PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2009; **8**: 50–61
Published online 3 April 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.331

Consequences of dichotomization

MAIN PAPER

Valerii Fedorov[1], Frank Mannino[1,*,†] and Rongmei Zhang[1,2]
[1] *Research Statistics Unit, Biomedical Data Sciences, GlaxoSmithKline Pharmaceuticals, Collegeville, PA, USA*
[2] *Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

Dichotomization is the transformation of a continuous outcome (response) to a binary outcome. This approach, while somewhat common, is harmful from the viewpoint of statistical estimation and hypothesis testing. We show that this leads to loss of information, which can be large. For normally distributed data, this loss in terms of Fisher's information is at least $1 - 2/\pi$ ***(or 36%). In other words, 100 continuous observations are statistically equivalent to 158 dichotomized observations. The amount of information lost depends greatly on the prior choice of cut points, with the optimal cut point depending upon the unknown parameters. The loss of information leads to loss of power or conversely a sample size increase to maintain power. Only in certain cases, for instance, in estimating a value of the cumulative distribution function and when the assumed model is very different from the true model, can the use of dichotomized outcomes be considered a reasonable approach. Copyright © 2008 John Wiley & Sons, Ltd.***

Keywords: *dichotomization; categorization; grouping*

1. INTRODUCTION

We consider the common situation in biomedical statistics where continuous outcomes are observed, but inference and estimation are done using a dichotomized version of the outcomes. The primary argument in favor of dichotomization is the ease and simplicity of reporting results (e.g. [1–3]. The problem with these arguments is the unnecessary confounding of analysis and reporting. We do not deny that in certain cases dichotomized reports can be simpler for non-statisticians to understand. However, very rarely, as we will later discuss, dichotomization leads to better statistical properties.

The term dichotomization can also refer to other practices, such as the dichotomization of covariates in analysis (e.g. [4]) or dichotomization for the purpose of reporting results. Additionally, this can refer to the use of a continuous background (latent) models to generate binary multivariate models with mutually dependent components (e.g. multivariate probit, see for instance, [5–7]). None of these situations will be discussed here as we will focus solely on the dichotomization of outcomes for statistical analysis.

*Correspondence to: Frank Mannino, Research Statistics Unit, Biomedical Data Sciences, GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, PA 19426, USA.
† E-mail: frank.v.mannino@gsk.com

The dichotomization can be viewed as a particular case of grouping (e.g. [5, 8–10]). For instance, more general grouping can be defined as $Z = i - 1$ if $Y \in \Omega_i$, where $\cup_{i=1}^{n} \Omega_i$ constitutes the whole support set for Y and all Ω_i's are disjoint. This polychotomization, is also referred to in the literature as responder analysis (e.g. [11, 12]), where patients are broken into categories based on whether their continuous outcome meet some thresholds of response.

Similar to many other publications (e.g. 13–15]), the major recommendation of this paper is to avoid dichotomization whenever possible on the data analysis stage, but the technique can be a useful tool for the reporting of final results to non-statistical communities. Some of the results presented here are not new and have been known for quite some time (e.g. [16]). However, we hope that systematic presentation combined with some of our own findings will lead to a better understanding of the pros and cons of dichotomization.

2. NORMAL CASE

2.1. Model

We first consider a normally distributed random variable Y with $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2$, i.e. with cumulative Gaussian distribution function $F(y|\mu, \sigma^2)$. Let

$$Z = \begin{cases} 0, & Y \leqslant c \\ 1, & Y > c \end{cases} \tag{1}$$

where c is a predefined cut point. We will assume here that c is always known. Random variable Z has a Bernoulli distribution with $p = F(c|\mu, \sigma^2)$. The likelihood function for a single observation is

$$\begin{aligned} L(\mu|Z) &= F^Z(c|\mu, \sigma)[1 - F(c|\mu, \sigma)]^{1-Z} \\ &= \Phi^Z(u)[1 - \Phi(u)]^{1-Z} \end{aligned}$$

where $u = (c - \mu)/\sigma$ and $\Phi(u)$ is the standardized Gaussian distribution function. Observing that $\partial \Phi/\partial\mu = -(1/\sigma)\partial\Phi/\partial u$, one can derive that the Fisher information about μ of a single dichotomized observation is

$$I_d(u) = \text{var}\left(\frac{\partial}{\partial\mu} \ln L(\mu|Z)\right) = \frac{1}{\sigma^2} \frac{\phi^2(u)}{\Phi(u)[1 - \Phi(u)]} \tag{2}$$

where $\phi(u) = \partial\Phi(u)/\partial u$. Unless otherwise specified, we assume that σ^2 is known and we wish to estimate μ only. Actually, this is a necessity because we cannot estimate both σ^2 and μ using the dichotomized data. When σ is known, the estimation of u is equivalent to the estimation of μ. The information about μ from a single observation of Y is $I_o = 1/\sigma^2$. We are interested in the relative efficiency of the dichotomized analysis, defined as $R = I_d/I_o$. For the normally distributed data, this becomes

$$R(u) = \frac{\phi^2(u)}{\Phi(u)[1 - \Phi(u)]}$$

This can alternatively be expressed in terms of the percentage of information lost, $100[1 - R(u)]$. Additionally, the inverse of $R(u)$ is the factor by which the number of observations (sample size) must be increased to mitigate the impact of dichotomization. For n independent observations, the maximum likelihood estimators (MLEs) of μ are

$$\hat{\mu}_o = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \hat{\mu}_d = c - \Phi^{-1}\left(\frac{n_1}{n}\right)\sigma \tag{3}$$

for the continuous and dichotomized cases, respectively. Here, n_1 is the total number of observations Y below c. Recall that σ is known.

