Assessing an effect size from dichotomized data

T. O. Alakija[a], I. A. Adeleke[b] and R. Okafor[c]

[a]*Department of Statistics, Yaba College of Technology, Yaba, Lagos, Nigeria*; [b]*Department of Actuarial Sciences, University of Lagos, Lagos, Nigeria*; [c]*Department of Mathematics, University of Lagos, Lagos, Nigeria*

Dichotomizing continuous data usually results in the loss of information and has the consequence of reducing statistical power when the objective is to test for a statistical relationship between variables. This article determines the cost of dichotomizing continuous data on the value of an effect size that is, 'a quantitative measure of the strength of a phenomenon' using both simulated data and education data. The data set was estimated using the correlation coefficient, chi-square and odds ratio to investigate the relationship between continuous assessment score and examination score among students. The result shows that the value of the correlation coefficient decreased by 23.5% after dichotomizing the two variables as compared to before dichotomization, the value of the chi-square decreased by 98.9% while the value of the odds ratio increased by 11.2%. Hence, the effect of dichotomization on the strength of association between the two variables using the three different statistical methods differ.

Keywords: effect size; dichotomization; correlation; odds ratio; chi-square.

1. Introduction

Dichotomization converts continuous variables into categorical variables by grouping the values into two categories. It includes asserting that there is a straight line of effect between one variable and another which are often much easier to deal with statistically. Some advantages of forcing all individuals into two groups are that it greatly simplifies statistical analysis and leads to easy interpretation and presentation of results (Douglas and Patrick, 2006). Dichotomization has been used when researchers believe there are distinct groups of individuals or is interested in group differences rather than individual differences (Iacobucci et al., 2015). It involves utilizing a median split to create equal groups in an ex-post facto method which helps to simplify the statistical analyses and interpretation of the results. This however is believed to result in a loss of variability (MacCallum et al., 2002). Measurements of continuous variables are often made in medicine and other experimental sciences, aiding in the diagnosis and treatment of patients. In clinical practice it is helpful to label individuals as having or not having an attribute, such as being 'hypertensive' or 'obese' or having 'high cholesterol', depending on the value of a continuous variable. In academics, a student can be categorize as having good standing or not. It could be a pass or a fail. Categorization of continuous variables saves us the need for assumptions about the nature of the relation between the variable and the outcome. It makes the analysis and interpretation of results simple as it is quite easy to ascertain the work done and what the results are. Dichotomization however, leads to several problems and is usually criticized by statisticians. A binary split at the median leads to a comparison of groups of individuals with high or low values of the measurement, leading in the simplest case to a t test or chi-square test and an estimate of the difference between the groups with its confidence interval. There is, however, no good reason in general to suppose that there is an underlying dichotomy, and if one exists there is no reason why it should be at the median (MacCallum et al., 2002).

For some variables under dichotomization, there are recognised cut-points, such as $>$ 25kgm^2 to define 'overweight' based on body mass index. In the absence of a prior cut-point, the most common approach is to take the sample median. However, using the sample median implies that various cut-points will be used in different studies so that their results cannot easily be compared; seriously hampering meta-analysis of observational studies (Buettner et al., 1997). Use of the 'optimal' cut-point runs a high risk of a spuriously significant result. That is, the difference in the outcome variable between the groups will be overestimated, perhaps considerably; and the confidence interval will be narrow (Altman et al.,1994; Royston et al.,2006).

Dichotomization may increase the risk of a positive result being a false positive (Austin and Brunner, 2004). Individuals close to but on opposite sides of the cut-point are characterised as being very different rather than very similar. Also using two groups conceals any non-linearity in the relation between the variable and outcome. When regression is used to adjust for the effect of a confounding variable, dichotomization runs the risk that a substantial part of the confounding remains (Austin and Brunner, 2004; Royston et al., 2006). Nevertheless, all these approaches are preferable to performing several analyses and choosing that which gives the most convincing result.

Dichotomisation can result in the loss of information and power (Altman and Royston, 2006; MacCallum et al., 2002) which makes the statistical power to detect a relationship between the variable and subject outcome reduced. It has been emphasized that dichotomization is appropriate only when a threshold effect value truly exists. That is, if we can assume some binary split of the continuous covariate (Abdolell et al., 2002).

Rousson (2014) studied the consequences of dichotomizing continuous data on the value of an effect size in some classical settings. It turns out that the conclusions will not be the same whether using a correlation or an odds ratio to summarize the strength of association between the variables; she illustrated her work using a data set to investigate the relationship between motor and intellectual functions in children and adolescents.

