

December 2017

# Study Evaluating the Alterations Caused in an Exploratory Factor Analysis when Multivariate Normal Data is Dichotomized


Rosilei S. Novak

*Federal University of Paraná, Curitiba, Brazil, rosileisouzanovak@gmail.com*

Jair M. Marques

*Federal University of Paraná, Curitiba, Brazil, jair.m.marques@gmail.com*

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Novak, R. S., & Marques, J. M. (2017). Study Evaluating the Alterations Caused in an Exploratory Factor Analysis when Multivariate Normal Data is Dichotomized. *Journal of Modern Applied Statistical Methods*, 16(2), 604-617. doi: 10.22237/jmasm/1509496380

This Emerging Scholar is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in *Journal of Modern Applied Statistical Methods* by an authorized editor of DigitalCommons@WayneState.

# Study Evaluating the Alterations Caused in an Exploratory Factor Analysis when Multivariate Normal Data is Dichotomized

**Rosilei S. Novak**  
Federal University of Paraná  
Curitiba, Brazil

**Jair M. Marques**  
Federal University of Paraná  
Curitiba, Brazil

---

The relationships resulting from the dichotomization of multivariate normal data is a question that causes concern when using exploratory factor analysis. The relationships in an exploratory factor analysis are examined when multivariate normal data, generated by Monte Carlo methods, is dichotomized.

*Keywords:* Exploratory factor analysis, normal multivariate data, dichotomized data, Monte Carlo simulation

---

## Introduction

The dichotomization of multivariate normal data is widely used when working with exploratory factor analysis (EFA). The alteration on the variables facilitates the representation, reduces field expenses, and assists on the interpretations of the results. However, this process can lead to information loss from the real data.

The phi correlation coefficient was used for dichotomized data, since the objective was to analyze the impact of the substitution of the tetrachoric correlation coefficient by the phi correlation coefficient. In situations where the tetrachoric correlation matrices are singular, they are not appropriate for factor analysis (Embreson & Reise, 2013, p. 37).

Distortion is frequently verified on original data when data are dichotomized. MacCallum, Zhang, Preacher, and Rucker (2002) presented a practical analysis of dichotomization, illustrating with numerical examples the consequences caused on the original variables. Kubinger (2003) highlighted a problem in psychological

---

*Rosilei Souza Novak is a PhD in the Department of Numerical Methods in Engineering. Email her at: [rosileisouzanovak@gmail.com](mailto:rosileisouzanovak@gmail.com). Jair Mendes Marques is an Associate Professor of Numerical Methods in Engineering. Email him at: [jair.m.marques@gmail.com](mailto:jair.m.marques@gmail.com).*

studies, where hundreds of tests were developed based on factor analysis using dichotomic variables, leading to compromised results. Fedorov, Mannino, and Zhang (2008) stated dichotomization is a transformation of a continuous result into a binary result. This uncommon approach is prejudicial to hypothesis tests and statistical estimates. Their work was based on Fisher's approach, showing that this transformations leads to a great loss of information when data present normal distribution. In terms of information, this loss affects at least  $1 - 2 / \pi$  (or 36%) of the original data. Pearson and Mundform (2010) considered the distortion of original data when it is dichotomized, because the effects caused by this transformation is unknown.

The magnitude of the real loss caused by dichotomization on real data is still unknown in empirical studies. The aim in this study is to shed light on this question. To do so, MATLAB functions were developed to generate random multivariate normal samples using the Monte Carlo simulation method. Those samples were then dichotomized and factor analysis was performed on each normal sample and its corresponding dichotomized sample. Finally, significance tests were performed to compare means and variances between multivariate normal and dichotomized samples.

## Methods

This research was conducted with the aid of MATLAB R12 software. A total of 12,600 random multivariate normal samples were generated by the Monte Carlo simulation method. Afterwards, those samples were dichotomized. The 12,600 samples were generated considering the variation of number of variables (14 cases), as shown in Table 1, and to each of these cases the sample size have varied in 9 different situations (2, 3, 4, 5, 10, 20, 30, 40, and 50 times the number of variables). Once each sample was generated 100 times, the final result was  $14 \times 9 \times 100 = 12,600$  samples.

Shown in Table 1 are the simulations, where the vectors represent the number of variables per factor. For instance, the vector [3 2] represents 2 factors with 5 variables (3 variables on the first factor and 2 variables on the second factor). The criterion used to select the number of variables was a minimum of 5 and a maximum of 50, and the criterion used to select the number of factors was a minimum of 2 and a maximum of 10.

## EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

**Table 1.** Total classification of number of variables per factor

Simulation	Variables	Variables per factor
1	5	[3 2]
2	6	[3 3]
3	7	[4 3]
4	8	[5 3]
5	9	[4 3 2]
6	10	[5 3 2]
7	15	[7 4 2 2]
8	20	[10 5 3 2]
9	25	[10 5 5 3 2]
10	30	[10 8 5 3 2 2]
11	35	[15 5 5 3 3 2 2]
12	40	[15 5 5 5 3 3 2 2]
13	45	[15 5 5 5 4 4 3 2 2]
14	50	[14 6 5 5 4 4 3 3 3 3]

In order to perform this study, two functions were created in MATLAB R12: *Matrizc1* and *Simula1*. The first, *Matrizc1*, generates an iteration according to sample size, number of variables involved, and number of factors, based on 100 random multivariate normal samples. Those samples were then dichotomized. The dichotomization of the multivariate normal samples was performed considering three conditions:  $P(z \leq z_c) = 0.25$  (1<sup>st</sup> dichotomization point),  $P(z \leq z_c) = 0.50$  (2<sup>nd</sup> dichotomization point) and  $P(z \leq z_c) = 0.75$  (3<sup>rd</sup> dichotomization point).

Only multivariate normal samples with the following requirements were considered: Phi correlation matrix, with  $MSA > 0.5$  and communalities  $\geq 0.7$ . The samples which did not matched the established requirements were discarded and replaced, until the total amount of 100 samples was reached.

The second function, *Simula1*, performed factor analysis to each of the 100 samples individually (multivariate normal samples and its corresponding dichotomized samples), obtaining (to each of the 100 samples generated) the MSA (mean sample adequacy) for the percentage of variance explained by the first factor, by the total variance and by the communalities (evaluation of the common proportion of variance of each variable shared with common factors).

The factor analysis was performed using the main components method. The number of factors was determined using the Kaiser criterion. The rotation method used was Varimax orthogonal. Factorial loads were not considered, once many oscillations occurred among samples, disallowing comparisons.

After obtaining the factor analysis results, the same function, Simula1, performed significance tests to compare means and variances of the samples. The following statistical significance tests were performed: Student's *t*-test (comparing means of multivariate normal and dichotomized data, variance explained by the first factor and total variance explained),  $T^2$  Hotelling test (comparing mean vectors of multivariate normal data and dichotomized data from the communalities), Snedecor *f*-test (comparing the variances of MSA multivariate normal data and dichotomized data, variance explained by the first factor and total variance explained), with the objective of determining the adequate Student's *t*-test. Finally, the multivariate chi-square test (comparing the covariance matrix of communalities vectors between multivariate normal and dichotomized data) with the objective of determining the adequate  $T^2$  Hotelling test.

All tests were applied considering a significance level of 0.05. Once the tests are all bilateral, the significant results present  $p < 0.025$ .

In a summarized manner, the methodology was developed in sequence, according to the following stages:

- Generate 100 multivariate normal samples;
- Generate dichotomized samples corresponding to the multivariate normal samples;
- Perform factor analysis on each of the generated samples (normal and dichotomized);
- Calculate, for the 100 multivariate normal and dichotomized samples, the MSA means, variance explained by the first factor, total variance explained and the vectors of the communalities means;
- Perform statistical tests comparing the results obtained through factor analysis of the multivariate normal and dichotomized data.

This article does not present an extensive list of all the simulations conducted. Instead, a representative group was selected, shown in Table 2. The simulations chosen are 1 and 14 with sizes 2, 5, 20, and 50 times the number of variables.

**Table 2.** Reduced classification of number of variables per factor

Simulation	Variables	Variables per factor
1	5	[3 2]
14	50	[14 6 5 5 4 4 3 3 3 3]

## Results

The results of the study are presented in tables, showing the results of the MSA, the proportion of variance explained by the first factor, and the total proportion of variance related to the four types of sample simulations: one small sample (size equal to 2 times the number of variables), two intermediate samples (sizes equal to 5 and 20 times the number of variables), and one large sample (size equal to 50 times the number of variables), in three different dichotomization points. Tables showing the communalities results also are presented, describing only one type of sample simulation, considering sample sizes equal to the cases already seen (2, 5, 20, and 50 times the number of variables) and three different points of dichotomization. The tables referring to the communalities results are extensive, as they show mean vectors. Since no relevant oscillations occurred among simulations, only the tables referring to the first sample simulation are presented, relating the group behavior.