Many sources have referenced the minimal information lost when dichotomizing a normal distribution at the mean (median) cut point, which equals $(1 - 2/\pi) = 36.33\%$ (e.g. [16]). However, the choice of cut point must be made prior to analysis when the mean is unknown. This use of the mean as the cut point is the least damaging option, while all other cut points lead to a greater loss of information (see Figure 1, top left plot). This fact is often overlooked when considering the effect of dichotomization. Even if we increase the sample size to compensate for the 36% loss of information, the study is still underpowered unless the selection of $c = \mu$ is precise.

The sample size necessary to match the precision of the dichotomized estimator to that of the

DOI: 10.1002/pst

Pharmaceutical STATISTICS

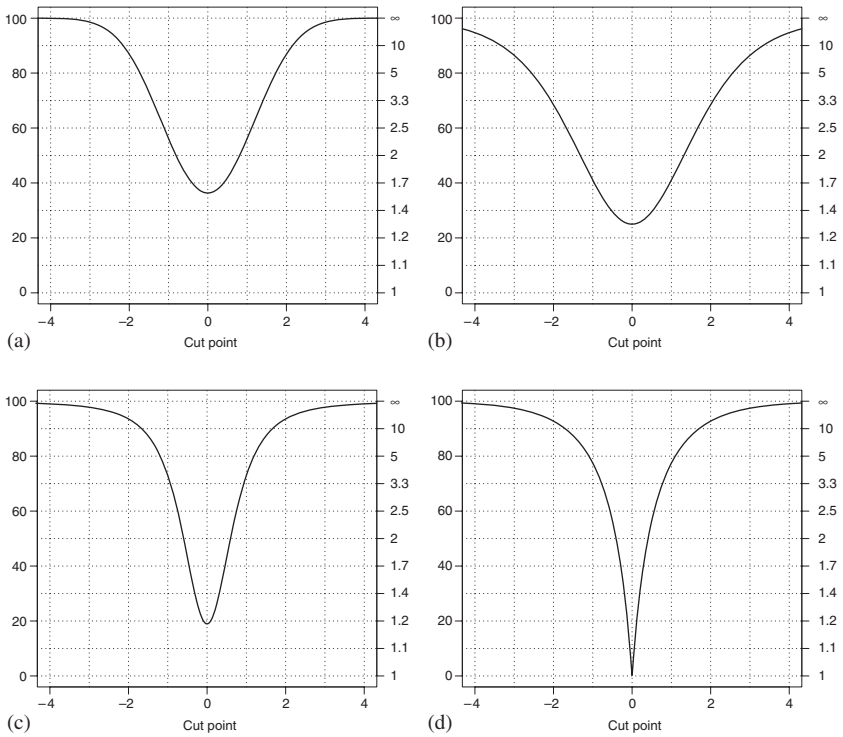

Figure 1. Percentage of information lost when dichotomizing, based on various cut points. The plots correspond to normal (a), logistic (b), Cauchy (c), and double exponential (d) distributions. The right vertical axes correspond to the factor of sample size increase needed to mitigate the loss.

continuous estimator can be seen in the right axis on the top left plot in Figure 1. The best-case scenario ($c = \mu$) would require a sample size $\pi/2 = 1.571$ times larger to compensate for dichotomization.

We can also consider the case where the cut point is not preselected, but the data are dichotomized into two groups of equal size. The estimate of the μ in this case is simply the median of the data, which has a variance of $\pi\sigma^2/(2n)$ [17 (Chapter 13)]. This leads to an asymptotic relative efficiency of $2/\pi$ compared with the mean estimator. *A posteriori* choice of a cut point is essentially equivalent to the ideal prior choice of a cut point, which leads to a 36% loss of information. While this eliminates the risk of greater information loss, the empirical cut point lacks any real-life implications and is not even useful for simple reporting of results. In addition, as Altman [13] discusses, these data-driven choice of cut points does not allow for comparisons of different studies.

2.2. Regression and probit model

Let $E(Y|x) = \mu(x,\theta) = \theta^{\mathrm{T}} f(x)$ and $\mathrm{var}(Y|x) = \sigma^2$, where vector function $f(x)$ is given, x is a covariate (for instance, dose), and θ are unknown parameters. The information matrix for a single observation under the dichotomized model is the direct generalization of (2) (see, for example, [18, 19]),

$$\mathbf{I}_d = \frac{1}{\sigma^2}\frac{\phi^2(u(x,\theta))}{\Phi(u(x,\theta))[1-\Phi(u(x,\theta))]} f(x) f^{\mathrm{T}}(x)$$
$$= \frac{1}{\sigma^2}\omega(x,\theta) f(x) f^{\mathrm{T}}(x)$$

where $u(x,\theta) = (c - \mu(x,\theta))/\sigma$ and $\omega(x,\theta) = \phi^2(u(x,\theta))/[\Phi(u(x,\theta))[1-\Phi(u(x,\theta))]]$. For continuous

DOI: 10.1002/pst

outcomes, assuming σ is known, the information matrix is

$$\mathbf{I}_o = \frac{1}{\sigma^2} f(x) f^{\mathrm{T}}(x)$$

The total Fisher information matrices for a given design $\xi_n = \{r_i, x_i\}^n$ are

$$\mathbf{I}_d(\xi_n) = \sum_{i=1}^{n} \frac{r_i \omega(x_i, \theta)}{\sigma^2} f(x_i) f^{\mathrm{T}}(x_i) \quad \text{and}$$

$$\mathbf{I}_o(\xi_n) = \sum_{i=1}^{n} \frac{r_i}{\sigma^2} f(x_i) f^{\mathrm{T}}(x_i)$$

where x_i are the design points and r_i are the weights of these design points. Thus, there is no problem in comparison with these two matrices. From the previous section, note that for all x_i, $\omega(x_i, \theta) \leqslant 2/\pi$ or $\mathbf{I}_d \leqslant 2\mathbf{I}_o/\pi$.