In statistics analysis, the effect size can usually be measured in different ways. The correlation, chi-square and odds ratio are established concepts of inferential statistics to measure in a symmetric way the strength of an association (or an 'effect size') between two variables. In practice, one usually calculates a correlation if the variables are continuous, calculate chi-square if the variables are categorical and one calculates an odds ratio if the variables are binary. In this paper, we want to see whether or not the dichotomization of continuous data often has the effect of decreasing the value of a correlation coefficient, chi-square and odds ratio(OR). This research is therefore, aimed at measuring the effect size of dichotomization of continuous data using correlation, chi-square and odds ratio.

2. Dichotomization and effect size

Effect size is a statistical concept that measures the strength of the relationship between two variables on a numeric scale and it tells the differences in data regardless of sample size. Effect sizes show practical or meaningful differences instead of simply statistical differences. Effect size can be thought of as a measurement of the amount of impact an independent variable has on a dependent variable (Murphy and Myors, 1998, p. 12).

Kristopher et al.(2005) presented in their work, the use of the Extreme Groups Approach: A Critical Re-examination and New Recommendations, that the analysis of continuous variables sometimes proceeds by selecting individuals on the basis of extreme scores of a sample distribution and submitting only those extreme scores to further analysis The authors illustrate the effects extreme groups approach can have on power, standardized effect size, reliability, model specification, and the interpretability of results.

Estimates of effect size can be classified as unstandardized or standardized. Unstandardized effect size estimates reflect the magnitude of an effect in raw units of whatever is being measured while the Standardized effect size estimates like (r_{xy},R^2,ω^2,η^2,cohen's d) are expressed in common metrics unrelated to the raw scales of measurement of the observed variables(Kristopher et al., 2005).

Generally, dichotomization of continuous variable can increase or decrease an effect size depending on the method of measuring the association between the two groups. When X

and Y are continuous random variables, one of the appropriate estimator of an effect size of a linear relationship between X and Y is the Pearsons correlation coefficient, r (rho), which measures the strength of the linear relationship between two variables on a continuous scale A typical example for quantifying the association between two variables measured on an interval/ratio scale is the analysis of relationship between a students continuous assessment (CA) score and examination score. Each of these two characteristic variables is measured on a continuous scale. When a continuous variable is dichotomized, it will give a lesser value of correlation than if not dichotomized. We are excluding the frequently difficult technique of creating a dichotomous variable by arbitrarily dichotomizing originally continuous scores into two groups; below the median and above the median known as the median split.

If one variable is measured continuous and the second variable is dichotomous (has two outcomes), then the point-biserial correlation coefficient is appropriate. The point-biserial correlation is mathematically equivalent to the Pearson correlation, that is, if we have one continuously measured variable X and a dichotomous variable Y, $r_{xy} = r_{pb}$. This can be shown by assigning two distinct numerical values to the dichotomous variable. Other combinations of data types (or transformed data types) may require the use of more specialized methods to measure the association in strength and significance.

The chi-square test for association (contingency) is a standard measure for association between two categorical variables. The chi-square test, unlike Pearson's correlation coefficient or Spearman rho, is a measure of the significance of the association rather than a measure of the strength of the association.

To reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2×2 contingency table (Yates, 1934).This reduces the chi-squared value obtained and thus increases its p-value (Sokal and Rohlf, 1981).The effect of Yates' correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5. Unfortunately, Yates' correction may tend to over-correct.

The phi coefficient (also referred to as the 'mean square contingency coefficient' and denoted by φ (or $r\varphi$)) is a measure of association for two binary variables. Introduced by Karl Pearson (Cramer, 1946)),this measure is similar to the Pearson correlation coefficient in its interpretation. In fact, a Pearson correlation coefficient estimated for two binary variables will return the phi coefficient (Guilford, 1936). The square of the Phi coefficient is related to the chi-squared statistic for a 2×2 contingency table (Everitt, 2002). Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal. Phi is related to the point-biserial correlation coefficient and Cohen's d and estimates the extent of the relationship between two variables (2×2) (Aaron et al., 1998).

The OR is one of three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population. If each subject in a population either does or does not have a property 'A', (e.g. 'fail exam'), and also either does or does not have a property 'B' (e.g. 'fail test') where both properties are appropriately defined, then a ratio can be formed which quantitatively describes the association between the presence/absence of 'A' (fail exam) and the presence/absence of 'B' (fail test) for subjects in the population. This ratio is the OR and can be computed following these steps. For a given individual that has 'B' compute the odds that the same individual has 'A'. For a given individual that does not have 'B' compute the odds that the same individual has 'A'. Divide the odds from step 1 by the odds from step 2 to obtain the OR (Cornfield, 1951; Mosteller and Frederick, 1968; Edwards,1963).

3. Materials and method

3.1 *Data Description*

The data used for this research are from the raw test scores and examination scores for a particular course of some students in Yaba College of Technology for the just concluded semester. The test score is a proxy for the CA and the semester examination scores represent exam scores. It is expected that students with high CA will also score high in examination. The CA score is graded over 30 while the examination score is graded over 70.

