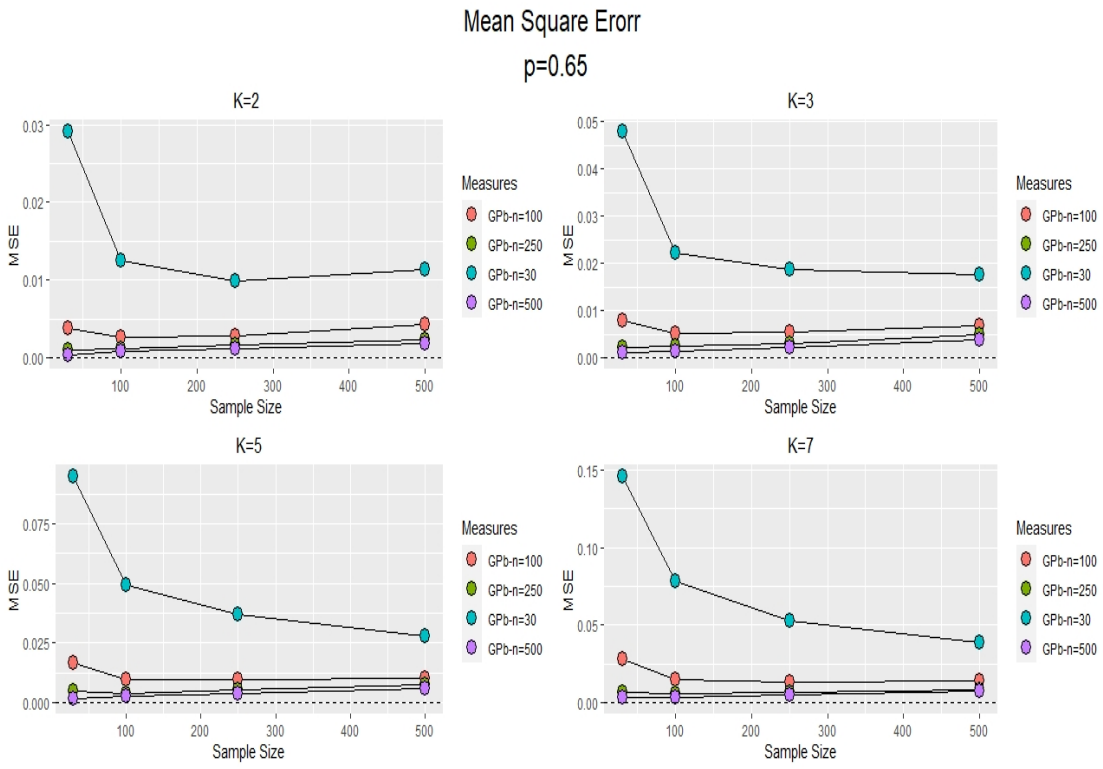


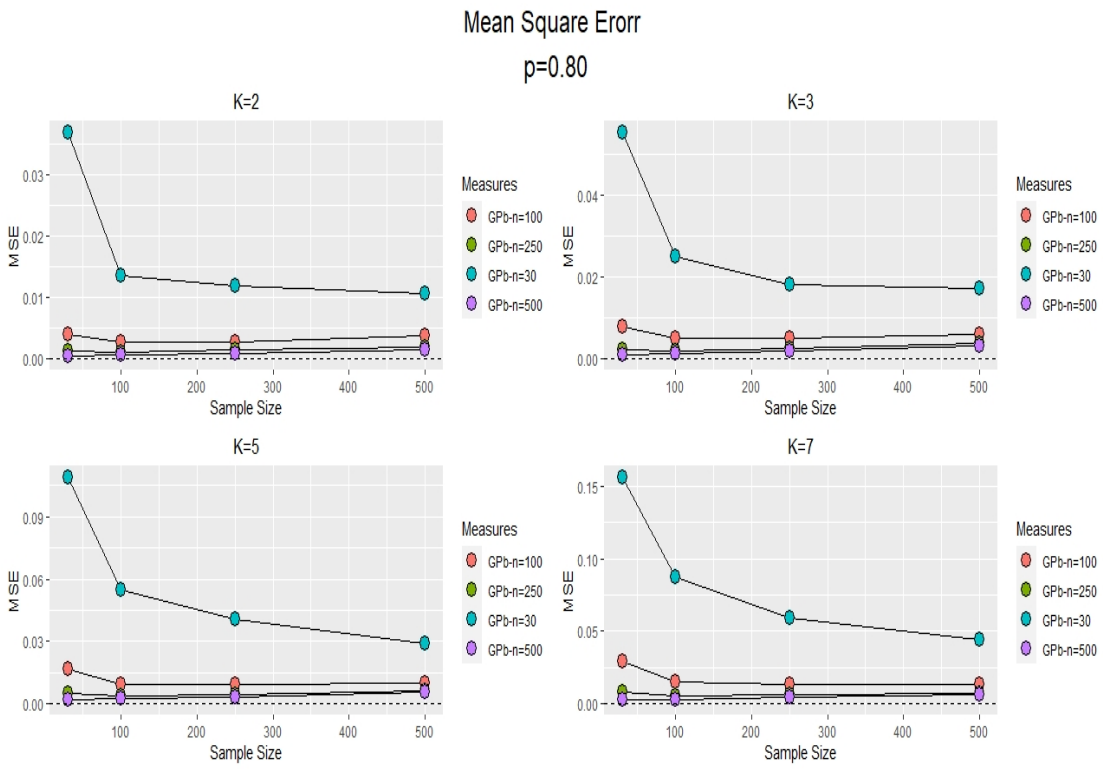

Bias
K=5
0.2
0.1
0.0
-0.1
-0.2
bias
100
200
300
400
500
Sample size
Measures
Eta2
GPb


Bias
K=7
0.2
0.1
0.0
-0.1
-0.2
bias
100
200
300
400
500
Sample size
Measures
Eta2
GPb

