DETERMINING THE CUTOFF BASED ON A CONTINUOUS VARIABLE TO DEFINE TWO POPULATIONS

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirement for the Degree of
DOCTOR OF PHILOSOPHY

By
Shu Li
May, 2012

Examining Committee Members:

Milton Parnes, Advisory Chair, Statistics
Damaraju Raghavarao, Advisory Chair, Statistics
Pallavi Chitturi, Statistics
Ivan SF Chan, Merck & Co.
Alexandra L Hanlon, University of Pennsylvania, School of Nursing

ABSTRACT

DETERMINING THE CUTOFF BASED ON A CONTINUOUS VARIABLE TO DEFINE TWO POPULATIONS

Shu Li

DOCTOR OF PHILOSOPHY

Temple University, May 2012

Advisor Chair: Drs. Milton Parnes and Damaraju Raghavarao

In clinical research, it is sometimes desirable to dichotomize a continuous variable so that the information expressed using a dichotomous variable is more straightforward for clinicians to interpret and communicate. The distribution of a continuous variable can differ between two populations defined by the case status. Under such a scenario, the dichotomization process can be based on distributions of the continuous variable in two distinct populations. The resulting dichotomous variable can be used as an endpoint in future studies. Even though dichotomization has not been extensively studied, dichotomization has been commonly carried out in clinical trials. We developed a methodology on dichotomization based on maximizing the correlation between the two populations and the dichotomous variable. We have investigated several commonly assumed distributions (e.g., normal, log-normal and gamma distribution) of the continuous variable for the two populations and developed a numerical algorithm for the proposed method to determine the optimal cutoff point. The two populations can differ in form and/or parameters. The proposed method of finding the optimal cutoff was also extended to adjust for covariates. In real world scenarios where the two samples from the two populations are not completely

identified, we recommended using the EM method to first estimate the parameters associated with the two populations before applying the proposed method to find the optimal cutoff point. The performance of the proposed method with the numerical algorithm and the EM method has been studied for several theoretical distributions and using simulated data. These methods were also applied to a varicella vaccine example.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my co-advisers Dr. Milton Parnes and Dr. Damaraju Raghavarao, for their excellent guidance and extraordinary patience throughout the work on my dissertation at Temple University. Their encouragement brought forth my best efforts and professional growth. They taught me invaluable lessons about the key components of quality statistical research. A special mention needs to be given to Dr. Ivan SF Chan, who is a committee member. He was also my colleague and manager during the time I worked at Merck, where I have gained invaluable experience that helped me a great deal with my current work and dissertation. This dissertation was motivated by the projects that I was involved in at Merck. This dissertation will not be done without his guidance and help. I would also like to thank the other members of my committee, Dr. Pallavi Chitturi and Dr. Alex Hanlon for their contributions, which substantially improved this work.

I am fortunate to have a friendly work environment and constant support of my management and colleagues at Johnson and Johnson. I truly appreciated the time spent in sharing the statistical knowledge and the working experience with them.

Most of all I am grateful to my husband Weidong Xiao, who has supported me through the pursuit of the doctoral degree in Statistics. His patience, love, and support made it possible for me to achieve this goal. Also I would like to thank my children Jennifer, Hannah and Eileen for their support, patience and tolerance in my frustration at times. Without their support, I would not have been able to accomplish this goal. I would also like to thank my parents-in laws and my parents for their generous support. The care they gave to my children allowed me to focus on the Ph.D work. I offer you all with my love and gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BTN	Benefit associated with true negative
BTP	Benefit associated with true positive
CFN	Cost associated with false negative
CFP	Cost associated with false positive
EM	Expectation and maximization
FPR	False positive rate
MR	Misclassification rate
PDF	Probability density functions
ROC	Receiver operating characteristic
TNR	True negative rate
TPR	True positive rate

CHAPTER 1

1 INTRODUCTION

1.1 Background

It is argued that continuous variables capture more information than binary variables. For example, as is well known, if a continuous variable is normally distributed, the Pitman efficiency of sign test compared with the t-test is about 64 percent ($\frac{2}{\pi}$) (Cochran et al., 1961). This is deemed to be the optimistic analogy to the loss in efficiency in dichotomizing when the median was used as the cutoff point. Others have also pointed out that sometimes, the dichotomization of a continuous variable is arbitrary (Senn 2003). Because of these reasons, dichotomizing a continuous variable has not been extensively studied. However, in some cases, a binary variable would constitute a more clinically meaningful endpoint and it is a common practice in a clinical setting to dichotomize a continuous variable. For example, we define hypertensive or not, diabetic or non-diabetic, fever or no fever, based on laboratory tests or temperature which is continuous in scale. Their use allows us to describe our perceptions that we can understand. This can further facilitate communication and subsequent follow-up or treatment. In studies with psoriasis trials, for example, disease severity is commonly assessed by psoriasis area and severity index (PASI), an index used for assessing and grading the severity of psoriatic lesions and their response to therapy. The PASI produces a numeric score that can range from 0 to 72. Instead of using the score as a continuous variable, the gold standard endpoint used in clinical trials to assess the response to therapy is PASI 75 (at least 75% improvement from baseline in PASI score). Of course, there are cases of inappropriate definitions of response based on dichotomizing a continuous variable. Senn (2009) has criticized several inappropriate uses of the arbitrary dichotomy. Lewis (2004) has defended the use of dichotomy by saying that analysis using dichotomous variables is more direct, requires less in the way of assumptions and can account more robustly for some aspects of missing data. However, he did acknowledge that, because of the potential loss of power, the significance test could be done in other scales (e.g., continuous or ordinal scale), but the analysis based on the dichotomous variables can serve as a means of estimating the practical clinical consequence of effects shown to be statistically significant by other means. Yue and Xie (2008) raised an interesting scenario in which dichotomization may provide better results when the continuous variable is contaminated. Lastly, both FDA and EMA have endorsed the use of the responder analysis. Sometimes, the definition of responder is defined in the guidance documents. Therefore, in practice, continuous and dichotomous variables can be used complementarily to achieve both statistical and clinical goals.