3.2 *Methods of dichotomization*

- Median split – dichotomizing variables at the median to create equal 'high' and 'low' group. Median split involves when we divide a sample into two groups based on whether each score on a continuous predictor variable is above or below the median prior to conducting analyses (Iacobucci et al., 2015a). This is the most common method for dichotomization (MacCallum et al., 2002).
- Quartile splits – dichotomizing at the quartiles could be used to create groups who are 'high' or 'low' using the third or first quartile (respectively). However, this would cut out data between the first and third quartile range. Alternatively, if the groups consist of the third quartile and higher as one group and all data points lower than the third quartile, this will result in unequal sample sizes as well.
- Mean split – dichotomizing variables at the mean which simply involves dividing data into groups which could result in unequal groups if the data is skewed or there exists the presence of outliers. The mean spilt method involves splitting the data based on their characteristics of interest and then finding the mean of both groups for example, using the smoker variable, Patients are divided into smokers and non-smokers and then the mean weight for each group is calculated.

3.3 *Evaluating an effect size by correlation measures*

The (Pearson) correlation r between two variables X and Y is defined as the covariance between X and Y divided by the product of their standard deviations. Equivalently, this is the least squares regression slope of Y on X, multiplied by the standard deviation of X, and divided by the standard deviation of Y, yielding:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \beta\sqrt{\frac{Var(X)}{Var(Y)}} \quad (3.1)$$

where $\beta = Cov(\text{X,Y})/Var(\text{X})$. This definition applies that as long as X and Y are continuous and also when X *and/or* Y are binary. If X is binary (with possible values 0 and 1), the regression slope is simply a mean difference. If both X and Y are binary, the regression slope is a difference of proportions. When X is binary and Y is continuous, r is sometimes called the 'point biserial correlation', whereas r is known as the 'phi coefficient' when X and Y are binary (Rousson, 2014).The value of the effect size of Pearson r correlation varies between -1 to +1. The effect size is low if the value of r varies around 0.1, medium if r varies around 0.3, and large if r varies more than 0.5. (Cohen, 1992).

3.4 *Evaluating an effect size by chi-square measures*

The Chi-square (χ^2) between two variables X and Y is defined as the ratio of the square of the difference between the observed and expected value and the expected value. The χ^2 is mostly used for count data. If the variables are categorical, then χ^2 is appropriate. However, there are different measures of the χ^2, we have interval- interval, interval-ordinal, ordinal-ordinal and binary-binary and the Yates corrected χ^2. In practice, the χ^2 is computed as

$$\chi^2 = \sum_{i=1}^{r}\sum_{i=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.2}$$

where O_{ij} is the observed frequency at row i and column j from a contingency table or cross tabulation of variables X and Y. E_{ij} is the expected frequency at row i and column j.

3.5 *Evaluating an effect size by odds-ratio measures*

The OR is another suitable effect size. It is appropriate when the research question focuses on the degree of association between two binary variables. The OR is another concept to the correlation for measuring the strength of association between two quantitative variables X and Y. The OR has been generalized to quantitative (including continuous) variables by Agresti (1980). The OR ω between two quantitative variables X and Y is defined as:

$$\omega = \frac{Pr\{Y_1 > Y_2 \mid X_1 > X_2\}}{Pr\{Y_1 < Y_2 \mid X_1 > X_2\}} \tag{3.3}$$

where (X_1, Y_2) and (X_2, Y_2) are two independent observations of (X, Y). The numerator is referred to as the 'probability of concordance' and the denominator as the 'probability of discordance'.

Let X and Y be two continuous variables. In this research, we shall restrict our attention to the case where (X, Y) is uniformly distributed with correlation r, where we consider without loss of generality that X and Y have minimum of 0 and maximum 1.

Since Y is continuous, the probability to have a tie $(Y_1 = Y_2)$ is zero such that:

$$\omega = \frac{Pr\{Y_1 > Y_2 \mid X_1 > X_2\}}{1 - Pr\{Y_1 > Y_2 \mid X_1 > X_2\}} \tag{3.4}$$

One has then:

$$\Pr\{Y_1 > Y_2 \mid X_1 > X_2\} = \frac{Pr\{X_1 > X_2 \text{ and } Y_1 > Y_2\}}{Pr\{X_1 > X_2\}} \tag{3.5}$$

$$\Pr\{Y_1 > Y_2 \mid X_1 > X_2\} = \frac{Pr\left\{\frac{X_1 - X_2}{\sqrt{2}} > 0 \text{ and } \frac{Y_1 - Y_2}{\sqrt{2}} > 0\right\}}{Pr\left\{\frac{X_1 - X_2}{\sqrt{2}} > 0\right\}} \tag{3.6}$$

