February 2004: Dichotomize Continuous Variables for Odds Ratio Analysis on the Basis of the Distribution of the Data, the Variance of the Odds Ratio, and the Odds Ratio—in this Order. (New Rule 4.13.)

Introduction

While dichotomization typically leads to loss of information (see Rule 4.11) there are occasions when dichotomizing a continuous variable makes the data more interpretable. The question then comes up, how should the data be dichotomized. Some surprising results crop up as will be shown below.

Rule of Thumb

Dichotomize continuous variables for odds ratio analysis on the basis of the distribution of the data, the variance of the odds ratio, and the odds ratio—in this order.

Illustration

Suppose data sets come from two logistic distributions that differ in location shift. The data are to be dichotomized and the odds ratio calculated. What is the optimum split? Surprisingly, every split produces the same odds ratio. However, the variance changes dramatically and, hence, the significance of the odds ratio.

Specifically, let X come from a logistic distribution with location parameter μ and Y from a logistic distribution with location parameter $\mu+\delta$. Suppose the cut point is at $X=x$. Then the logarithm of the odds ratio is δ *for all values of X.* That is, no matter where the cut point is chosen, the log odds is the same. This follows directly from the distributions. However, the variance of the log odds depends very much on the cut point and hence the significance of the log odds.

For example, suppose that 50 observations each are taken from two logistic distributions with means $\mu=0$ and $\delta=0.5$. Figure 1 shows the log odds ratio, ln(*O*), for a range of cut points ranging from -4 to +4. Superimposed on the log odds ratio is its standard error. The standard error is derived from the usual estimate of the variance,

$$\text{var}[\ln(O)] = \frac{1}{n}\left[\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right].$$

The quantities π_{ij} are the proportions that fall into each of the two categories below the cut point (j=0), above the cut point (j=1), for samples i=1, 2 so that $\pi_{11}+\pi_{12}=1$ and $\pi_{21}+\pi_{22}=1$.

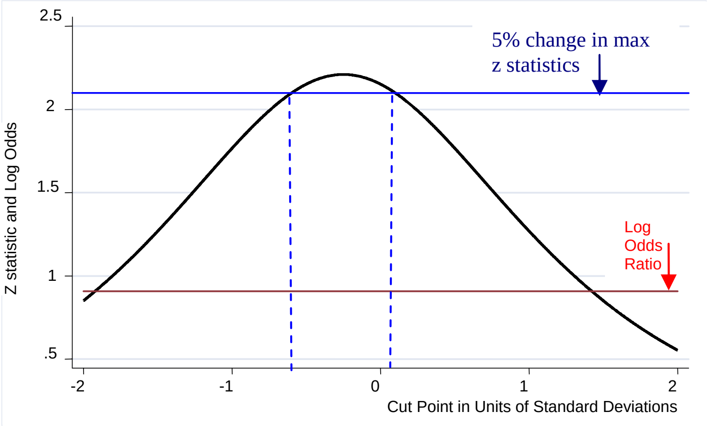

Figure 1. Two samples of 50 observations drawn from logistic distributions with means 0 and 0.5, and standard deviation 1. Cut points range from -2 to 2. The *z*-statistic is the log of the odds ratio divided by its standard error.

Figure 1 indicates that the significance of the odds ratio depends on the cut point whereas the odds ratio is unchanged over the whole range of the cut points. The figure also indicates that there is a rather narrow range of cut points over which the significance of the odds ratio changes less than 5%. This range is roughly from -0.6 to 0.1 standard deviations. This is a rather narrow range. We explore this approach in the discussion with other distributions.

Basis of the rule

As indicated for data from a logistic distribution the variance changes with the cut point but not the odds ratio. The importance of the shape of the distribution will be explored below.

Discussion and Extensions

In this discussion we discuss two additional scenarios, first a normal distribution and then a skewed distribution, the lognormal.

Consider the case of two normal distributions with means 0 and 0.5 and equal standard deviations. We assume again that 50 observations are drawn and cut points ranging from -2 to +2. Figure 2 presents the picture for this situation. The pattern is very similar to that of Figure 1 with the exception that the odds ratio actually increases with more extreme cut points whereas the significance of the odds ratio decreases. The top line at

Y=2.5 is the value of the z statistic for comparing two samples of size 50 from two normal distributions with means 0, 0.5 and standard deviations 1. This represents the precision of the sample data when the observations are not dichotomized. The loss in precision can be sorted into two components then: the loss due to dichotomization and the loss due to choice of cut point. Equivalently this could be described in terms of additional sample sizes needed due to dichotomization and due to cut point.