For any specific design, we may derive more precise results. For instance, it is known (see, for example, [19]) that for the simplest probit models where $f^{\mathrm{T}}(x) = (1, x)$ and a design region including points $|x - \mu|/\sigma = 1.138$, the locally D-optimal design is

$$\xi_2^* = \{r, x_i\}_{i=1}^2, \quad x_{1,2} = \mu \pm 1.138\sigma$$

For this design,

$$\mathbf{I}_d(\xi_2^*) = \omega^* \mathbf{I}_o(\xi_2^*)$$

where $\omega^* = 0.392$, requiring a 255% increase in sample size to compensate for the loss of information due to dichotomization.

2.3. Cumulative distribution function estimation

Occasionally, we wish to estimate a certain value of the distribution function

$$p = \mathrm{Prob}(Y \leqslant c) = \Phi\left(\frac{c - \mu}{\sigma}\right)$$

rather than the mean. One may choose to directly estimate p by counting the proportion of observations n_1 that are less than c and dividing by the total number of observations, which is the MLE for estimating p in the standard binary setting. Although intuitively this may seem like a good estimator, we are still dichotomizing the data and losing the same amount of information compared with the continuous estimator. The pair of estimators are defined as

$$\hat{p}_o = \Phi(\hat{u}) \quad \text{and} \quad \hat{p}_d = \frac{n_1}{n} \tag{4}$$

where $\hat{u} = (c - \hat{\mu})/\sigma$, under *the assumption that the normal model is correct*. The information loss as a function of c is identical to that seen when estimating μ in Figure 1(a). We must also note that when estimating very small or large cumulative distribution function (CDF) values, a finite sample size gives a significant chance of all outcomes belonging to one group, forcing the estimate of p to 0 or 1, i.e. making the estimation problem singular.

2.4. Link to hypothesis testing

We will consider a one-sided hypothesis test and look at the effect of dichotomization focusing on the loss of power. Recall that for a one-sided test,

$$z_{1-\alpha} + z_{1-\beta} = \frac{|\delta|}{\sqrt{\mathrm{var}(\hat{\delta})}}$$

where δ is the true effect we wish to test, α is the Type I error rate and β is the Type II error rate. For a given value of α, the power $(1 - \beta)$ is uniquely defined by the variance of the estimator. With a single arm and a test about the mean μ, our hypotheses are $H_o : \mu = \mu_1$ and $H_1 : \mu \geqslant \mu_1 + \delta$. The variances of estimators (3) are given by

$$\mathrm{var}(\hat{\mu}_o) = \frac{\sigma^2}{n} \quad \text{and} \quad \mathrm{var}(\hat{\mu}_d) \approx \frac{\sigma^2 \Phi(u)[1 - \Phi(u)]}{n\phi^2(u)}$$

for continuous and dichotomized estimators, respectively (see (2) for more comments). The first plot in Figure 2 shows the power for various cut points with a test based on estimators (3), with the continuous normal approximation used in the dichotomized case. We have chosen $n = 100$, $\alpha = 0.05$, $\sigma^2 = 1$, and $\delta = 0.29$ selected to obtain 90% power in the continuous case. The variance of $\hat{\mu}_o$ and therefore the power are independent of c.

Our hypothesis test concerning p can be expressed as $H_o : p = p_1$ and $H_1 : p \geqslant p_1 + \delta$. We focus on these hypotheses, acknowledging that there are other ways to formulate a test

DOI: 10.1002/pst

Pharmaceutical STATISTICS

about the probability p. The variances of estimators (4) are

$$\mathrm{var}(\hat{p}_o) \approx \frac{\phi^2(u)}{n} \quad \text{and} \quad \mathrm{var}(\hat{p}_d) = \frac{\Phi(u)[1-\Phi(u)]}{n}$$

The second plot in Figure 2 shows the power when $n = 100$, $\alpha = 0.05$, and $\delta = 0.1$ under the continuous (solid line) and dichotomized (dashed line) cases, using a normal approximation. Unlike a test about μ, the power under the continuous hypotheses depends upon the true value of p.

These methods can also be extended to the case of hypothesis tests with two arms. Ragland [20] discusses the effect on power when estimating a prevalence ratio p_1/p_2 or an odds ratio $p_1 \times (1-p_2)/(p_2(1-p_1))$.

Alternatively, for a given sample size and treatment effect, we can compare the possible combination of Type I and Type II error rates for the continuous and dichotomized outcomes. Figure 3 shows these values with a sample size of 100 for a test of μ and for a test of p. We can see that the continuous response (solid line) yields the best possible choices of α and β. The dashed lines represent different choices of the cut point for the dichotomized test. As we get farther away from the

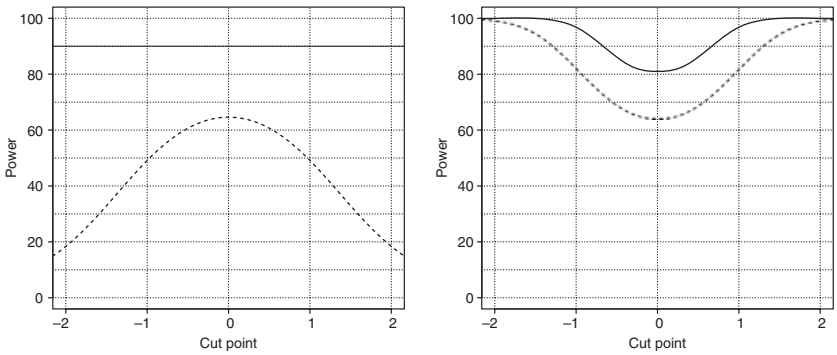

Figure 2. The plots show the power for a single arm hypothesis test about $\hat{\mu}$ (left plot) and p (right plot). The solid lines represent the continuous case and the dashed lines represent the dichotomized case. Note that when testing hypotheses about p, the power depends on the true value of p in both cases.