Since the binormal of (X, Y) implies that X_1, X_2, Y_1, and Y_2 are uniformly normally distributed, the variables $\frac{X_1 - X_2}{\sqrt{2}} > 0$ and $\frac{Y_1 - Y_2}{\sqrt{2}} > 0$ are also uniformly normally distributed and one may check that their correlation is still equal to r. One has thus:

$$\Pr\{Y_1 > Y_2 \mid X_1 > X_2\} = \frac{PrX > 0 \text{ and } PrY > 0}{PrX > 0} = \frac{A}{A + D} \tag{3.7}$$

where A and D are defined as:
A=Pr X_d=1 and Y_d=1
B=Pr X_d=0 and Y_d=1

C=Pr X$_d$=0 and Y_d=0
D=Pr X_d=1 and Y_d=0
with $P_X = P_Y = 0.5$. Since D = $\frac{1}{2}$ - A, one has PrY_1¿Y_2 — X_1¿X_2 = 2A and hence $\omega = \frac{2A}{1-2A}$.

4. Result and discussion

A sample of size $m = 100$ from a uniform distribution with parameters $a = 20$ and $b = 70$ and replicated n = 1000 times was used to simulate the examination score and a sample of size m = 100 from a uniform distribution with parameters $a = 0$ and b = 30 and replicated n = 1000 times was used to simulate the continuous assessment score. The two variables were sorted to have high collinearity. The simulation was repeated for the exam and CA using normal distribution with parameters with parameters $\mu = 37$ and $\sigma = 10$ for exam while parameters $\mu = 16$ and $\sigma = 9$ are used for the CA. The average scores after sorting for the exam and the CA were taken to represent the simulated data.

The descriptive statistics of the result of the simulated result is displayed on Table 1. While that of the real life data is displayed in table 2. Figure 1 and Figure 2 depict the unsorted simulated scatter plot for uniform and normal generated data.

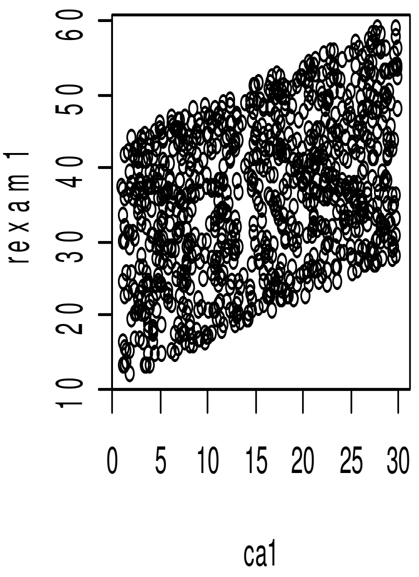

Figure 1: Simulated Uniform Distribution

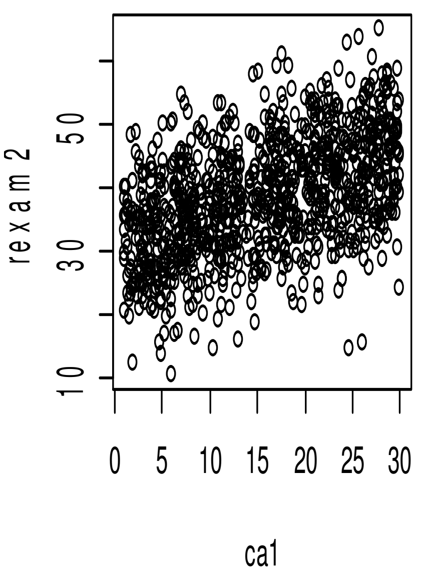

Figure 2: Simulated Normal Distribution

Figure 2 (left) depicts the scatter plot of the average sorted simulated data, showing a near perfect correlation and (right) depicts the scatter plot of the real-life data (actual result of the students). Figure 4(left and right)shows the boxplot of simulated CA and Exam respectively, while Figure 5 (left and right) shows that of the real-life data. Figure 3 shows the normalized score of the real-life data in a scatter plot portioned into A, B, C and D.

The actual data were normalized using the formula stated below as:

$$Y_i = \frac{Exam_i - Exam_{\min}}{Exam_{\max} - Exam_{\min}}$$

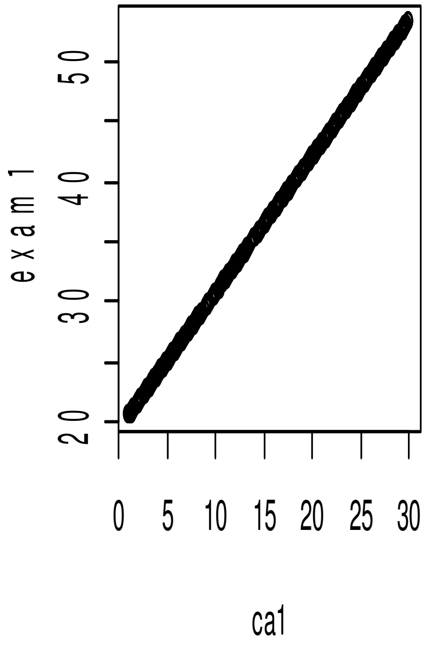

Figure 3: Simulated Regression

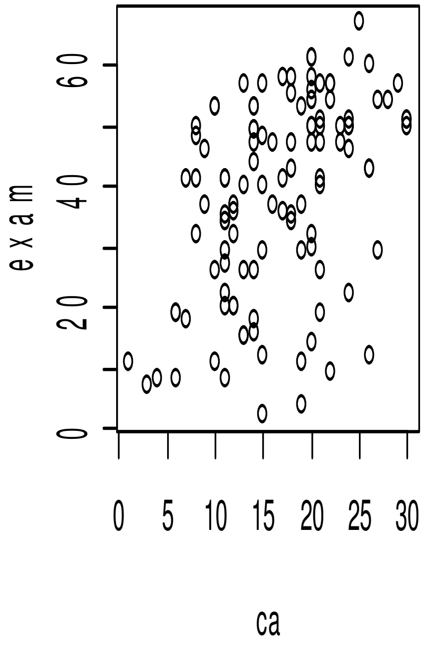

Figure 4: Actual Data- Exam on CA

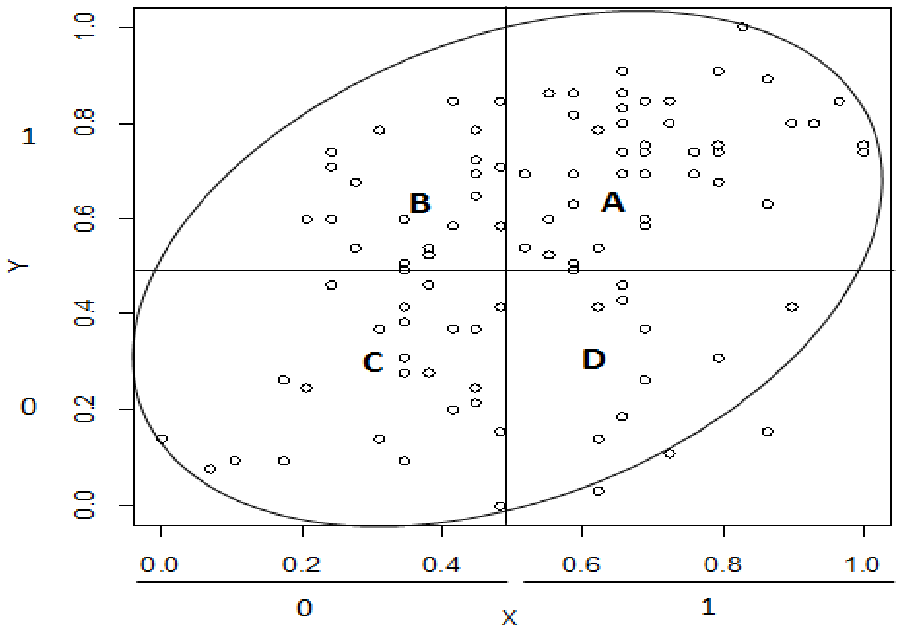

Figure 5: Normalized between 0 and 1

and

$$X_i = \frac{CA_i - CA_{\min}}{CA_{\max} - CA_{\min}}$$

where $Exam_i$ is the exam score for student i and CA_i is the CA score for student i. $Exam_{max}$ and $Exam_{min}$ are the highest exam and lowest exam scores respectively. CA_{max} and CA_{min} are the highest and the lowest CA scores respectively. Y_i is the normalized exam