This research topic is motivated by vaccine development, where it is useful to identify immune markers (e.g., antibody response) that could predict the clinical vaccine efficacy. Once the immune markers are identified, more efficient evaluations of new vaccines or manufacturing process changes can be performed using the immune markers. This will greatly increase the efficiency of the vaccine development since a traditional efficacy trial would take a long time to complete and/or require a large number of participants, particularly for a rare disease. Identifying immune markers are best done in a prospectively planned efficacy trial where the potential immune markers and clinical endpoints (e.g., diseases vaccine attempts to prevent) are both captured. Once the information is available, the relationship between immune responses and protective efficacy is often evaluated through the concept of correlate of protection by establishing a "protective level" of immune response above which a person is considered protected from the disease while those who do not achieve this level are considered to be still susceptible to disease. By examining vaccine failures (i.e., vaccinated subjects who come down with the disease) and non-failures, this "protective level" of an immune response can be used to determine whether a person is "completely protected" or "not protected and still susceptible to the disease". Once this level has been established, the dichotomous variable based on this level is used as the primary endpoint in subsequent clinical trials to evaluate the effectiveness of vaccines. On a population basis, this "protective level" corresponds to a 100% protection. Unfortunately, this ideal level of protection is usually non-existent. However, for most of the vaccines, on a population level, the protective efficacy tends to increase with higher immune response. For example, clinical trials with a live attenuated varicella (Oka/Merck) vaccine have shown a high protective efficacy against chickenpox as reported by Wibel et al. (1984), Kuter et al. (1991) and Vessey et al. (2001), as well as a clear inverse relationship between the frequency of chickenpox breakthrough among vaccinated subjects and the post-vaccination antibody titer measured by the glycoprotein enzyme-linked immunosorbent assay as described by White et al. (1992), Li et al. (2002) and Chan et al. (2002). However, no particular level of antibody can be considered as a clear-cut "protective level" since there exist vaccinated subjects who achieve this level but still contract the disease later. Therefore, in practice, a protective level that corresponds to a very high level of protection (say 95%) is considered as an approximate correlate of protection (Li et al., 2002).

In an attempt to identify immune markers that correlate with disease protection, we propose a method of determining the optimal cutoff point based on maximizing the correlation between the two populations (e.g., cases and non-cases) and a dichotomous variable (such as an immune marker). Since the correlation is a measure of association between the disease status and the dichotomous variable, maximizing this correlation may yield an optimal cutoff value that can be considered as a correlate of protection or an approximate correlate of protection.

1.2 A Mathematical Setup

In this section, we assume that 1) a continuous variable can be measured on both the disease case population and the non-case population, 2) the distributions of the continuous variable are different between the two populations either in form and/or parameters, and 3) the incidence rate of the disease is π. The first binary endpoint is defined by dichotomizing the continuous variable as the protected vs. non-protected based on a cutoff:

$Y = 0$, if a person is protected from the disease (ie. continuous variable $\geq$ the cutoff point),

$Y = 1$, if a person is not protected from the disease (ie. continuous variable $<$ the cutoff point).

The second binary endpoint is defined by whether a person developed a clinical disease case of interest:

Z = 0, if a person belongs to non-case population,

Z = 1, if a person belongs to disease case population.

Let

π = P (Z=1),

FPR = false positive rate = P (Y=1 | Z=0),

FNR = false negative rate = P (Y=0 | Z=1).

Based on these two binary variables, we construct the following 2x2 table.

Table 1 Cross-classification probability table based on two binary variables

Protection Status Based on the Dichotomized Variable	Disease Status		
	Non-case (Z = 0)	Case (Z = 1)	
Protected (Y =0)	(1-π)(1-FPR)	πFNR	π(FNR+FPR-1)+(1-FPR)
Not protected (Y=1)	(1-π)FPR	π(1-FNR)	π(1-FNR-FPR)+ FPR
	1-π	π	1

The correlation between Y and Z is:

$$\rho(\mathrm{Y, Z}) = \frac{\pi(1-FNR) - \pi\ [\pi(1-FNR-FPR)+FPR]}{\sqrt{(\pi(1-\pi)[\pi(1-FNR-FPR)+FPR][1-\pi(1-FNR-FPR)-FPR])}}$$

$$= \frac{\sqrt{\pi(1-\pi)}(1-FNR-FPR)}{\sqrt{[\pi(1-FNR-FPR)+FPR][1-\pi(1-FNR-FPR)-FPR)]}} \qquad (1)$$

It is assumed that there is some positive correlation between Y and Z. The optimal cutoff point is then defined as the one that maximizes the correlation between Y and Z given in equation (1). For a given cutoff, the misclassification rate is:

$$\mathrm{MR} = \pi * FNR + (1-\pi) * FPR \tag{2}$$

One can also define an optimal cutoff point by minimizing the misclassification rate which we will refer to as the misclassification method.

1.3 Research Contribution

The study of the optimal cutoff point is a relatively new topic that has not been extensively studied in the literature. Most of cutoff points in vaccine literature have been determined empirically and no cutoff based on a theoretical distribution has been established. Lakshmi (1995) in her Ph.D dissertation investigated finding a cutoff point by maximizing the asymptotic relative efficiency for testing the equality of the two proportions based on the sample sizes with and without misclassification. She considered normal and exponential distributions for the populations. Some of her thesis work was published in Damaraju (2009). Maximizing the asymptotic relative efficiency is known to be the same as maximizing the squared correlation between the actual and the predicted binary responses when the prevalence rate is 0.5. In Padgett (2011), she gave SAS programs to find the cutoff point by maximizing the correlation for given data. The present dissertation extends the above work for the scenario where all the case status and the non-case status have been observed as well as the scenario where only a sample of the cases have been observed and another sample contains a mix of the cases and non-cases. In our work, the incidence can range from

0 to 1. This work also considers other common distributions for both populations and allows different distributions or a mixture of distributions for the two populations.

Even though the research is motivated by vaccine development, this work can be used for other areas and provide a framework for finding an optimal cutoff if one needs to dichotomize a continuous variable. From a methodology perspective, our study has the following features: 1) We have assumed common distributions for the continuous variables; 2) The two populations (case vs. non-case) can be different in form and/or parameters; 3) We have developed a numeric algorithm based on the proposed method to solve for the optimal cutoff. The solution is relatively easy to implement; 4) Due to complex nature of the solution, we obtain the distribution and the confidence interval for the optimal cutoff point via a bootstrap technique; 5) This method can be extended to adjust for covariates that are related to either the case-status and/or the continuous variable; 6) In real world scenarios where one sample comes from the case population and the second sample is a mixture of the case and the non-case populations, the EM method can be used to first estimate the parameters associated with the two populations before applying the proposed method to find the cutoff point.