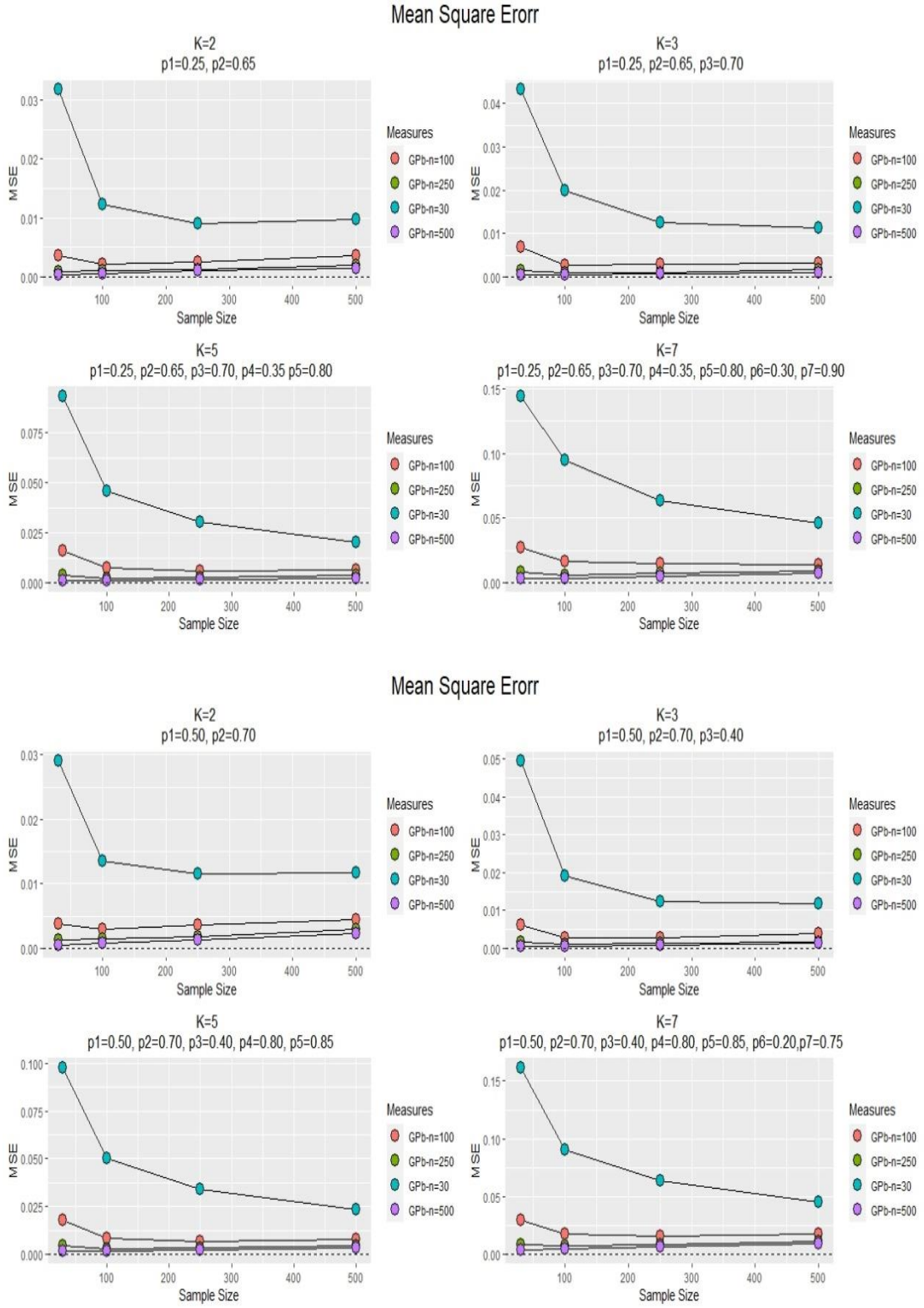
APPENDIX D: MSE OF ρ_B FOR TWO, THREE, FIVE AND SEVEN IID DICHOTOMOUS VARIABLES

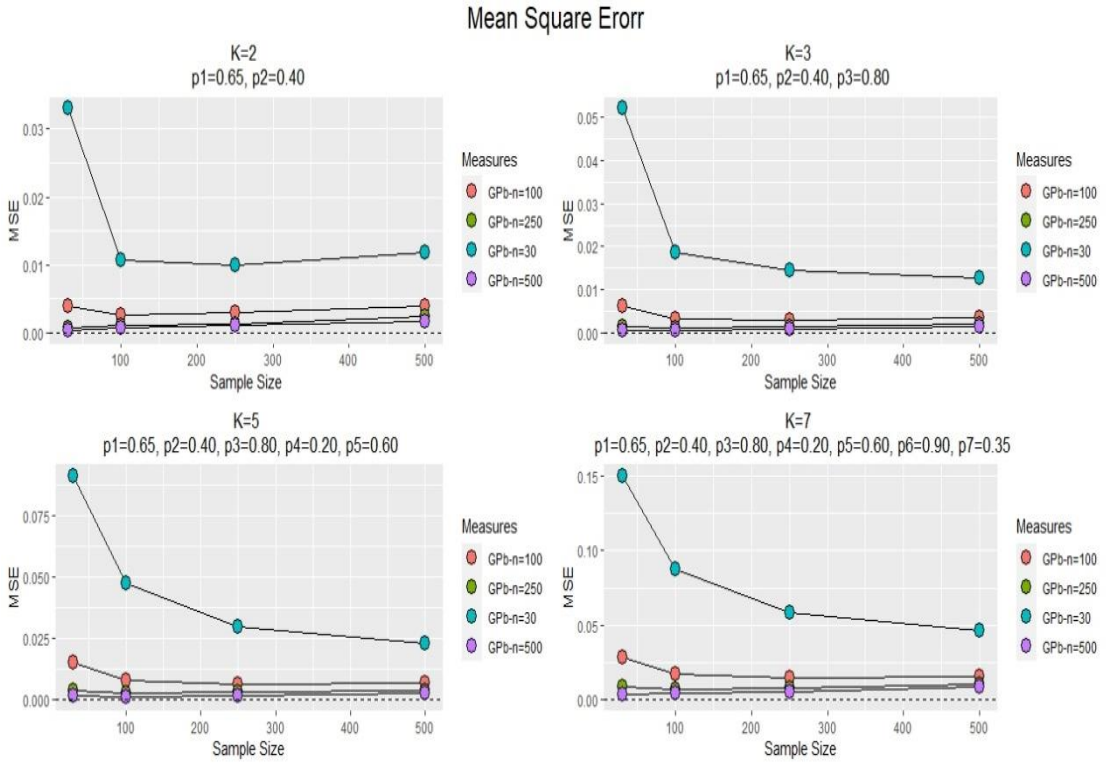


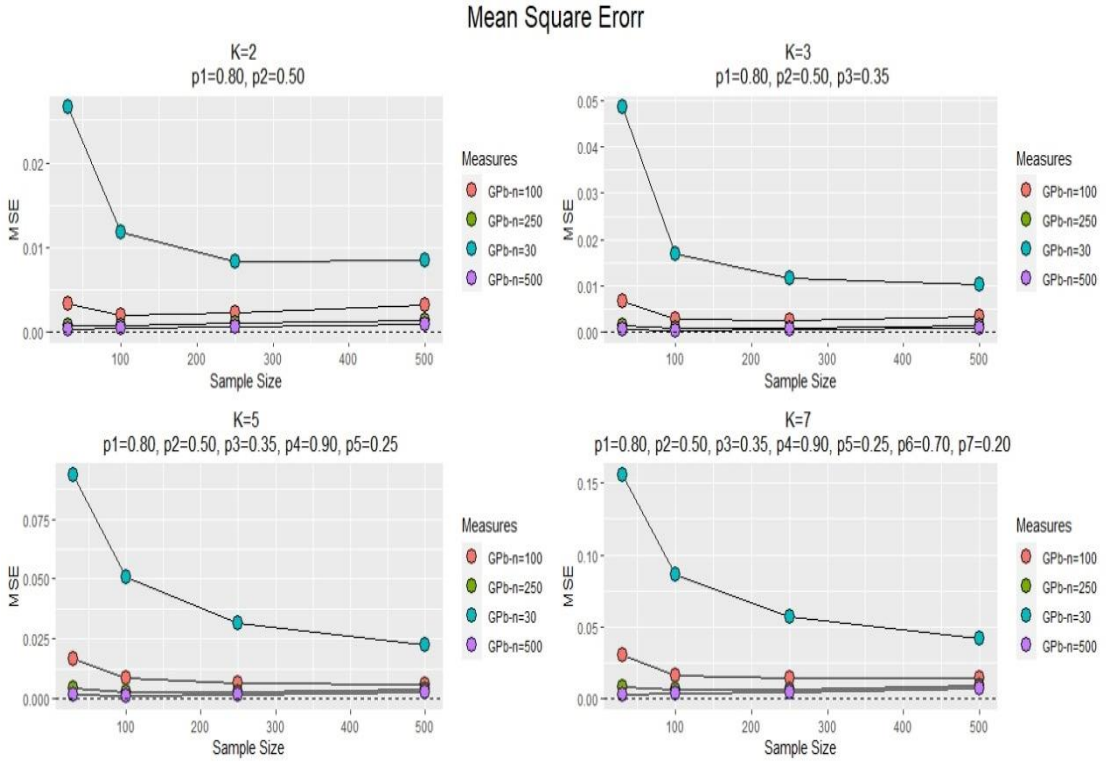

Mean Square Erorr
p=0.65
K=2
K=3
K=5
K=7
MSE
Sample Size
Measures
GPb-n=100
GPb-n=250
GPb-n=30
GPb-n=500


Mean Square Erorr
p=0.80
K=2
K=3
K=5
K=7
MSE
Sample Size
Measures
GPb-n=100
GPb-n=250
GPb-n=30
GPb-n=500

APPENDIX E: MSE OF ρ_{pB} FOR TWO, THREE, FIVE AND SEVEN IID DICHOTOMOUS VARIABLES




Mean Square Erorr
K=2
p1=0.65, p2=0.40
K=3
p1=0.65, p2=0.40, p3=0.80
K=5
p1=0.65, p2=0.40, p3=0.80, p4=0.20, p5=0.60
K=7
p1=0.65, p2=0.40, p3=0.80, p4=0.20, p5=0.60, p6=0.90, p7=0.35
MSE
Sample Size
Measures
GPb-n=100
GPb-n=250
GPb-n=30
GPb-n=500


Mean Square Erorr
K=2
p1=0.80, p2=0.50
K=3
p1=0.80, p2=0.50, p3=0.35
K=5
p1=0.80, p2=0.50, p3=0.35, p4=0.90, p5=0.25
K=7
p1=0.80, p2=0.50, p3=0.35, p4=0.90, p5=0.25, p6=0.70, p7=0.20
MSE
Sample Size
Measures
GPb-n=100
GPb-n=250
GPb-n=30
GPb-n=500