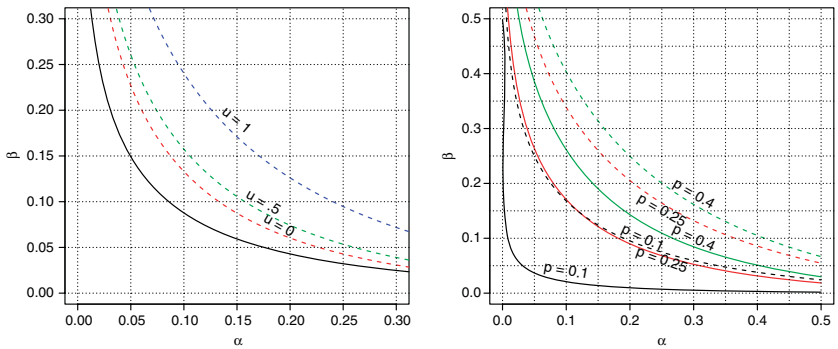

Figure 3. The Type I and Type II errors for a sample size of 100. The left plot shows a solid line for the hypothesis test about μ performed using the continuous data and dashed lines represent the dichotomized hypothesis test for several cut points. The plot on the right shows a hypothesis test about the CDF. The solid lines represent the continuous outcomes and the dashed lines represent the dichotomized outcomes for different true values of p.

DOI: 10.1002/pst

optimal cut point ($u = 0$) in either direction (remember that the effect of the cut point is symmetric), the possible choices of α and β get progressively worse. For any reasonable choice of α, the power lost through dichotomization is over 10% at the optimal cut point and over 20% for less ideal cut points (e.g. $u = 1.0$). For a test of μ, the sample sizes needed to match the power under the continuous outcomes for the cut points of 0, 0.5, and 1 are 158, 173, and 228, respectively. For a test of the CDF value, the choices of α and β even in the continuous case depend upon what CDF value we wish to estimate, but the same conclusions apply.

2.5. Unknown σ

We can also extend our analysis to the more realistic scenario when σ is unknown. In the dichotomized case, only $\Phi^{-1}((c-\mu)/\sigma)$ can be estimated directly. Consequently, μ and σ cannot be estimated separately. The continuous outcomes do not face this same problem. When σ is unknown, we focus on the estimation of $\hat{p}_o$ and compare it with $\hat{p}_d$ which requires no knowledge of σ. We now estimate $\hat{p}_o$ by plugging in estimators $\hat{\mu}=\sum_{i=1}^{n} y_i/n$ and $\hat{\sigma}=\sqrt{\sum_{i=1}^{n}(y_i-\hat{\mu})^2/(n-1)}$, i.e.

$$\hat{p}_o = \Phi\left(\frac{c-\hat{\mu}}{\hat{\sigma}}\right)$$

Noting that estimators $\hat{\mu}$ and $\hat{\sigma}$ are uncorrelated, their variances are σ^2/n and $2\sigma^2/n$, and using Taylor's expansion at the vicinity of the true values μ and σ, one can verify (see Appendix A) that the relative efficiency is approximately

$$R(u) \approx \frac{\phi^2(u)}{\Phi(u)[1-\Phi(u)]}\left[1+\frac{1}{2}u^2\right] \qquad (5)$$

This leads to a bimodal loss of information, relative to the cut point as seen in Figure 4. The minimum information lost is now slightly less than when σ is known, but always at least 33%. Even though we are estimating two parameters to calculate $\hat{p}_o$, it is still far superior to the dichotomized estimator.

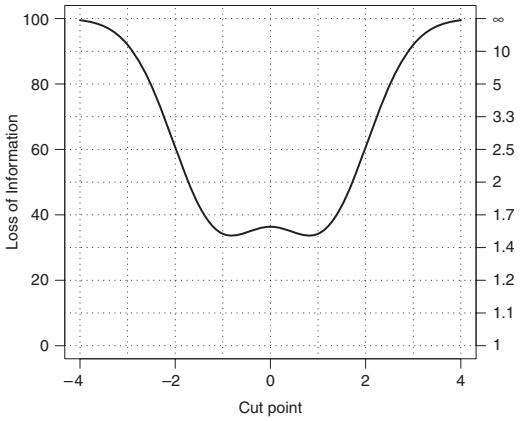

Figure 4. The loss of information under normally distributed data when estimating a CDF value p if σ is unknown.

2.6. Model misspecification

An argument can be made that using $\hat{p}_d = n_1/n$ may be beneficial for estimating p as it does not assume any distribution for Y, while $\hat{p}_o$ requires the assumption of a normal distribution. We have simulated patient observations under logistic and double exponential models (both with a standard deviation of 1) and then estimated $\hat{p}_d$ using dichotomized data and $\hat{p}_o = \Phi((c-\hat{\mu})/\hat{\sigma})$, i.e. wrongly assuming a normal distribution for continuous data. The values of $\hat{\mu}$ and $\hat{\sigma}$ are estimated using the MLEs under the (incorrect) assumption of a normal distribution. The variance of $\hat{p}_o$ is less than the variance of $\hat{p}_d$ but for most values of c, $\hat{p}_o$ is a biased estimator as the cumulative distribution functions are different for normal, logistic, and double exponential; hence, we will compare these values using the root mean-squared errors (RMSEs). At values of c when the distribution functions of the normal and logistic models are close (e.g. around μ) or when the sample size n is small, $\hat{p}_o$ outperforms $\hat{p}_d$. Figure 5 shows the RMSEs when the true model is logistic (left side) and double exponential (right side) for three different sample sizes. For the logistic model, we see that when $n = 20$, the continuous estimator is always superior. When $n = 70$, the two estimators are almost identical for certain values and when