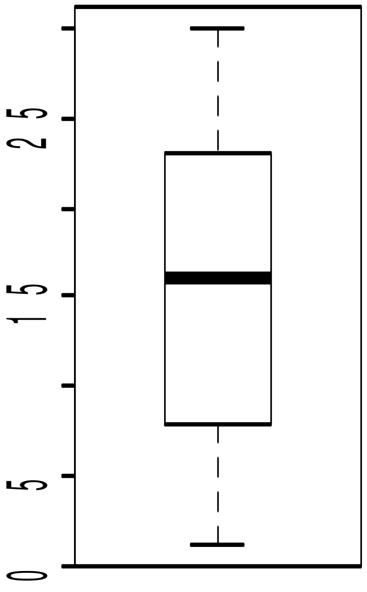

Figure 6: Box plot for simulated CA

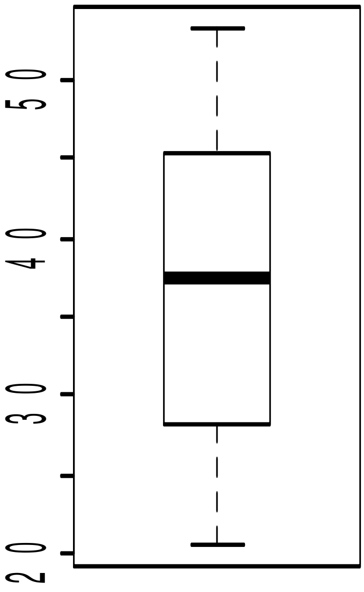

Figure 7: Box plot for simulated exam

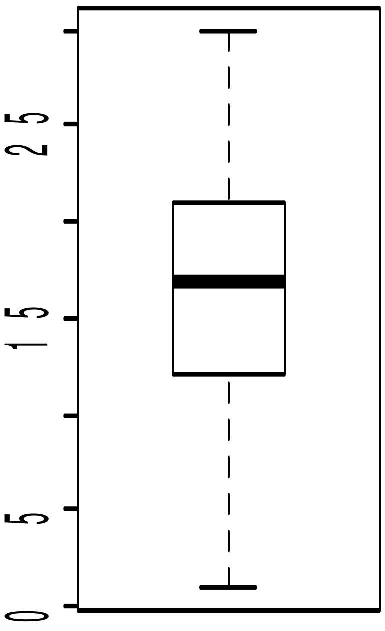

Figure 8: Box plot for actual CA

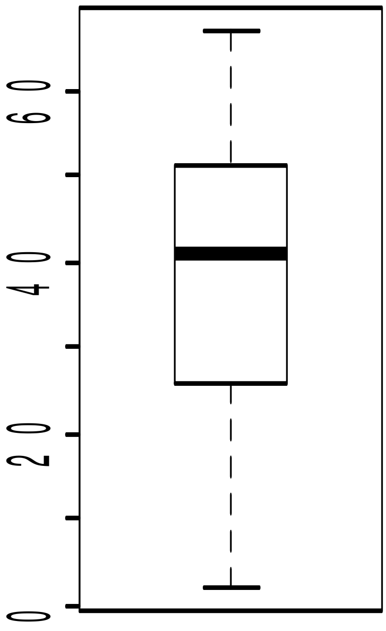

Figure 9: Box plot for actual exam

Figure 1. Box Plot for Actual CA (left) & Box Plot for Actual Exam (right)

score for student i while X_i is the normalized CA score for student i.

4.1 *Simulated data analysis*

Table 1: Descriptive Statistics of Simulated Data

Variable	Obs	Mean	Std.Dev.	Min	Max
exam	1000	36.99	9.66	20	53
CA	1000	15.67	8.52	1	30

Table 2: Descriptive Statistics of Normalized Simulated Data Between 0 and 1

Variable	Obs	Mean	Std.Dev.	Min	Max
Y_d	1000	0.5080	0.5002	0	1
X_d	1000	0.5200	0.4998	0	1

For the normalized variables, the minimum is zero and the maximum is 1. It is from these values that we dichotomized the continuous variables. Values below the mean are 0, while values greater than or equal to the mean are 1. So, the dichotomized variables are Y_d and X_d for Exam and CA respectively.

Table 3: Contingency Table of Simulated Exam and CA

		Exam	
		0	1
	0	480	0
CA	1	12	508

Table 4: Probability Distribution Table

		Exam	
		0	1
	0	0.4800	0.000
CA	1	0.1002	0.5080

A = 0.5080, B = 0.0000, C = 0.4800, D = 0.1200,
P_X = B + C = 0.0000 + 0.4800 = 0.4800,
P_Y = D + C = 0.1200 + 0.4800 = 0.6000,
A + B + C + D = 0.5080+ 0.0000 + 0.4800 + 0.1200 = 1.
Note that r is the correlation between X and Y, r_1 is the correlation between X_d and Y while r_2 is the correlation between X_d and Y_d.

$$r{=}0.997$$
$$r_1{=}0.847$$

$$r_2 = \frac{AC - BD}{\sqrt{(A+D)(B+C)(A+B)(D+C)}}$$

$$= \frac{(0.5080)(0.4800) - (0.0000)(0.1200)}{\sqrt{(0.5080+0.1200)(0.0000+0.4800)(0.5080+0.0000)(0.1200+0.4800)}}$$

$$= 0.7820$$
$$\text{OR} = \text{AC/BD} = \infty$$

The OR is the ratio of the probability of concordance to that of the probability of discordance.

Measuring Effect Size - Correlation Coefficient
Pearson correlation coefficient between Y and X = 0.997 with P-Value = 0.000
Pearson correlation coefficient between Y_d and X = 0.847 with P-Value = 0.000
Pearson correlation coefficient between Y and X_d = 0.847 with P-Value = 0.000
Pearson correlation coefficient between Y_d and X_d = 0.976 with P-Value = 0.000
The coefficient of correlation was reduced from 0.976 for both continuous variables to 0.976 after dichotomization of the continuous variables. The correlation coefficient was reduced by 2.1%.