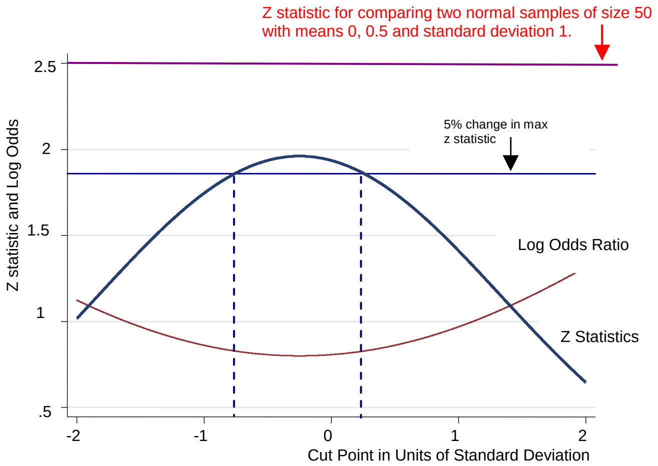

Figure 2. Two samples of 50 observations drawn from logistic distributions with means 0 and 0.5, and standard deviation 1. Cut points range from -2 to 2. The *z*-statistic is the log of the odds ratio divided by its standard error.

The second situation considered is where the distribution is skewed. For illustration we take the two normal distributions of Figure 2 and exponentiate them. This produces lognormal distributions with means and variances as follows:

Table 1. Lognormal sampling situation. The pooled standard deviation is the square root of the average variance.

Sample	Parameters	Normal	Lognormal
1	Mean	0	1.65
	S.D.	1	2.16
2	Mean	0.5	2.72
	S.D.	1	3.56
Pooled	S.D.	1	2.94

To make comparisons with the previous situations we cover a range of four standard deviations in the lognormal scale. A standard deviation is defined as the square root of the average variance, $2.94 = \sqrt{1/2(2.16^2 + 3.56^2)}$. Figure 2 displays the characteristics of this configuration. The interval with less than 5% reduction in the maximum Z statistic for the logarithm of the odds ratio is now considerably narrower than in the first two examples. The actual values are 0.26 and 0.72 standard deviations—a rather narrow range. The other interesting feature in Figure 3 is that in a small range of partitioning the z statistic for the odds ratio is approximately equal to the z statistic for the two sample test with heterogeneous variances. This occurs at about 0.44 standard deviations.

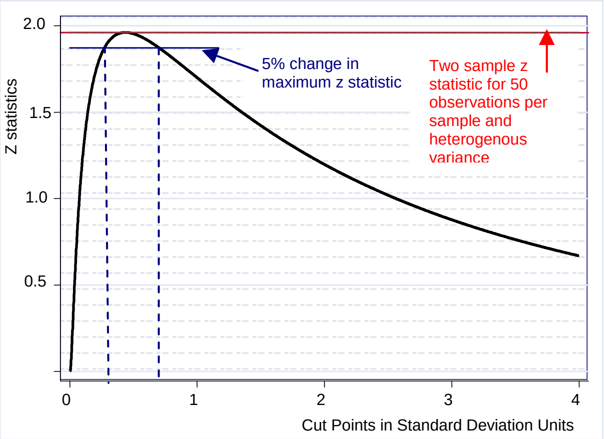

Figure 3. Two samples drawn from lognormal distributions as described in Table 1 with standard deviation 2.94.

The conclusions are that care must be taken in dichotomizing data in order to calculate odds ratios. The shape of the distribution, it variance, and the cut point must be taken into account. It appears from these analyses that the shape of the distribution is the most critical component.

This work could be extended by considering an ordinal partitioning of the data, not just a dichotomy. The ordered categories could be analyzed by a non-parametric test such as the Wilcoxon rank sum test. The optimal partitioning strategy could be investigated. Most likely there will be less loss of information in this case and also greater robustness in the cut points.