DOI: 10.1002/pst

Pharmaceutical STATISTICS

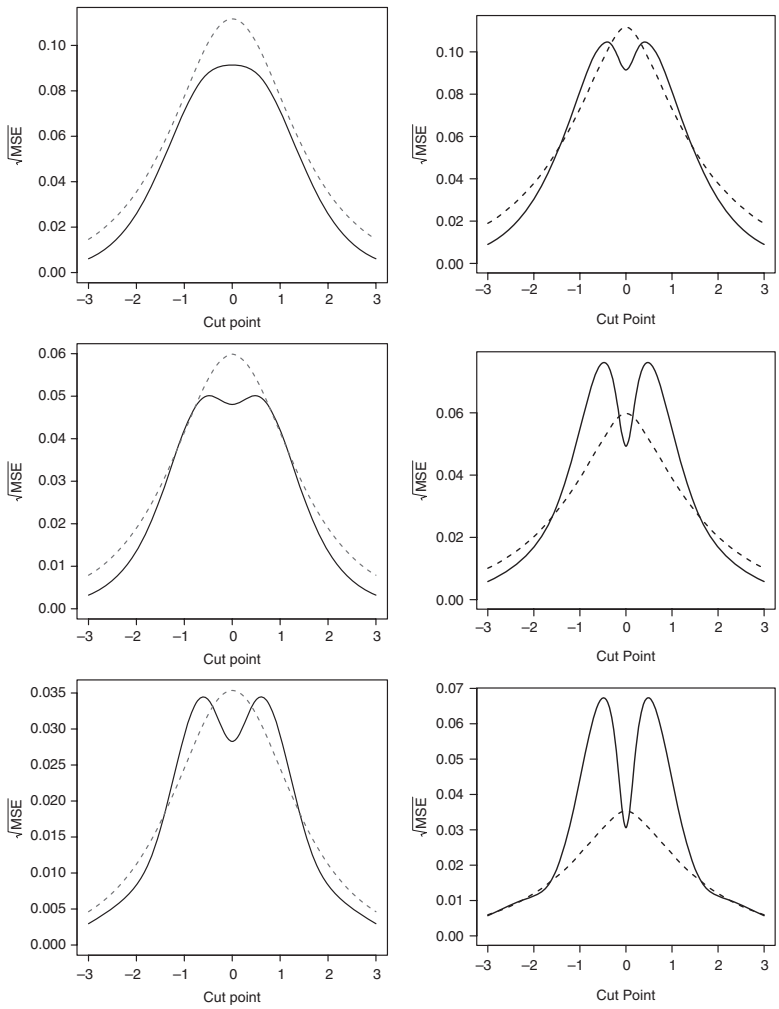

Figure 5. Comparison of estimators between dichotomized (dashed lines) and continuous (solid lines) normal models when the true distribution is logistic (left side) or exponential (right side). The plots show the root mean-squared errors as a function of the cut point for sample sizes of 20, 70, and 200.

$n = 200$ there are regions where $\hat{p}_d$ is superior to $\hat{p}_o$ and other regions where the opposite is true. Despite the wrong model being used in estimation, the continuous estimator is better than the dichotomized estimator over a large range of cut points. Thus, dichotomization can be recommended as a robust approach in the CDF estimation for large (several hundred) sample sizes. For the true double exponential model, the dichotomized estimator often performs better here due to the larger difference between the true and assumed models. This holds even for small sample sizes ($n = 20$). However, the double exponential distribution is an extreme case and this will very rarely be close to the true underlying model.

We also consider the scenario when the true distribution is a mixture of two normal distributions. Again we wish to estimate p when the true model is

$$Y \sim 0.25N(0,1) + 0.75N(4,4)$$

In Figure 6, we show the RMSE for the analysis under the incorrect normal model (solid lines), the

DOI: 10.1002/pst

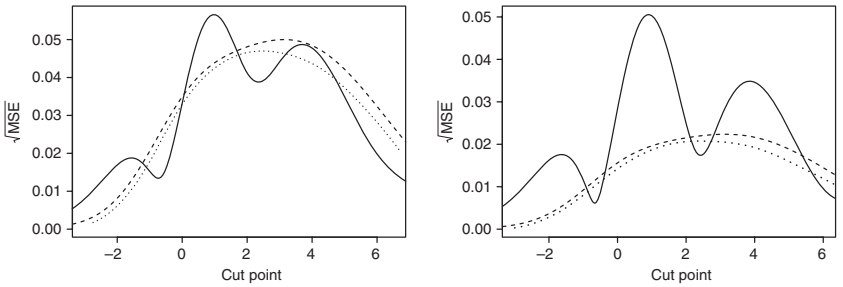

Figure 6. Comparison of estimators between dichotomized (dashed lines), continuous normal (solid lines), and normal mixture (dotted lines) models when the true distribution is a normal mixture. The plots show the root mean-squared errors as a function of the cut point for sample sizes of 100 and 500.

Table I. Various location-scale family distributions and the information under dichotomized data (see [21 (p. 121)] for further details).

Distribution	Density function	Var(Y)	Maximum $\frac{I_d}{I_o}$	I_o
Normal	$\frac{1}{\sqrt{2\pi\sigma_u^2}}e^{-u^2/2}$	σ^2	$\frac{2}{\pi}=0.637$	$\frac{1}{\sigma^2}$
Logistic	$\frac{e^{-u}}{\sigma(1+e^{-u})^2}$	$\frac{\pi^2\sigma^2}{3}$	0.75	$\frac{1}{3\sigma^2}$
Double exponential	$\frac{1}{2\sigma}e^{-\lvert u\rvert}$	$2\sigma^2$	1.0	$\frac{1}{\sigma^2}$
Cauchy	$\frac{1}{\pi\sigma}\frac{1}{1+u^2}$	does not exist	$\frac{8}{\pi^2}=0.811$	$\frac{1}{2\sigma^2}$

dichotomized outcomes (dashed lines), and analysis under the true mixture model (dotted lines) for sample sizes of 100 and 500. Using the true mixture distribution in analysis is the best practice. However, if we do not know that our data follow this distribution, then we will naturally choose from either a normal model or the dichotomized model. For a small sample size, the dichotomized estimator provides little improvement over the incorrect normal distribution. However, at large sample sizes the use of the dichotomous outcomes is superior.