1.4 Dissertation Structure

The remaining chapters of this dissertation are organized as follows. Chapter 2 provides the literature review including the receiver operating characteristic (ROC) curve and other methods of finding the cutoff points. Chapter 3 presents a number of new results for the problem of finding the optimal cut-off point based on maximizing the correlation, under a variety of common distributions of the continuous variable. It also presents the extension of the research to incorporating covariates that may be related to the case-status and/or the continuous variable. Chapter 4 presents how to estimate the parameters based on the expectation-maximization (EM) method in the real world scenarios where one sample comes from the case population and the second sample is a mixture of the case and the non-case populations. Chapter 5 presents some numerical results via simulations and/or based on a real-world vaccine example. The proposed method is also compared with the traditional methods such as the ROC-curve method or the misclassification method. Chapter 6 concludes by discussing the contributions of this work and offering suggestions for future research.

CHAPTER 2

2 LITERATURE REVIEW

2.1 Introduction

The literature review consists of two parts. Section 2.2 gives a general review of some empirical methods of defining the optimal cutoff. Section 2.3 reviews the basics of the traditional ROC curve, its evaluation criteria, and its application to finding the optimal cutoff. This section also discussed other methods such as minimizing misclassification rate or maximizing the asymptotic relative efficiency for testing the equality of the two proportions based on the sample size with and without misclassification.

2.2 Methods for defining optimal cutoff

As mentioned in Section 1.1, the cutoff for vaccines is complicated by the fact that subjects who are vaccinated with high level of antibodies still come down with the disease vaccines try to prevent. However, for some diseases (e.g., diphtheria, tetanus, haemophilus influenza type b, hepatitis B), these protective levels have been proposed (Kayhty et al., 1983; Anderson et al., 1984; White et al., 1992; Olin et al., 1995; Shahin et al., 1992; Keith et al., 1996; Crosnier et al., 1981; and Crosnier et al., 1981). For the most part, these protective levels were established by correlating the clinical vaccine efficacy and immune response or other clinical judgment.

Klugman et al. (1996) has used a combination of the protective efficacy together with the distribution of antibodies in the vaccinated and the control group to derive an estimate of the protected level of the antibodies. For example, they proposed to define a protective level at which the relative ratio of antibody responses (the proportion of subjects exceeding the protective level in the vaccinated group/the proportion of subjects exceeding the protective level in the control group) exceeds the relative risk of disease (the proportion of subjects developing disease in the control group/the proportion of subjects developing disease in the vaccinated group). In the paper, they also defined the protective level based on the clinical efficacy and the antibody distribution in the vaccinated group. For example, if the efficacy is 60%, the protective level then corresponds to the 40^{th} percentile of the antibody levels in the vaccinated group. Li et al. (2002) have proposed to determine the cutoff as such that above that level most individuals (e.g., > 95%) are protected from the disease based on the data. Kohberger (1998) has also proposed similar ideas as Li et al. where the protective level can be determined by a desirable and somewhat arbitrary risk level p of contracting the disease. In their research, the authors proposed to first build a logistic model based on data and then set a very low risk level and solve for the cutoff.

Lakshmi (1995) in her Ph.D dissertation investigated finding a cutoff point by maximizing the asymptotic relative efficiency for testing the equality of the two proportions based on the sample size with and without misclassification. Specifically, in her dissertation, she defined a screening variable X which can be used to classify an experimental unit into either normal (N) or abnormal ($\overline{N}$) categories based on a threshold point. Let π be the incidence of abnormal units and θ be the probability that an experimental unit be classified as abnormal based on the screening variable X. She showed that the asymptotic relative efficiency for testing the null hypothesis of $\pi \leq \pi_0$ against the alternative hypothesis of $\pi > \pi_0$ with and without misclassification is

$$R^2 = \frac{[1 - l_1 - l_2]^2}{1 - [l_1 - l_2]^2},$$

where l_1 and l_2 are false positive and negative rates, respectively. When π_0=0.5, R is the correlation between the actual and predicted binary responses. Damaraju (2009), Pagett (2011) provide some details.

2.3 Receiver Operating Characteristic (ROC) curve

Receiver operating characteristic (ROC) methodology has been used in some areas in determining the optimal cutoff. ROC methodology is based on statistical decision theory and was developed in the electronic signal detection and problems in radar in the early 1950s (Metz 1986). An ROC type of plot was used in the late 1950s to describe the ability of an automated Pap smear analyzer to discriminate between smears with and without malignant cells. ROC analyses were later widely used in other fields. Most notably, it is widely used in the medical field in the laboratory setting to describe the diagnostic accuracy of different tests. The ROC curve depicts the overlap of the diseased vs. non-diseased population by plotting the sensitivity (i.e., true positive) as the Y-axis and 1-specificity (i.e., false positive) as the X-axis based on the range of values as the cutoff point. Therefore, it shows the tradeoff between the sensitivity and specificity. If there is no overlap between the two populations, there will be cutoff points to enable the ROC curve to pass through the upper left corner (i.e., both sensitivity and specificity equal to 1). For tests (values) with no discrimination, the ROC curve will be a 45 degree line from the lower left corner to the upper right corner. ROC curves can also be used to compare different tests with different accuracies. Since for a given specificity, one would prefer better sensitivity and vice versa, therefore tests with higher area under the ROC curve are preferred.

Metz (1978) suggested that a graphical approach based on the ROC curve could be used to derive the optimal cut-off point. Stein (2005) proposed to select an optimal cut-off based on ROC analysis by minimizing the cost function or maximizing the net profit value in the banking industry. In those cases, it makes certain assumptions about the default rates and costs associated with the false negative and the false positive as well as benefit associated with the true negative and the true positive. Similarly, one can develop a cost function associated with the disease being studied and then find the optimal cutoff that minimizes the cost function (Zhou et al., 2002) which we will refer to as the ROC-curve method. For example, we can define a cost function as follows:

$$C_s = \pi * CFN * FNR - \pi * BTP * TPR + (1-\pi) * CFP * FPR - (1-\pi) * BTN * TNR,$$

Where, CFN = cost associated with false negative

BTP = benefit associated with true positive

CFP = cost associated with false positive

BTN = benefit associated with true negative

TPR = true positive rate = P (Y=1 | Z=1)

TNR = true negative rate = P (Y=0 | Z=0)

FPR = false positive rate = P (Y=1 | Z=0)

FNR = false negative rate = P (Y=0 | Z=1)

π = disease incidence

Quantifying this using ROC curve, we get

FPR = k, TNR = 1 – k, TPR = ROC (k), FNR = 1 – ROC (k).