Measuring Effect Size - Chi-square
The Pearson Chi-Square for Y and X = 23205.010, with degrees of freedom = 1620
The Pearson Chi-Square for Y_d and X_d= 953.096, with degrees of freedom = 1.
The Chi-Square was reduced from 23205.010 for both continuous variables to 953.096 after dichotomization of the continuous variables. The Chi-Square was reduced by 95.89%.

Measuring Effect Size - OR
The OR between Y_d and X_d using binary regression is 6420498.580 with P-value = 0.000. The OR between Y_d and X_d using binary regression is 68388435053.045 with P-value = 0.000. The OR was increased from 6420498.580 for one continuous and one dichotomized variable to 68388435053.045 after dichotomization of the continuous variables. The odds ratio was increased by 99.99%.

4.2 *Application Data*

Table 5: Descriptive Statistics of Raw Scores

Variable	Obs	Mean	Std.Dev.	Min	Max
exam	113	38.12	16.26	2	67
CA	113	16.66	6.39	1	30

Table 6: Descriptive Statistics of Normalized Data Between 0 and 1

Variable	Obs	Mean	Std.Dev.	Min	Max
exam	113	0.5558	0.25012	0	1
CA	113	0.5401	0.22049	0	1

For the normalized variables, the minimum is zero and the maximum is 1. It is from these values that we dichotomized the continuous variables. Values below the mean are 0, while values greater than or equal to the mean are 1. So, the dichotomized variables are Y_d and X_d for Exam and CA respectively.

Table 7: Contingency Table of Exam and CA

		Exam	
		0	1
	0	29	24
CA	1	13	47

Table 8: Probability Distribution Table

		Exam	
		0	1
	0	0.2566	0.2124
CA	1	0.1150	0.4159

A = 0.4159, B = 0.2124, C = 0.2566, D = 0.1150
P_X = B + C = 0.2124 + 0.2566 = 0.4690
P_Y = D + C = 0.1150 + 0.2566 = 0.3717

A + B + C + D = 0.4159 + 0.2124 + 0.2566 + 0.1150 = 1

Note that r is the correlation between X and Y, r_1 is the correlation between X_d and Y while r_2 is the correlation between Y_d and X_d.

$$r{=}0.446$$
$$r_1{=}0.395$$

$$r_2 = \frac{AC - BD}{\sqrt{(A+D)(B+C)(A+B)(D+C)}}$$

$$= \frac{(0.4159)(0.2566) - (0.2124)(0.1100)}{\sqrt{(0.4159+0.1150)(0.2124+0.2566)(0.4159+0.2124)(0.1150+0.2566)}}$$

$$= 0.341$$
$$\text{OR} = \text{AC/BD} = 4.369$$

The OR is the ratio of the probability of concordance to that of probability discordance.

Measuring Effect Size - Correlation Coefficient
Pearson correlation coefficient between Y and X = 0.446 with P-Value = 0.000
Pearson correlation coefficient between Y_d and X = 0.356 with P-Value = 0.000
Pearson correlation coefficient between Y and X_d = 0.395 with P-Value = 0.000
Pearson correlation coefficient between Y_d and X_d = 0.341 with P-Value = 0.000
The coefficient of correlation was reduced from 0.446 for both continuous variables to 0.341 after dichotomization of the continuous variables. The correlation coefficient was reduced by 23.54%.

Measuring Effect Size - Chi-square
The Pearson Chi-Square for Y and X = 1239.215, with degrees of freedom = 1107
The Pearson Chi-Square for Y_d and X_d= 13.163, with degrees of freedom = 1.
The Chi-Square was reduced from 1239.215 for both continuous variables to 13.163 after dichotomization of the continuous variables. The Chi-Square was reduced by 98.9%.

Measuring Effect Size - OR
The OR between Y_d and X_d using binary regression is 3.882 with P-value = 0.000.
The OR between Y_d and X_d using binary regression is 4.369 with P-value = 0.000.
The OR was increased from 3.882 for one continuous and one dichotomized variable to 4.369 after dichotomization of the continuous variables. The odds ratio was increased by 11.15%.

5. Conclusion and recommendation

The practice of dichotomization has been strongly criticized in the literature, mainly because of the loss of statistical power mentioned, but also because of the interpretational problems caused by dichotomization. This is especially true with normally distributed variables, where a majority of the observations lie near the median. In that case, many observations that are close apart on the continuous scale will not be in the same category after a median dichotomization. In this research, our purpose is to study the effect of dichotomization of continuous variable on the value on an effect size. For this, we have retrieved and gathered several formulae to calculate a correlation, chi-square and an OR in classical settings involving dichotomized continuous variables. We observe that a dichotomization made at the median of a normal distribution will decrease the value of our result using correlation by 23.54%, it will decrease the result of the chi-square value by 98.9% and will increase the OR by 11.2%. The decrease of the correlation and chi-square as well as the increase of the OR is still more important if the dichotomization is made away from the median.