3. OTHER DISTRIBUTIONS

3.1. Location-scale family

The location-scale family (see [21 (pp. 20–21)] consists of distributions with densities given by the function

$$f(y|\mu,b)=\frac{1}{b}\phi(u)$$

with $u=(y-\mu)/b$. The most popular of them are presented in Table I. The logistic and normal distributions provide the logit and probit transformations, respectively, which are quite popular in biostatistics. Figure 1 contains the plots of information lost for several members of the family (see Table I). For all members of this family, $R\leqslant 1$ for all u (see Appendix B for proof). This inequality holds for any distribution and follows from the fact that $\{z_i\}_1^n$ is not in general a sufficient statistic for estimating μ. The information based on any function of the original data X will be less than or equal to the information derived from X. Moreover, the values of the information will only be equal when the function

DOI: 10.1002/pst

Pharmaceutical STATISTICS

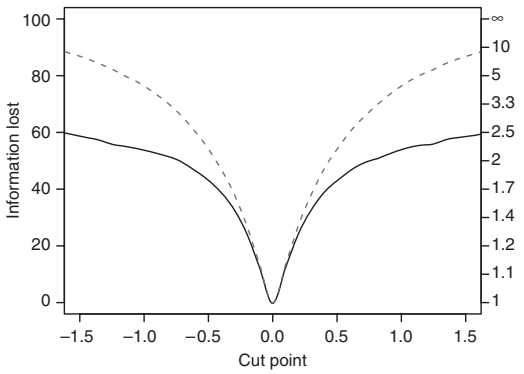

Figure 7. The information lost from dichotomizing a double exponential distribution for the estimation of a CDF value. The solid line represents the case when we estimate both μ and σ and the dashed line represents the case when σ is known. For the case of unknown σ, the values were estimated through simulations, using the MLEs $\hat{\mu}_o = \text{median}(Y)$ and $\hat{\sigma} = n^{-1}\sum_{i=1}^{n}|y_i - \hat{\mu}_o|$ [23 (p. 172)].

of X is sufficient. The fact that the dichotomized data are not sufficient is further proof that $I_d < I_o$ [22, p. 244].

The double exponential distribution presents an interesting case. The maximum likelihood estimator, $\hat{\mu}$, is the median of Y [23 (Chapter 24)]. If the cut point is equal to μ, there is no information lost when dichotomizing (see Figure 1(d)). The information lost increases very sharply as $|c - \mu|$ increases, and as μ is not known, the dichotomization will likely be very damaging. The same effect is seen when estimating p from a double exponential distribution (see Figure 7).

3.2. Discrete distributions

We may also want to consider the loss of information when dichotomizing a discrete distribution. Consider the Poisson distribution

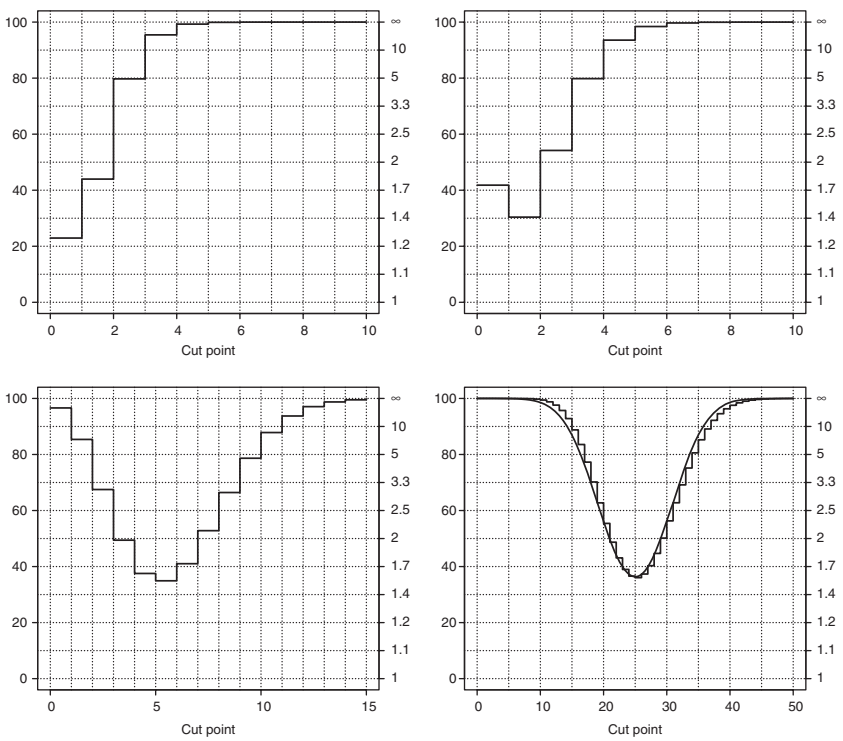

Figure 8. Information lost when dichotomizing a Poisson distribution. The plots correspond to different values of λ (0.5, 1.0, 5.0, 25.0). The last plot also shows the information lost using the normal approximation to a Poisson distribution.