The tables show sample size, means or mean vectors, and  $p$ -values (resulting from significance tests comparing means or MSA mean vectors, proportion of variance explained by the first factor, total proportion of variance and the communalities between multivariate normal data and dichotomized data). The tables do not show variances or covariance matrices and  $p$ -values (resulting from the significance tests comparing variances or covariance matrices), even though the tests performed assisted in the selection of the adequate mean and mean vector tests.

### Results Obtained for the MSA

Tables 3 and 4 relate the sample size with its corresponding means, and the results from the test of mean difference (pMc) of the MSA from the multivariate normal samples and its corresponding dichotomized samples, in three points of dichotomization. It can be observed in Table 3 the results of the MSA for the sample [3 2] (2 factors and 5 variables). The differences identified between the MSA means from normal and dichotomized data were always significant.

The means from dichotomized data were always larger than the means from normal data, except for the sample with size 250 (1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> points of dichotomization), sample classified as large. There is no great influence from the points of dichotomization on the results.

**Table 3.** Means and tests of significance of the MSA for samples with 5 variables and 2 factors – vector [3 2]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization		2 <sup>nd</sup> point of dichotomization		3 <sup>rd</sup> point of dichotomization	
		Mean	pMc	Mean	pMc	Mean	pMc
10	Normal	0.7974	0.00000	0.5587	0.00000	0.5603	0.00000
	Dichotomized	0.8368		0.6192		0.6042	
25	Normal	0.5492	0.00000	0.5600	0.00000	0.5697	0.00015
	Dichotomized	0.5893		0.5968		0.5961	
100	Normal	0.5397	0.00000	0.5480	0.00000	0.5440	0.00000
	Dichotomized	0.5816		0.5920		0.5782	
250	Normal	0.6607	0.00000	0.6545	0.00000	0.6543	0.00000
	Dichotomized	0.6398		0.6431		0.6384	

**Table 4.** Means and tests of significance of the MSA for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization		2 <sup>nd</sup> point of dichotomization		3 <sup>rd</sup> point of dichotomization	
		Mean	pMc	Mean	pMc	Mean	pMc
100	Normal	0.6471	0.00000	0.6466	0.00000	0.6435	0.00000
	Dichotomized	0.6673		0.6876		0.6704	
250	Normal	0.6978	0.00000	0.6986	0.00000	0.6999	0.00000
	Dichotomized	0.8064		0.8175		0.8061	
1000	Normal	0.7128	0.00000	0.7125	0.00000	0.7129	0.00000
	Dichotomized	0.8596		0.8655		0.8587	
2500	Normal	0.7141	0.00000	0.7147	0.00000	0.7145	0.00000
	Dichotomized	0.8688		0.8740		0.8689	

Table 4 shows the MSA results for the sample [14 6 5 5 4 4 3 3 3 3], 10 factors and 50 variables. The differences between the MSA means from normal data and dichotomized data were always significant. The MSA was always higher for dichotomized data. The three points of dichotomization do not have influence on the results.

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

**Table 5.** Means and tests of significance of the MSA for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization			2 <sup>nd</sup> point of dichotomization			3 <sup>rd</sup> point of dichotomization		
		Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N
10	Normal	56.1392	0.00000	0.8541	55.1227	0.02110	0.8856	55.8176	0.00000	0.8440
	Dichotomized	47.9528			48.8185			47.1141		
25	Normal	53.0267	0.00000	0.8676	53.0125	0.00000	0.8747	53.3529	0.00000	0.8582
	Dichotomized	46.0085			46.3744			45.7901		
100	Normal	52.1435	0.00000	0.8454	52.1435	0.00000	0.8454	52.2820	0.00000	0.8351
	Dichotomized	44.0870			44.0870			43.6632		
250	Normal	52.3523	0.00000	0.8265	52.2140	0.00000	0.8438	52.2068	0.00000	0.8285
	Dichotomized	43.2731			44.0626			43.2547		