To minimize the cost function, we can differentiate C_s with respect to k and set it to zero resulting in the following condition:

$$S = \frac{dROC(k)}{dk} = \frac{(1-\pi)(CFP + BTN)}{\pi(CFN + BTP)}.$$

The cost function is minimized to determine the optimal cutoff which we refer to as the ROC-curve method. In the process, a graphical ROC is used in the literature to determine the cutoff point. As mentioned in Section 1.2, one can also minimize the misclassification rate as defined by equation (2) to derive a cutoff point (the misclassification method). If we assume the benefit of the true positive and the true negative is 0, and the cost for the false negative and the false positive is the same, it can be shown that the ROC method will be identical to the misclassification method.

CHAPTER 3

3 MAXIMIZING CORRELATION TO FIND THE OPTIMAL CUTOFF

3.1 Introduction

This chapter describes a number of new results for the study of an optimal cutoff based on different distributions of the continuous variable when the goal is to maximize the correlation. Section 3.2 provides the framework for the selection of optimal thresholds based on different types of distributions in the continuous variable. Section 3.3 discusses a numerical algorithm to find the cutoff since the closed form of the solution from Section 3.2 is not available. Section 3.4 discusses the bootstrap method to derive the distribution and confidence intervals around the identified cutoff point. Section 3.5 illustrates the proposed method based on the mathematical setup in Section 1.2 by plotting different parameters (the correlation, the misclassification rate, FNR, TPR) with different cutoff points. Section 3.6 extends the proposed method to adjust for covariates that could potentially affect the protection and/or the continuous variable.

3.2 Determination of the cutoff

This section will provide the derivation of the cutoffs based on different distributions on the basis of maximizing the correlation discussed in Section 1.2, Equation 1. In this chapter, first, the two populations are assumed to follow the same form of distributions with different parameters. This is also done for the two populations that follow two different forms of distributions (i.e., one with normal distribution and the other with gamma distribution). The distributions considered here are normal, log-normal and gamma distributions since these are the most commonly assumed distributions of data seen in clinical trials. In all of the scenarios, we first calculate the FPR and the FNR based on a given cutoff. We then calculate the correlation per Equation 1. One can calculate the correlations based on all possible cutoffs with an increment of a very small amount and then determine the optimal cutoff when the correlation is the maximum. We will refer the cutoff determined by this process as the true cutoff. Alternately, since the correlation is not a linear or quadratic function of the cutoff, in order to maximize the correlation, Taylor series expansion is used to simplify the correlation function to a quadratic function of the cutoff. Then the usual method of maximizing the quadratic function is used to determine the cutoff point. This method is referred to as the proposed method and numerical algorithm.

3.2.1 Normal distributions

We assume that the two populations follow normal distributions as follows: Population 1 (case population) follows the normal distribution with a mean of μ_1 and a standard deviation of σ and population 2 (non-case population) follows the normal distribution with a mean of μ_2 and the same standard deviation of σ with $\mu_1 < \mu_2$. It should be noted that the same standard deviation is assumed for simplicity. The same approach can be taken for distributions with different standard deviations. Under the above assumptions, the probability density function of the case population is:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right],$$

and the probability density function of the non-case population is:

$$f(x_2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_2-\mu_2)^2}{2\sigma^2}\right].$$

Then,

$$FNR = P\,(X_1 \geq t \mid \text{case})$$

$$= \int_t^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right]dx = 1-\Phi\left(\frac{t-\mu_1}{\sigma}\right),$$

$$FPR = P\,(X_2 < t \mid \text{non - case})$$

$$= \int_{-\infty}^{t}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right]dx = \Phi\left(\frac{t-\mu_2}{\sigma}\right).$$

Therefore,

$$\rho(Y, Z) = \frac{\pi(1-\pi)(1-FNR-FPR)}{\sqrt{[\pi(1-FNR-FPR)+FPR][1-\pi(1-FNR-FPR)-FPR]}}.$$

$$= \frac{\sqrt{\pi(1-\pi)}\left[\Phi(\frac{t-\mu_1}{\sigma})-\Phi(\frac{t-\mu_2}{\sigma})\right]}{\sqrt{\left\{\pi\left[\Phi(\frac{t-\mu_1}{\sigma})-\Phi(\frac{t-\mu_2}{\sigma})\right]+\Phi(\frac{t-\mu_2}{\sigma})\right\}\left\{1-\pi\left[\Phi(\frac{t-\mu_1}{\sigma})-\Phi(\frac{t-\mu_2}{\sigma})\right]-\Phi(\frac{t-\mu_2}{\sigma})\right\}}}.$$

Let $a = \Phi(\frac{t-\mu_1}{\sigma})-\Phi(\frac{t-\mu_2}{\sigma})$

$$c = \pi\left[\Phi(\frac{t-\mu_1}{\sigma}) - \Phi(\frac{t-\mu_2}{\sigma})\right] + \Phi(\frac{t-\mu_2}{\sigma}),$$

$$b = \sqrt{c(1-c)}.$$

Then,

$$\rho\,(Y, Z) = \frac{\pi(1-\pi)a}{\sqrt{\pi(1-\pi)c(1-c)}} = \sqrt{\pi(1-\pi)}\,\frac{a}{b}.$$

To maximize ρ (Y, Z), we maximize $\frac{a}{b}$.

To determine the cutoff point, we approximate $\frac{a}{b}$ in a quadratic form using Taylor series expansion centered at C_0.

Now:

$$\frac{a}{b} \approx \frac{a}{b}\Big|_{t=c_0} + \frac{d}{dt}(\frac{a}{b})\Big|_{t=c_0}(t-c_0) + \frac{1}{2}\frac{d^2}{dt^2}(\frac{a}{b})\Big|_{t=c_0}(t-c_0)^2.$$

We know:

$$\frac{d}{dt}(\frac{a}{b}) = \frac{b\frac{da}{dt} - a\frac{db}{dt}}{b^2},$$

$$\frac{d^2}{dt^2}(\frac{a}{b}) = \frac{b^2\frac{d^2a}{dt^2} - ab\frac{d^2b}{dt^2} - 2b\frac{da}{dt}\frac{db}{dt} + 2a(\frac{db}{dt})^2}{b^3},$$

where,

$$\frac{da}{dt} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_1)^2}{2\sigma^2}\right] - \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right],$$

$$\frac{dc}{dt} = \pi\frac{da}{dt} + \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right],$$

$$\frac{db}{dt} = \frac{1}{2\sqrt{c(1-c)}}(1-2c)\frac{dc}{dt} = \frac{1-2c}{2\sqrt{c(1-c)}}\left\{\pi\frac{da}{dt} + \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right]\right\},$$

$$\frac{d^2a}{dt^2} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_1)^2}{2\sigma^2}\right]\left[-\frac{(t-\mu_1)}{\sigma^2}\right] + \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right]\left[\frac{(t-\mu_2)}{\sigma^2}\right],$$

$$\frac{d^2b}{dt^2} = \frac{1}{2}\left[\frac{-2\frac{dc}{dt}c(1-c) - \frac{1}{2}(1-2c)^2\frac{dc}{dt}}{[c(1-c)]^{\frac{3}{2}}}\right]\left\{\pi\frac{da}{dt} + \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right]\right\}$$

$$+\frac{1-2c}{2\sqrt{c(1-c)}}\left\{\pi\frac{d^2a}{dt^2} - \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(t-\mu_2)^2}{2\sigma^2}\right]\frac{(t-\mu_2)}{\sigma^2}\right\}.$$