DOI: 10.1002/pst

defined as $P(Y = y|\lambda) = e^{-\lambda}\lambda^y/y!$. We defined the dichotomized variable Z as in the normal case. Our possible cut points will be restricted to integers, as these are the only possible values the Poisson distribution can take. Here, the relative efficiency of the dichotomized observations is

$$\frac{I_d}{I_o} = \frac{\lambda\left(\sum_{i=0}^{c}(1/i!)\lambda^i(i/\lambda - 1)\right)^2}{\left(\sum_{i=0}^{c}(1/i!)\lambda^i\right)\left(\sum_{i=c+1}^{\infty}(1/i!)\lambda^i\right)}$$

This is plotted for four different values of λ (0.5, 1.0, 5.0, 25.0) in Figure 8. While the minimum value of information loss varies, depending upon the true value of λ, we see that it is always a significant amount. As λ gets large, the Poisson distribution can be approximated by the normal distribution and the earlier results can be used to look at the relative efficiency of dichotomizing the Poisson distribution. This approximation is shown with the information lost under the normal distribution for the fourth plot in the figure.

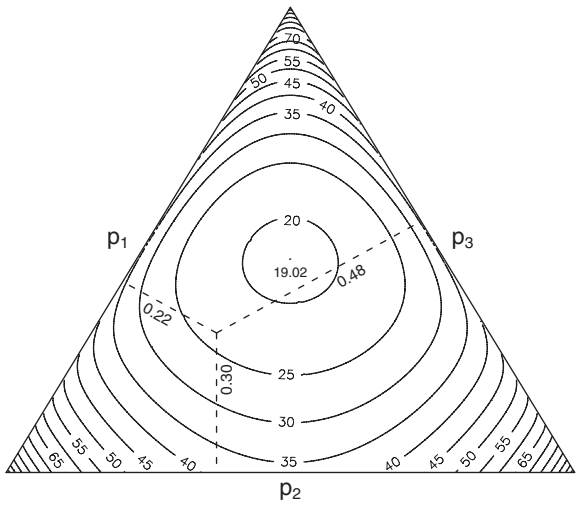

Figure 9. Information lost when trichotomizing a normal distribution. Two cut points, c_1 and c_2, must be chosen, leading to three areas of probability, $p_1 = \Phi(u_1)$, $p_2 = \Phi(u_2) - \Phi(u_1)$, $p_3 = 1 - \Phi(u_2)$ that correspond to the length of the line from the each edge to the point. The example point shows $c_1 = -0.772$ and $c_2 = 0.0504$. This leads to a relative efficiency of 0.7642.

4. DISCRETIZATION INTO MULTIPLE GROUPS

The reduction of information from dichotomization can also be extended to several classes. Using the normal distribution as an example, we can trichotomize our observation by choosing two cut points, c_1 and c_2. If chosen properly, this will lead to a better estimator than dichotomizing, with only 19.0% of information lost. However, suboptimal choices of cut points will quickly lead to very poor estimators. The plot in Figure 9 shows the isolines of the information lost for various cut point values. The optimal choice is $u_1 = (c_1 - \mu)/\sigma = -0.612$ and $u_2 = -u_1$, leading to regions of probability of 0.27, 0.46, and 0.27. The knowledge of optimal cut points does not help too much on the data analysis stage, as we do not know μ prior to the experiment. However, this knowledge may lead to better reporting of results to laymen.

As we increase the number of classes, the amount of information lost with the best cut points decreases [16]. However, the often-argued benefit of dichotomizing lies in the simplicity of results. Discretizing data into four or more groups, while not being as damaging if the cut points are well chosen, has less benefit of simplicity.

5. CONCLUSIONS

The knowledge of losing information from dichotomizing a continuous outcome is nothing new. However, many previous writings report on the optimal choice of cut points, which depends upon the parameters we wish to estimate. If we are lucky, the chosen cut point is near the optimal point, but the consequences of dichotomizing become more dire as we deviate from the optimal point. We focus our study on the evaluation of losses caused by dichotomization given cut points. While the analysis of dichotomized outcomes may be easier, there are no benefits to this approach when the true outcomes can be observed and the 'working' model is flexible enough to describe the

DOI: 10.1002/pst

Pharmaceutical STATISTICS

population at hand. Thus, dichotomization should be avoided in most cases. Only when we wish to estimate a CDF value, our working model poorly approximates reality, and our sample size is large will the biasedness of model-based estimators overpower the improvement in variance. In this case, the dichotomized estimator may lead to better results, but further study-specific consideration is needed. We also want to emphasize that while analysis should be done using actual outcomes, some aspects of this analysis can be reported on a dichotomized scale.

ACKNOWLEDGEMENTS

We would like to thank Frank Rockhold, Stephen Senn, and Steve Snapinn for their useful comments and suggestions.