**Table 6.** Means and tests of significance of the variance explained by the first factor for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization			2 <sup>nd</sup> point of dichotomization			3 <sup>rd</sup> point of dichotomization		
		Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N
100	Normal	23.2264	0.00000	0.7476	23.2459	0.00000	0.7760	23.1972	0.00000	0.7486
	Dichotomized	17.3648			18.0395			17.3660		
250	Normal	23.1705	0.00000	0.7505	23.2400	0.00000	0.7739	23.1826	0.00000	0.7528
	Dichotomized	17.3907			17.9855			17.4532		
1000	Normal	23.1376	0.00000	0.7530	23.1392	0.00000	0.7770	23.1397	0.00000	0.7511
	Dichotomized	17.4239			17.9811			17.3806		
2500	Normal	23.0981	0.00000	0.7522	23.1010	0.00000	0.7738	23.1173	0.00000	0.7511
	Dichotomized	17.3765			17.8774			17.3650		



### Results Obtained for the Proportion of Variance Explained by the First Factor

Tables 5 and 6 show, for the three points of dichotomization, the means, the  $p$ -values of the test of mean difference (pC1c) for the results obtained from the proportion of variance explained by the first factor, and the proportions of the means of the variance explained by the first factor of the dichotomized samples, in comparison to the means of normal samples (D/N).

Table 5 shows the results from the test of mean difference for the proportion of variance explained by the first factor (pC1c) for the samples [3 2], with 5 variables and 2 factors. It can be noted that all differences are significant.

The results are always larger for normal samples. The variance explained by the first factor of the multivariate normal samples was always higher than 52%, and the variance explained by the first factor of the dichotomized samples always lower than 49%. It can be observed that the means of the dichotomized samples correspond, at least, to 82.65% of the mean of the multivariate normal samples ( $n = 250$ , 1<sup>st</sup> point of dichotomization), and a maximum of 88.56% ( $n = 10$ , 2<sup>nd</sup> point of dichotomization). The points of dichotomization on the comparison results show similar results.

Table 6 shows the results from the test of mean difference for the proportion of variance explained by the first factor (pC1c) for the samples [14 6 5 5 4 4 3 3 3 3], with 50 variables and 10 factors. The differences were all significant, with results always larger for normal samples.

The variance explained by the first factor of the multivariate normal samples was always higher than 23%, and the variance explained by the first factor of the dichotomized samples always lower than 19%. The means of the dichotomized samples corresponds to a minimum of 74.76% of the mean from the multivariate normal sample ( $n = 100$ , 1<sup>st</sup> point of dichotomization), and a maximum of 77.70% ( $n = 1000$ , 2<sup>nd</sup> point of dichotomization). The three points of dichotomization do not have influence on the results.

### Results Obtained for the Total Variance Explained

Tables 7 and 8 show, for the three points of dichotomization, the means of multivariate normal samples and dichotomized samples,  $p$ -values from the test of mean difference for the total variance explained (pCc), and the proportions of the means of the total variance explained of dichotomized samples, in comparison to the means of normal samples (D/N). Table 7 shows the results from the test of mean

## EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

difference for the total variance explained (pCc) for samples with 5 variables and 2 factors [3 2]. All the differences are significant, with results always larger for normal samples. The total variance explained of the factors of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 84%. The means from dichotomized samples correspond to a minimum of 83.87% of the mean from the multivariate normal samples ( $n = 250$ , 1<sup>st</sup> point of dichotomization), and a maximum of 89.57% ( $n = 10$ , 2<sup>nd</sup> point of dichotomization). The points of dichotomization do not have influence on the results.

Table 8 shows the results from the test of mean difference for the total variance explained (pCc) for samples with 50 variables and 10 factors [14 6 5 5 4 4 3 3 3 3]. All the differences are significant, with results always larger for normal samples.

The total variance explained of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 74%. The means from dichotomized samples correspond to a minimum of 79.72% of the mean from the multivariate normal sample ( $n = 2500$ , 3<sup>rd</sup> point of dichotomization), and a maximum of 84.25% ( $n = 100$ , 2<sup>nd</sup> point of dichotomization). Results do not vary among the three points of dichotomization.

Table 9 shows the results from the test of mean difference for the total variance explained (pCc) for samples with 50 variables and 10 factors [14 6 5 5 4 4 3 3 3 3]. All differences were significant, with results always larger for normal samples. The total variance explained of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 74%. The means from dichotomized samples correspond to a minimum of 79.72% of the mean from the multivariate normal sample ( $n = 2500$ , 3<sup>rd</sup> point of dichotomization), and a maximum of 84.25% ( $n = 100$ , 2<sup>nd</sup> point of dichotomization). The points of dichotomization do not have influence on the results.