The square of a correlation can be interpreted as the percentage of variance of one variable that can be linearly predicted by the other. The value of a chi-square can be regarded as the measure of the discrepancies between observed and expected frequencies. An OR can be interpreted as the ratio of the probabilities of concordance and discordance. Which is the most relevant information? It is probably a matter of taste. Interestingly, the usual practice which consists of calculating a correlation in the case of continuous variables, the calculation of chi-square in the case of categorical variable and calculating an OR in the case of binary variables, is the one for which the chosen measure of strength of association cannot be improved by changing the nature of the scales of the variables, a correlation being at its highest when the scales are continuous, a chi-square is best for categorical data while an odds ratio being at its highest when the scales are binary.

There are still other ways to quantify an association between two variables, which may provide still other messages regarding the effect of dichotomization on the strength of association. Instead of considering the ratio of the probabilities of concordance and discordance, one may consider their difference, obtaining Kendalls t in the case of continuous variables, in the case of a binormal distribution with correlation r.

In conclusion, while it is true, from an inferential statistics point of view, that the consequences of dichotomizing continuous data will be in most cases a 'loss of statistical power', it is not clear, from a descriptive statistics point of view, whether it leads at the same time to a 'loss of effect size', the conclusion depends on the measure of association, which is used to quantify this effect. It is therefore recommended to study many literature to ascertain the best method of measuring association between two variables. We however recommend the use of correlation for two continuous variables, chi-square for categorical variables and odds ratio for two binary variables.

Acknowledgement

The authors would like to thank the Editor-in-Chief and the anonymous reviewers for their valuable comments which have greatly improved the quality of the earlier manuscript.

References

Aaron, B., Kromrey, J. D. and Ferron, J. M. (1998). Equating r-based and d-based effect-size indices: problems with a commonly recommended formula. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433353)

Abdolell, M., LeBlanc, M., Stephens, D. and Harrison, R. V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat. Med.*, 21(22):3395-409.

Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994). Dangers of using 'optimal' cut points in the evaluation of prognostic factors. *J. Natl. Cancer Inst.*, 86: 829–35. [PubMed]

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, 332(7549): 1080.

Agresti, A. (1980). Generalized Odd-Ratio for Ordinal Data. *Biometrics*, 36: 59-67.

Austin, P.C. and Brunner L.J. (2014). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Stat. Med.* 23: 1159-78. [PubMed]

Buettner, P., Garbe, C., Guggenmoos-Holzmann, I. (1997). Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma. *J. Clin. Epidemiol.*, 50: 1201-10. [PubMed]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychology Measure*, 7: 249-53

Cornfield, J (1951). *"A Method for Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix"*. Journal of the National Cancer Institute. 11: 12691275. PMID 14861651.

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Davenport, E. and El-Sanhury, N. (1991). Phi/Phimax: review and synthesis. *Educational and Psychological Measurement*, 51: 821828.

Del Priore G, Zandieh P, Lee M.J. (1997). Treatment of continuous data as categoric variables in obstetrics and gynecology. 89: 351-4. [PubMed]

Edwards, A.W.F. (1963). The measure of association in a 2x2 table. *Journal of the Royal Statistical Society*, Series A, 126 (1): 109114. doi:10.2307/2982448.

Everitt B.S. (2002) *The Cambridge Dictionary of Statistics*, CUP.

Guilford, J. (1936). *Psychometric Methods.* New York: McGrawHill Book Company, Inc.

Iacobucci, D., Posavac, S.S., Kardes, F.R., Schneider, M.J. and Popovich, D.L. (2015). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology*, 25: 652-665.

Sokal, R. R. and Rohlf, F.J. (1981). Biometry: *The Principles and Practice of Statistics in Biological Research.* Oxford: W.H. Freeman.

Kristopher, J., Preacher, *et al.* (2005). The use of the extreme groups approach: a critical reexamination and new recommendations. *Journal of the American Psychological Association*, 10(2): 178-192.

MacCallum, R. C., Zhang, S., Preacher, K. J. and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1): 19-40.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63(321): 128. doi:10.2307/2283825.

Murphy, K.R. and Myors, B. (1998). *Statistical power analysisA simple and general model for traditional and modern hypothesis tests*: Mahwah, N.J., Lawrence Erlbaum Associates, Inc.

Rousson, V. (2014). Measuring an Effect Size From Dichotomized Data: Contrasted Results Whether Using a Correlation or an Odds Ratio. *Journal of Educational and Behavioral Statistics* 39(2): 144163. DOI: 10.3102/1076998614524597

Royston, P., Altman, D. G. and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25: 127-41. [PubMed]

Tabachnick, B.G. and Fidell, L.S. (2007). *Using multivariate statistics*, fifth edition. New York, NY: Pearson Education, Inc.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. Supplement to the *Journal of the Royal Statistical Society*, 1(2): 217235.