Hence,

$$\frac{a}{b} \approx \frac{a}{b}\Big|_{t=c_0} + \frac{d}{dt}\left(\frac{a}{b}\right)\Big|_{t=c_0}(t-c_0) + \frac{1}{2}\frac{d^2}{dt^2}\left(\frac{a}{b}\right)\Big|_{t=c_0}(t-c_0)^2.$$

and the cutoff that maximizes the correlation is:

$$\text{Cutoff} = C_0 - \frac{\frac{d}{dt}\left(\frac{a}{b}\right)\Big|_{t=c_0}}{\frac{d^2}{dt^2}\left(\frac{a}{b}\right)\Big|_{t=c_0}}.$$

Note that we need to substitute $\frac{d^2}{dt^2}\left(\frac{a}{b}\right)\Big|_{t=c_0}$ and $\frac{d}{dt}\left(\frac{a}{b}\right)\Big|_{t=c_0}$ in the above equation.

3.2.2 Log-normal distribution

We assume the two populations follow log-normal distributions, that is, the log of the continuous variable for Population 1 (case population) follows a normal distribution with a mean of μ_1 and a standard deviation of σ and the log of the continuous variable for population 2 (non-case population) follows a normal distribution with a mean of μ_2 and the same standard deviation of σ. Again, the same standard deviation was assumed for simplicity. Therefore, the case population follows:

$$\mathrm{f}(\mathrm{x}_1) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_1 - \mu_1)^2}{2\sigma^2}\right]}{x_1},$$

Similarly, the non-case population follows:

$$\mathrm{f}(\mathrm{x}_2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_2 - \mu_2)^2}{2\sigma^2}\right]}{x_2}.$$

Then,

$$\mathrm{FNR} = \mathrm{P}\ (\mathrm{X}_1 \geq \mathrm{t} \mid \mathrm{case})$$

$$= \int_t^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x - \mu_1)^2}{2\sigma^2}\right]}{x} dx,$$

$$\mathrm{FPR} = \mathrm{P}\ (\mathrm{X}_2 < \mathrm{t} \mid \text{non-case})$$

$$= \int_0^{t} \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x - \mu_2)^2}{2\sigma^2}\right]}{x} dx.$$

The same approach as with the normal distribution can be followed and the optimal cutoff can be determined. It is also noted that the cutoff for the log-normal distribution can be obtained based on the normal distribution due to the relationship between the normal and the log-normal distribution. That is, if we determine the cutoff point for the normal distribution, the cutoff for the log-normal distribution can be obtained by the exponentiation of the cutoff obtained from the normal distribution.

3.2.3 Gamma distribution

We assume the two populations follow a gamma distributions below: Population 1 (case population) follows a gamma distribution with α_1 and β_1 and population 2 (non-case population) follows a gamma distribution with α_2 and β_2. Therefore, the case population follows:

$$f(x_1) = \frac{1}{\Gamma(\alpha_1)\beta_1^{\alpha_1}} x_1^{\alpha_1 - 1} \exp(-\frac{x_1}{\beta_1}),$$

Similarly, the non-case population follows:

$$f(x_2) = \frac{1}{\Gamma(\alpha_2)\beta_2^{\alpha_2}} x_2^{\alpha_2 - 1} \exp(-\frac{x_2}{\beta_2}).$$

Then,

$$\text{FNR} = P\ (X_1 \geq t \mid \text{case})$$

$$= \int_t^{+\infty} \frac{1}{\Gamma(\alpha_1)\beta_1^{\alpha_1}} x^{\alpha_1 - 1} \exp(-\frac{x}{\beta_1}) dx$$

$$\text{FPR} = P\ (X_2 < t \mid \text{non-case})$$

$$= \int_0^{t} \frac{1}{\Gamma(\alpha_2)\beta_2^{\alpha_2}} x^{\alpha_2 - 1} \exp(-\frac{x}{\beta_2}) dx$$

Again, a similar approach as in Section 3.2.1 can be followed.

3.2.4 Normal distribution for cases and gamma distribution for non-cases

We assume that the two populations follow the distributions as below:

Population 1 (case population) follows the normal distribution with mean of μ and standard deviation of σ and population 2 (non-case population) follows the gamma distribution with α and β . Under the above assumptions, the probability density function of the case population is:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_1-\mu)^2}{2\sigma^2}\right],$$

and the probability density function of the non-case population is:

$$f(x_2) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x_2^{\alpha_1 - 1} \exp(-\frac{x_2}{\beta}).$$

Then,

$$\begin{aligned} FNR &= P(X_1 \geq t \mid \text{case}) \\ &= \int_t^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx, \end{aligned}$$

$$\begin{aligned} FPR &= P(X_2 < t \mid \text{non - case}) \\ &= \int_0^t \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-\frac{x}{\beta}) dx. \end{aligned}$$

Again, a similar approach as in Section 3.2.1 can be followed.

3.3 Numerical algorithm

Since there is no closed form for the cutoff obtained in Section 3.2, we developed a numerical algorithm to solve for the cutoff.

Based on the above formulas, it is not hard to see that the cutoff would converge to C_0 when C_0 is close to the cutoff since the first derivative will be zero with the cutoff which maximizes the correlation. But how could we find the optimal cutoff? We have adopted the following algorithm.

1. Choose an initial value of C_0 which should be between the means of the two populations.
2. Calculate the cutoff based on C_0 using the formula

$$\text{Cutoff} = C_0 - \frac{\frac{d}{dt}(\frac{a}{b})|_{t=c_0}}{\frac{d^2}{dt^2}(\frac{a}{b})|_{t=c_0}}.$$

3. Calculate the absolute convergence criterion

$$|\varepsilon| = |Cutoff - C_0|.$$

4. Compare the absolute error $|\varepsilon|$ with the pre-specified error tolerance. If $|\varepsilon|$ > pre-specified tolerance (e.g., 0.0001), replace C_0 with the calculated cutoff and repeat Steps 2 – 4; else if $|\varepsilon|$ <= pre-specified tolerance or the number of the iterations has exceeded the maximum number of iterations (e.g., 1000), stop the algorithm.
5. At each step, for each cutoff point, the correlation can also be calculated and it can be confirmed that the final cutoff corresponds to the maximum correlation.