### Results for the Communalities

Table 9 shows the comparisons of the communalities from multivariate normal and dichotomized data, only for the first sample simulation and the second dichotomization point. Only one point of dichotomization is presented (50/50), which is most widely used by researchers, since the dichotomization do not have influence on the results.

NOVAK & MARQUES

**Table 7.** Means and tests of significance of the total variance explained for samples with 5 variables and 2 factors – vector [3 2]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization			2 <sup>nd</sup> point of dichotomization			3 <sup>rd</sup> point of dichotomization		
		Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N
10	Normal	93.6434	0.00000	0.8706	93.3969	0.00000	0.8957	94.3784	0.00000	0.8710
	Dichotomized	81.5291			83.6629			82.2129		
25	Normal	89.9031	0.00000	0.8800	89.5597	0.00000	0.8869	89.9636	0.00000	0.8728
	Dichotomized	79.1224			79.4313			78.5291		
100	Normal	88.6243	0.00000	0.8615	89.1904	0.00000	0.8744	88.8288	0.00000	0.8529
	Dichotomized	76.3543			77.9938			75.7635		
250	Normal	87.0965	0.00000	0.8387	86.9444	0.00000	0.8563	86.9663	0.00000	0.8412
	Dichotomized	73.0546			74.4515			73.1579		

**Table 8.** Means and tests of significance of the total variance explained for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 <sup>st</sup> point of dichotomization			2 <sup>nd</sup> point of dichotomization			3 <sup>rd</sup> point of dichotomization		
		Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N
100	Normal	87.7839	0.00000	0.8273	87.8335	0.02110	0.8425	87.7222	0.00000	0.8254
	Dichotomized	72.6315			73.9998			72.4105		
250	Normal	86.9701	0.00000	0.8077	87.0215	0.00000	0.8262	86.9936	0.00000	0.8076
	Dichotomized	70.2458			71.8986			70.2610		
1000	Normal	86.5961	0.00000	0.7988	86.5931	0.00000	0.8186	86.5737	0.00000	0.7985
	Dichotomized	69.1756			70.8880			69.1296		
2500	Normal	86.4823	0.00000	0.7973	86.4811	0.00000	0.8166	86.4973	0.00000	0.7972
	Dichotomized	68.9531			70.6282			68.9640		

## EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

**Table 9.** Mean vectors and tests of significance of the communalities for samples with 5 variables and 2 factors – vector [3 2] – 2<sup>nd</sup> point of dichotomization

Size ( <i>n</i> )	Sample	Mean vector	pHc	Prop. D/N
10	Normal	[0.9279, 0.8841, 0.9872, 0.9338, 0.9368]	0.00000	0.8796
	Dichotomized	[0.8329, 0.7777, 0.8912, 0.8428, 0.8386]		0.9027
25	Normal	[0.7928, 0.9015, 0.9070, 0.9078, 0.9689]	0.00000	0.8289
	Dichotomized	[0.6572, 0.8065, 0.8097, 0.8152, 0.8830]		0.9066
100	Normal	[0.7740, 0.8928, 0.9129, 0.9113, 0.9685]	0.00000	0.7870
	Dichotomized	[0.6092, 0.8031, 0.8079, 0.8060, 0.8735]		0.9019
250	Normal	[0.8684, 0.8758, 0.8631, 0.9100, 0.8299]	0.00000	0.8199
	Dichotomized	[0.7580, 0.7437, 0.7555, 0.7849, 0.6805]		0.8753

This table shows mean vectors for the communalities of normal and dichotomized data referring to samples [3 2], with 5 variables and 3 factors, its respective p-values for the results from the test of mean difference for the mean vectors (pHc) and the minimum and maximum proportions given by the mean vectors of the communalities of the dichotomized samples in comparison to normal samples (D/N). The first value corresponds to the minimum proportion and the second to the maximum proportion. The table shows significant differences among the mean vectors, with communalities results always larger for normal samples.

The mean vectors from the dichotomized samples correspond to a minimum of 78.70% of the mean from the multivariate normal samples ( $n = 100$ ), and a maximum of 90.66% ( $n = 25$ ).

According to the analyses performed, the results do not show great variation for the three points of dichotomization.