REFERENCES

1. Farrington DP, Loeber R. Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behavior and Mental Health* 2000; **10**:100–122.
2. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
3. Lewis JA. In defence of dichotomy. *Pharmaceutical Statistics* 2004; **3**:77–79.
4. Christ M, Laule K, Klima T, Hochholzer W, Breidthardt T, Perruchoud AP, Mueller C. Multimarker strategy for risk prediction in patients presenting with acute dyspnea to the emergency department. *International Journal of Cardiology* 2007; DOI: 10.1016/j.ijcard.2007.03.119.
5. Pearson K. On the systematic fitting of curves to observations and measurements. *Biometrika* 1902; **1**:265–303.
6. Lesaffre E, Molenberghs G. Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine* 1991; **10**:1391–1403.
7. Fedorov V, Wu Y. Generalized probit model in design of dose finding experiments. In *mODa 8 – Advances in model-oriented design and analysis*, Lopez-Fidalgo J, Rodriguez-Diaz JM, Torsney B (eds). Physica-Verlag: Wurzburg, 2007; pp. 67–74.
8. Kulldorff G. *Estimation from grouped and partially grouped samples*. Wiley: New York, 1961.
9. Haitovsky Y. *Regression estimation from grouped observations*. Griffin: London, 1973.
10. Heitjan DF. Inference from grouped continuous data. *Statistical Science* 1989; **4**:164–179.
11. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, Van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute* 2000; **92**:205–216.
12. Miaskowski C, Dodd M, West C, Paul SM, Schumacher K, Tripathy D, Koo P. The use of a responder analysis to identify differences in patient outcomes following a self-care intervention to improve cancer pain management. *Pain* 2007; **129**:55–63.
13. Altman DG. Statistics in medical journals: some recent trends. *Statistics in Medicine* 2000; **19**:3275–3289.
14. Senn S. Disappointing dichotomies. *Pharmaceutical Statistics* 2003; **2**:239–240.
15. Senn S. An unreasonable prejudice against modelling? *Pharmaceutical Statistics* 2005; **4**:87–89.
16. Cox DR. Note on grouping. *Journal of American Statistical Association* 1957; **52**:543–547.
17. Johnson NL, Kotz S, Balakrishnan N. *Continuous univariate distributions*, Vol. 1. Wiley: New York, 1994.
18. Wu CFJ. Optimal design for percentile estimation of a quantal response curve. In *Optimal design and analysis of experiments*, Dodge Y, Fedorov VV, Wynn HP (eds). North-Holland: Amsterdam, 1988; pp. 213–223.
19. Torsney B, Musrati AK. On the construction of optimal designs with applications to binary response and to the weighted regression models. In *Model-oriented data analysis*, Muller WG, Wynn HP, Zhigljavsky AA (eds). Physica-Verlag: Wurzburg, 1992; pp. 37–52.
20. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 1992; **3**:434–440.
21. Lehman EL. *Theory of point estimation*. Springer: New York, 1983.
22. Lindgren BW. *Statistical theory*. Chapman & Hall: New York, 1993.
23. Johnson NL, Kotz S, Balakrishnan N. *Continuous univariate distributions*, Vol. 2. Wiley: New York, 1995.

APPENDIX A: PROOF OF RELATIVE EFFICIENCY WHEN ESTIMATING μ AND σ

$$R(u) = \frac{\mathrm{var}(\hat{p}_o)}{\mathrm{var}(\hat{p}_d)} = \frac{\mathrm{var}\left(\Phi\left(\frac{c-\hat{\mu}}{\hat{\sigma}}\right)\right)}{\mathrm{var}\left(\frac{n_1}{n}\right)}$$

Using the Taylor approximation, we obtain

$$R(u) \approx \frac{\left[\frac{1}{\sigma}\phi\left(\frac{c-\mu}{\sigma}\right)\right]^2\frac{\sigma^2}{n} + \left[\frac{c-\mu}{\sigma^2}\phi\left(\frac{c-\mu}{\sigma}\right)\right]^2\frac{\sigma^2}{2n}}{\frac{\Phi(u)[1-\Phi(u)]}{n}}$$
$$= \frac{\phi^2(u)}{\Phi(u)[1-\Phi(u)]}\left[1+\frac{1}{2}u^2\right]$$

APPENDIX B: PROOF OF LOSS OF INFORMATION FOR LOCATION-SCALE FAMILY

For the symmetric distributions in the location-scale family, we can show that dichotomization always leads to a loss of information as the inequality $I_d \leqslant I_o$ is a direct corollary of the Cauchy–Schwartz inequality. Indeed, observing that $\partial\phi(u)/\partial\mu = -\partial\phi(u)/\partial y$, $\mathrm{d}u = (1/b)\,\mathrm{d}y = -(1/b)\,\mathrm{d}\mu$, $u = (y-\mu)/b$, and $\phi(u) = \int_{-\infty}^{u} \times (\partial\phi(y)/\partial y)\,\mathrm{d}y = -\int_{-\infty}^{u}(\partial\phi(u')/\partial\mu)(1/b)\,\mathrm{d}u'$, accordingly to the Cauchy–Schwartz inequality ($\int_R fg\,\mathrm{d}v \leqslant (\int_R f^2\,\mathrm{d}v)^{1/2}(\int_R g^2\,\mathrm{d}v)^{1/2}$), and assuming without loss of generality that $u \leqslant \frac{1}{2}$ one can verify that

$$\phi^2(u) = \left[\frac{1}{b}\int_{-\infty}^{u}\frac{\partial\phi(u')}{\partial\mu}\mathrm{d}u'\right]^2$$
$$= \frac{1}{b^2}\left[\int_{-\infty}^{u}\frac{1}{\sqrt{\phi(u')}}\frac{\partial\phi(u')}{\partial\mu}\sqrt{\phi(u')}\,\mathrm{d}u'\right]^2$$
$$\leqslant \frac{1}{b^2}\int_{-\infty}^{u}\left[\frac{\partial}{\partial\mu}\ln\phi(u')\right]^2\phi(u')\,\mathrm{d}u'\int_{-\infty}^{u}\phi(u')\,\mathrm{d}u'$$
$$= \int_{-\infty}^{u}\left[\frac{\partial}{\partial\mu}\ln\phi(u')\right]^2\phi(u')\,\mathrm{d}u'\int_{-\infty}^{u}\phi(u')\,\mathrm{d}u'$$

or

$$\frac{\phi^2(u)}{\Phi(u)} \leqslant \int_{-\infty}^{u}\left[\frac{\partial}{\partial\mu}\ln\phi(u')\right]^2\phi(u')\,\mathrm{d}u'$$
$$\leqslant \frac{b^2}{2}E\left(\frac{\partial}{\partial\mu}\ln\phi(u)\right)^2 = \frac{b^2 I_o}{2}$$

Dividing by $b^2[1-\Phi(u)]$, we conclude that

$$I_d = \frac{\phi^2(u)}{\Phi(u)[1-\Phi(u)]} \leqslant \frac{I_o}{2[1-\Phi(u)]} \leqslant I_o$$

DOI: 10.1002/pst