The above numerical algorithm converged to the optimal cutoff fairly quickly in the examples and simulations discussed in Chapter 5.

3.4 Distribution and 95% confidence intervals for the cutoff point

Since there is no closed-form formula to establish the 95% confidence interval for the cutoff, we relied on the bootstrapping technique to find the distribution and the 95% CIs for the cutoff. For example, if we have data for the two populations, we can randomly draw N samples with replacement from the combined populations. For each sample, we can estimate the relevant parameters and the cutoff point can then be derived based on the estimated parameters using the proposed method and numerical algorithm. Subsequently, a distribution of the cutoff can be obtained. The mean of the N cutoffs will be estimated as the optimal cutoff and the 2.5 percentile and 97.5 percentile of the N cutoffs are taken as the lower and upper limit of the 95 percent CI, respectively.

3.5 Illustrations of the proposed method

Initial studies were run to see whether such a cutoff point exists for different distributions. In these studies, we have assumed different distributions for the two populations as follows: 1) both populations follow the normal distributions; 2) both populations follow the log-normal distributions; 3) both populations follow the gamma distributions; 4) the case population follows a normal distribution and the non-case population follows a gamma distribution. It was confirmed that for each of the 4 scenarios, there exists a cutoff point which maximizes the correlation between Y and Z. Below for each scenario, we have first plotted the probability density functions (PDF) for each population with the two populations overlaid on one figure (Figures 1, 4, 7 and 10). When plotting the PDFs, the probability was calculated for each potential cutoff based on the continuous variable in an increment of 0.01. Similarly, based on different cutoff points, we have calculated the correlations based on Equation 1 and also the misclassification rate based on Equation 2 to see whether the cutoff which maximizes the correlation also minimizes the misclassification rate (Figures 2, 5, 8 and 11). Lastly, we plotted the FNR and the FPR along with the correlation to investigate the tradeoff between the FNR and the FPR (Figures 3, 6, 9 and 12). For the normal distribution, the log-normal distribution, the gamma distribution (exponential distribution), or a mix of the normal distribution and the gamma distribution, it is clear that there exists a cutoff which maximizes the correlation. When the incidence rate is 0.5, the cutoff which maximizes the correlation either minimizes the misclassification rate (normal and log-normal distributions) or almost minimizes the misclassification rate (the gamma distributions or a mix of the normal and the gamma distributions). However, this is not true when the incidence is different from 0.5. When the incidence rate is 0.5 and the correlation is at the maximum, the FPR and the FNR are about the same. Again, this is not true when the incidence is different from 0.5. It is also noted that different values of π would only minimally impact the cutoff point and the correlation while the further apart the two distributions are from each other, the higher the correlation is.

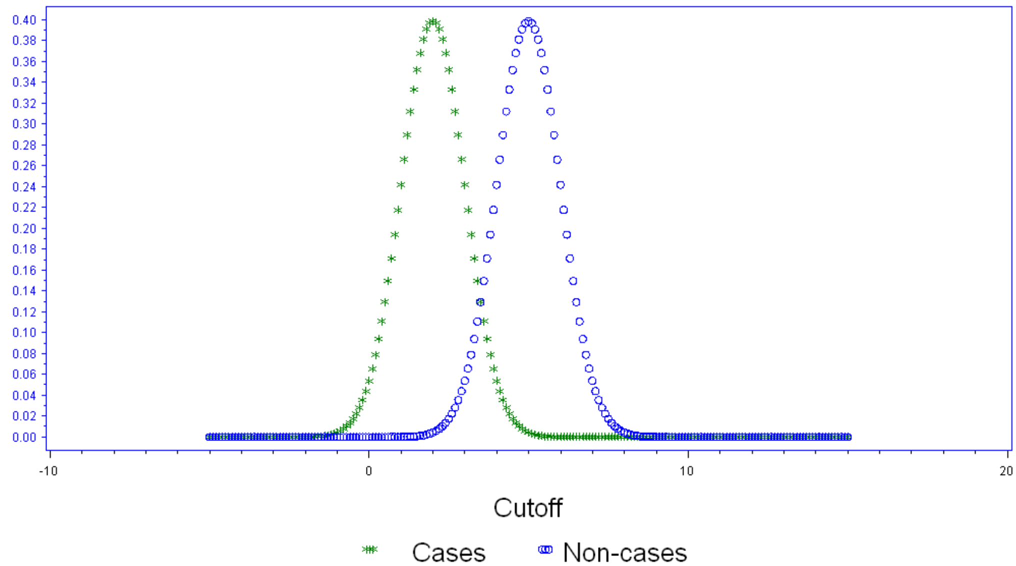

Figure 1 Probability density functions (PDFs) for case and non-case populations based on normal distributions (μ_1=5, μ_2=2, σ=1)

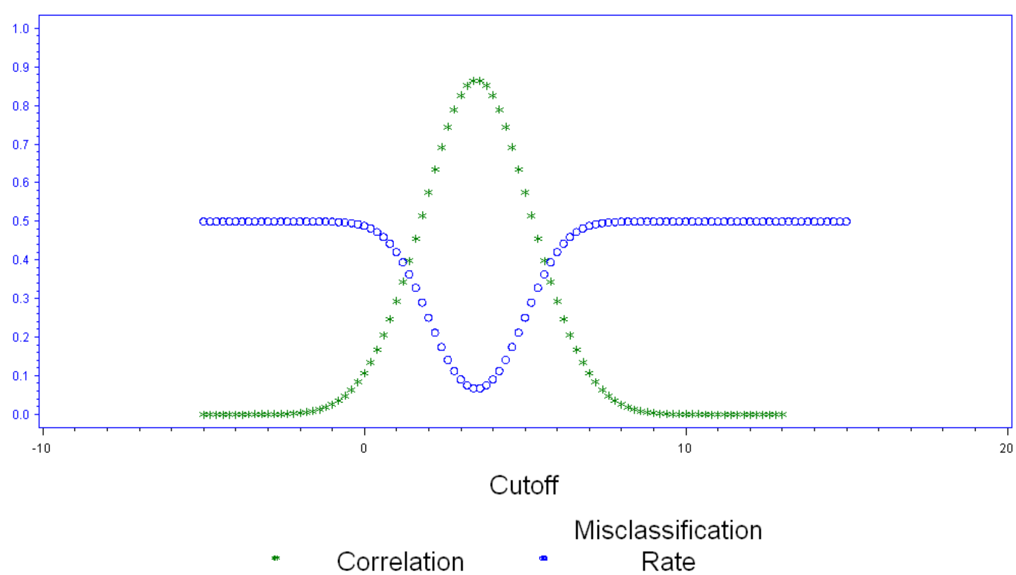

Figure 2 Correlation and misclassification rate using different cutoffs based on the normal distributions (μ_1=2, μ_2=5, σ=1)

