The danger of dichotomizing continuous variables: A visualization

Oliver Kuss

Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, University of Halle-Wittenberg, Halle (Saale), Germany

Address for correspondence:
Oliver Kuss, Institute of Medical Epidemiology, Biostatistics, and Informatics, Medical Faculty, Martin-Luther-University of Halle-Wittenberg, Magdeburger Str. 8, 06097 Halle (Saale), Germany, Phone: +49-345-557-3582, Fax: +49-345-557-3580, e-mail: oliver.kuss@medizin.uni-halle.de

Abstract
Four rather different scatterplots of two variables X and Y are given, which, after dichotomizing X and Y, result in identical fourfold-tables misleadingly showing no association.

Keywords
Dichotomization, Scatterplot, Fourfold-table

Statistical analysis is about reducing complexity in data while keeping information loss to a minimum. Statisticians have been extremely successful using this idea, but sometimes and unfortunately reducing complexity can result in obscuring most relevant information in the data.
In medical research, clinicians typically like the idea of dichotomizing continuous variables. They argue that this simplifies analysis and interpretation of results, moreover, clinical decision making frequently requires two categories, such as normal/abnormal or treat/do not treat (1). Warnings against these practices have often been issued, because dichotomizing results in the loss of efficiency, lower statistical power, lower reliability, and inflated type I and type II errors (2), but these warnings are frequently ignored.

In 1973, Anscombe (3) gave four scatterplots describing very different associations between two variables X and Y that, however, yield identical means and variances for X and Y. Moreover, identical correlation coefficients and regression lines emerge when assessing the association of X and Y.

Inspired by this idea, 5 scatter plots are given here that yield an identical fourfold table when dichotomized (figure 1). In each of these plots there are 100 observations with observed values for two continuous variables X and Y. Dichotomizing both X and Y at the cut-point 0 in each of these 5 cases yields a fourfold-table with having 25 observations in each cell, pointing to a null association (e.g., an odds ratio of 1) between the dichotomized versions of X and Y. This is perfectly right in figure 1 a) where the scatterplot shows no association between X and Y, and so inference from both the continuous and the dichotomized version of X and Y agrees.

However, dichotomizing X and Y and claiming no association from the resulting fourfold-table overlooks a quadratic association between X and Y in b), a more complicated sinusoidal non-linear association in c), heteroscedasticity in d), and a mixture of two populations in e). All of this additional information in the original scatterplots gets lost if only the fourfold table of the dichotomized variables is considered.

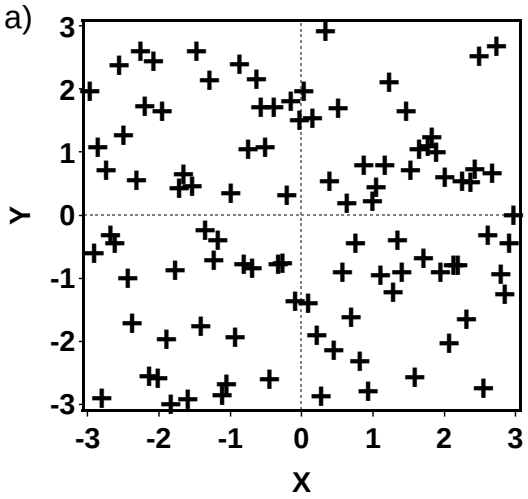
Simply dichotomizing continuous variables without previously referring to the original distributions by plotting them and checking consequences of dichotomization is a bad idea (4) and should be discouraged.

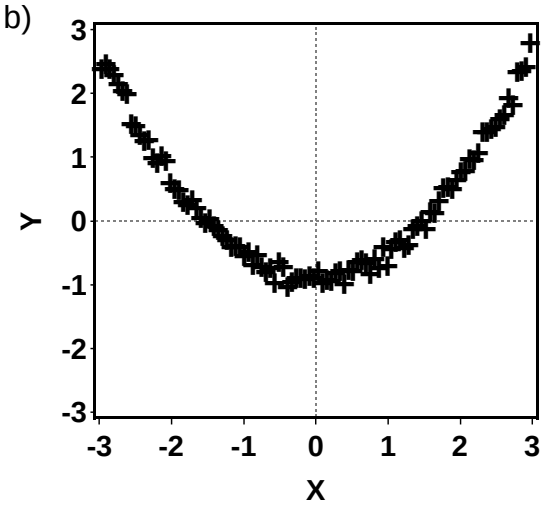
Table 1:

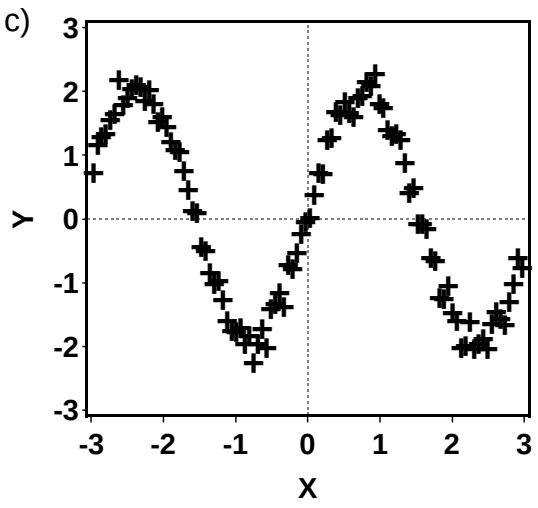
Resulting fourfold table from dichotomizing X and Y at the value 0 from all plots in figure 1.

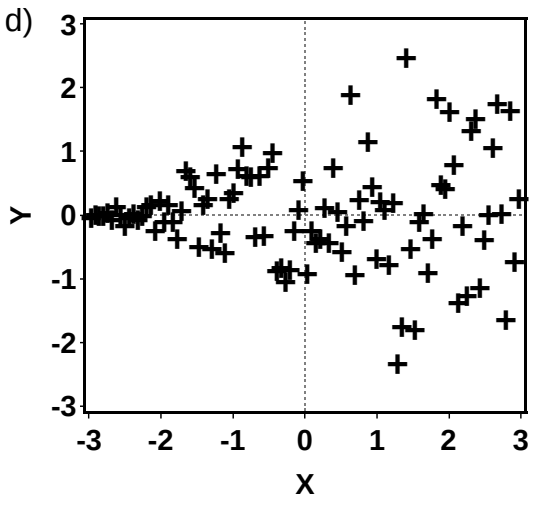
		Y positive?	
		Yes	No
X positive?	Yes	25	25
	No	25	25

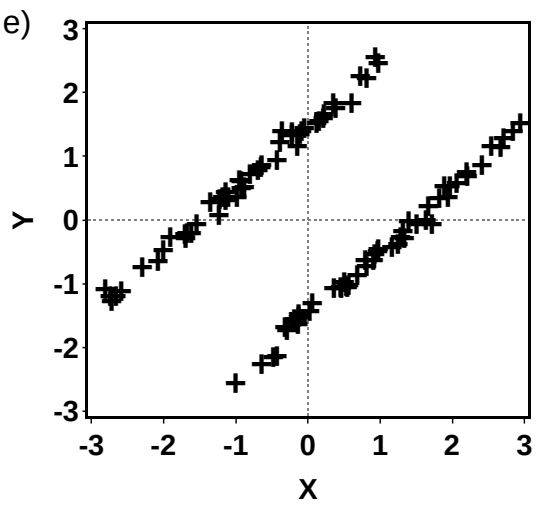
Figure 1:


a)
3
2
1
0
-1
-2
-3
Y
-3 -2 -1 0 1 2 3
X


b)
3
2
1
0
-1
-2
-3
Y
-3 -2 -1 0 1 2 3
X


c)
3
2
1
0
-1
-2
-3
Y
-3 -2 -1 0 1 2 3
X


d)
3
2
1
0
-1
-2
-3
Y
-3 -2 -1 0 1 2 3
X


e)
3
2
1
0
-1
-2
-3
Y
-3 -2 -1 0 1 2 3
X

Reference List

(1) Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol 2011 Mar;32(3):437-40.

(2) Chen H, Cohen P, Chen S. Biased odds ratios from dichotomization of age. Stat Med 2007 Aug 15;26(18):3487-97.

(3) Anscombe FJ. Graphs in Statistical Analysis. American Statistician 1973;27(1):17-21.

(4) Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006 Jan 15;25(1):127-41.