### Conclusion

For the cases studied, the following conclusions can be drawn on the relationships resulting from an EFA between multivariate normal and dichotomized data:

- 1) For the MSA there is no regularity of values for normal data and its corresponding dichotomized data. The results suggest that, with the increase of factors and number of variables, the MSA for dichotomized data presents values higher than the values for normal

- data. The differences among the MSA means, with few exceptions on small samples, were always significant.
- 2) For the variance explained by the first factor, the total variance, and the communalities, the differences among the mean values for the normal and dichotomized data were always significant, and the values for normal data were always higher in comparison with the values for dichotomized data. Therefore, normal data always explains dichotomized data more efficiently.
  - 3) With regard to the points of dichotomization, in the acquisition of dichotomized data, it can be concluded that its results are very similar, having no influence on the analyses performed.

According to the results obtained for the MSA on the 378 simulations performed (number of variables  $14 \times$  samples sizes  $9 \times$  dichotomization  $3 = 378$ ), it can be verified that, for the cases involving 2 or 3 factors (simulations 1, 2, 3, 4, 5, and 6), the comparison between the MSA means of multivariate normal data and its respective dichotomized data have not presented significant differences in 6 (8%) cases. In 78 (48%) cases the MSA mean was significantly higher for multivariate normal data and in other 78 (48%) cases it was higher for dichotomized data. In the cases involving 4 to 10 factors and 8 to 50 variables (simulations 7 to 14), were verified 6 (3%) cases where the difference between the MSA means of multivariate normal data and its respective dichotomized data have not presented significant differences. In 21 (10%) cases the MSA mean was significantly higher for multivariate normal data and in the other 189 (87%) cases it was higher for dichotomized data.

In the cases involving 2 or 3 factors, it was verified that the differences were not influenced by the sample size, and in cases with 4 to 10 factors, MSA mean was higher for multivariate normal data only in small samples, with 2, 3 or 4 times the number of variables. Therefore, it can be concluded specially cases where the factor number is higher than 3 and the sample size corresponds to at least 5 times the number of variables, resulted in a higher MSA mean for dichotomized data. These results show that, in this situation, dichotomized data are adequate for the application of factor analysis.

According to the results obtained for the test of mean difference of the variance explained by the first factor between multivariate normal samples and its corresponding dichotomized samples, it was verified that all the 378 cases studied showed significant differences, with multivariate normal data means always higher. For samples with 2 or 3 factors (5 to 10 variables), the minimum ratio between D/N

## EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

(proportion of the dichotomized data mean to the multivariate normal mean) was 77.72% and the maximum 91.42%, as for samples with 4 to 10 factors (15 to 20 variables), the ratio was of 64.02% and 80.91% respectively. Therefore, for smaller numbers of factors (2 or 3) the loss of explanation by the first factor when data is dichotomized is less intense than in cases involving higher numbers of factors (4 to 10)

The test of mean difference for the total variance explained of the 378 cases studied have presented significant differences, with means always higher for multivariate normal data. For samples with 2 or 3 factors (5 to 10 variables) the D/N ratio ranged from 81.24% to 89.98%, and for samples with 4 to 10 factors (15 to 50 variables) ranged from 78.87% to 87.16% respectively. Therefore, similarly to the case of the variance explained by the first factor, the total variance explained also presents better results for smaller numbers of factors and variables.

The comparisons between mean vectors of the communalities of multivariate normal samples and the corresponding mean vectors of the communalities of dichotomized samples resulted in significant differences for the 378 cases studied. The components of these vectors were always higher for multivariate normal data. The D/N ratio between components for samples with 2 or 3 factors (5 to 10 variables) ranged from 71.07% to 93.67%, and for samples with 4 to 10 factors (15 to 50 variables) ranged from 55.79% to 94.54%. Therefore, it can be concluded for samples with smaller numbers of factors, the communalities results for dichotomized data, in relation to multivariate normal data, presents better results.

The substitution of multivariate normal data by its corresponding dichotomized data, using the phi correlation coefficient to calculate the correlation matrix, as an alternative to the tetrachoric correlation coefficient (since it is not possible to use this coefficient), will be always viable within the conditions analyzed for the MSA, variance explained by the first factor, total variance explained and communalities.

## References

Embreson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. New York, NY: Psychology Press.

Fedorov, V., Mannino, F., & Zhang, R. (2008). Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1), 50-61. doi: 10.1002/pst.331

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 45(1). Retrieved from

[http://www.pabst-publishers.de/psychology-science/1-2003/abstract\\_06.html](http://www.pabst-publishers.de/psychology-science/1-2003/abstract_06.html)

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19-40. doi: 10.1037/1082-989x.7.1.19

Pearson, R. H., & Mundform, D. J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods*, 9(2), 359-368. doi: 10.22237/jmasm/1288584240