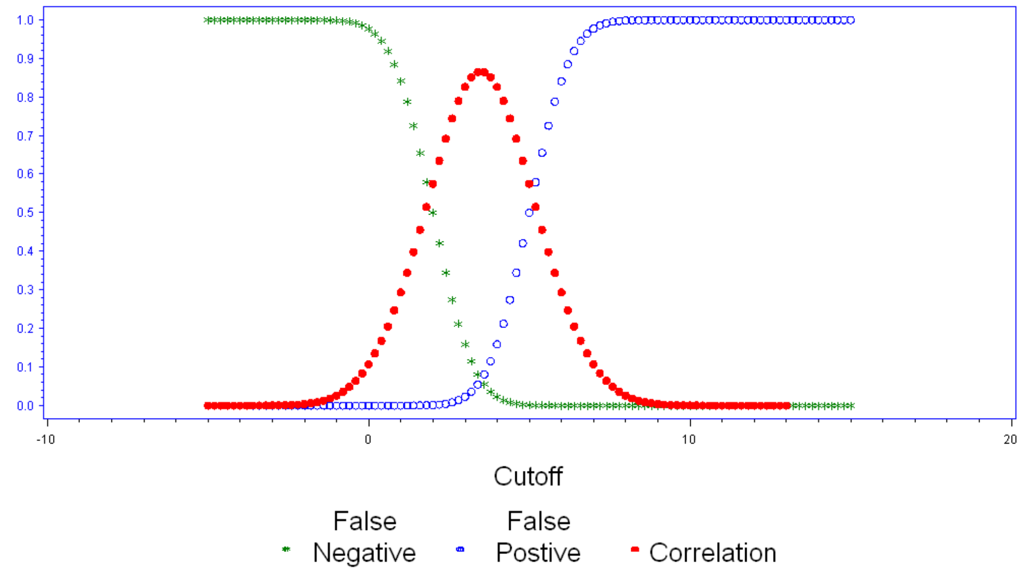

Figure 3 Correlation and false negative and false positive rate using different cutoffs based on normal distributions (μ_1=2, μ_2=5, σ=1)

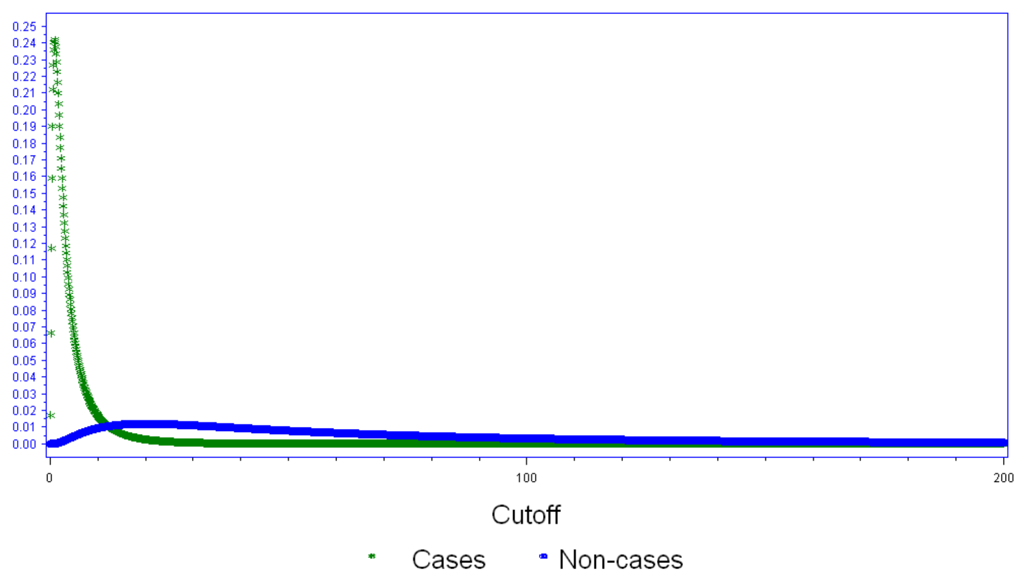

Figure 4 PDFs for case and non-case population based on lognormal distributions (μ_1=1, μ_2=4, σ=1)

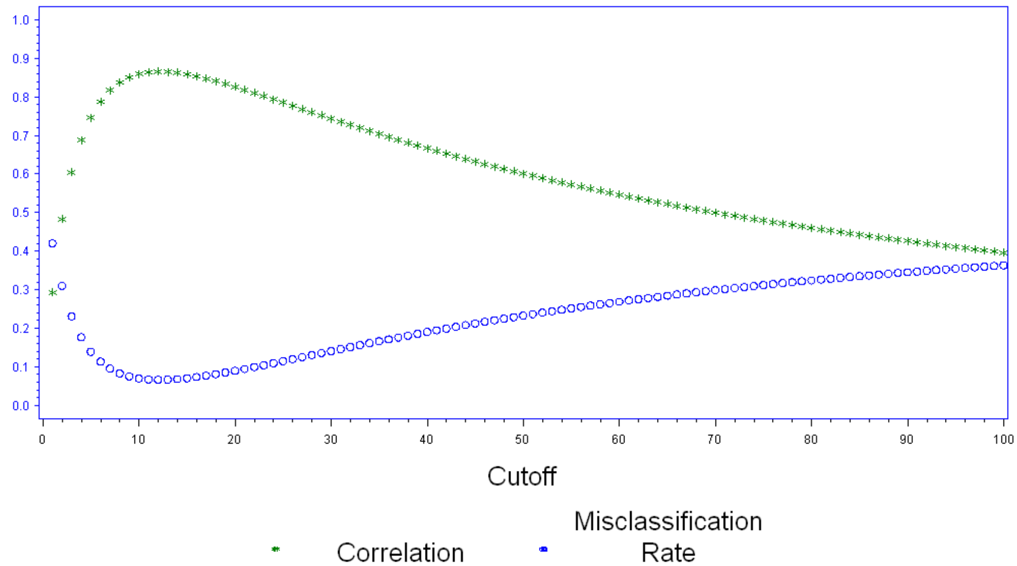

Figure 5 Correlation and misclassification rate using different cutoffs based on log-normal distributions (μ_1=1, μ_2=4, σ=1)

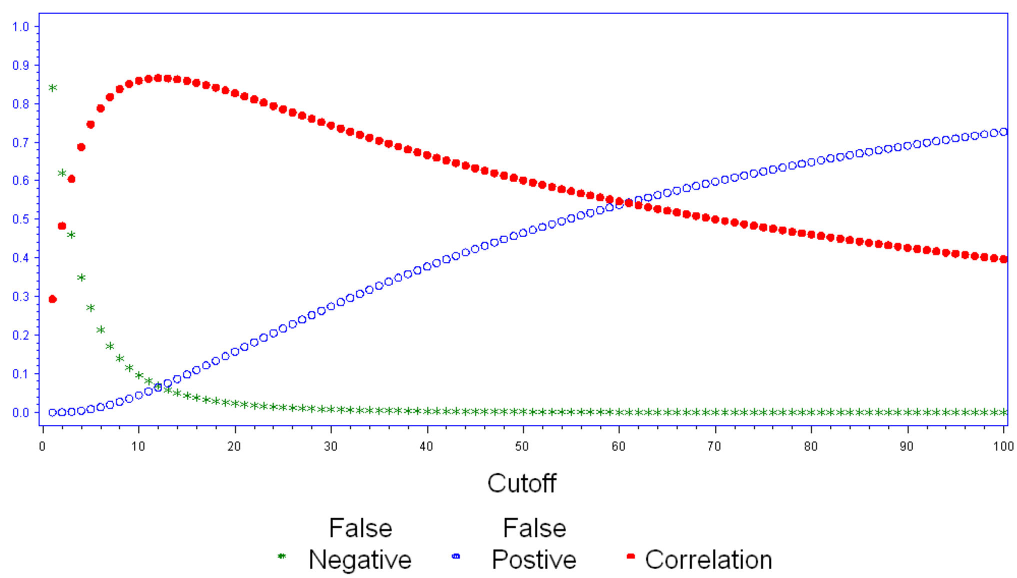

Figure 6 Correlation and false negative and false positive rate using different cutoffs based on log-normal distributions (μ_1=1, μ_2=4, σ=1)

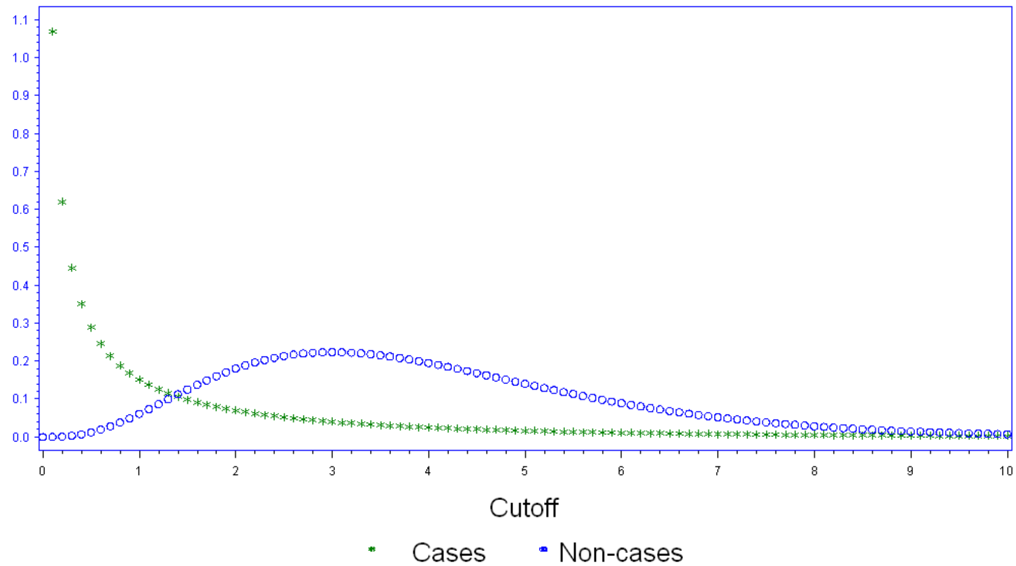

Figure 7 PDFs for case and non-case populations based on gamma distributions (α_1=0.25, β_1=4, α_2=4, β_2=1)

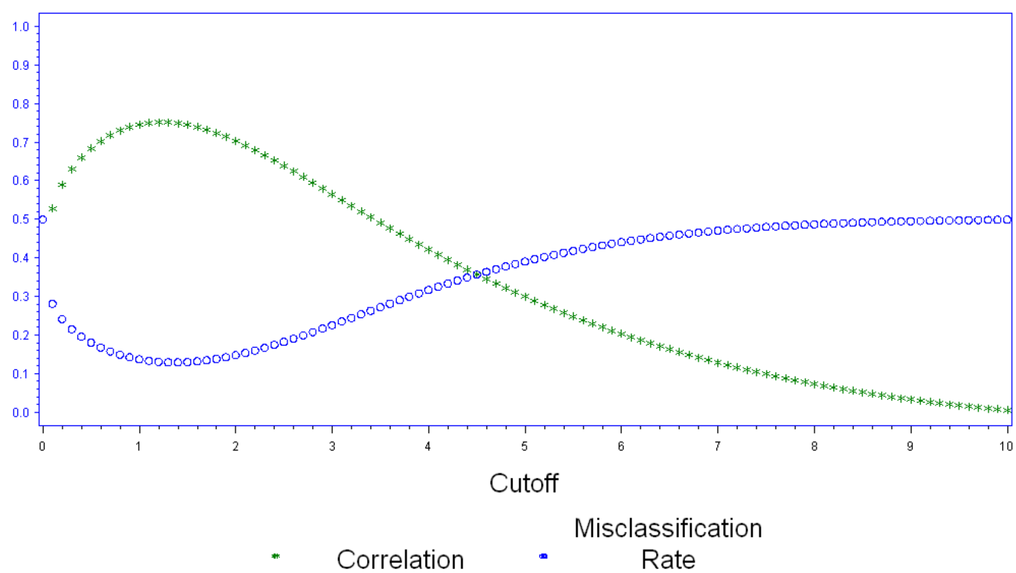

Figure 8 Correlation and misclassification rate using different cutoffs based on gamma distributions (α_1=0.25, β_1=4, α_2=4, β_2=1)

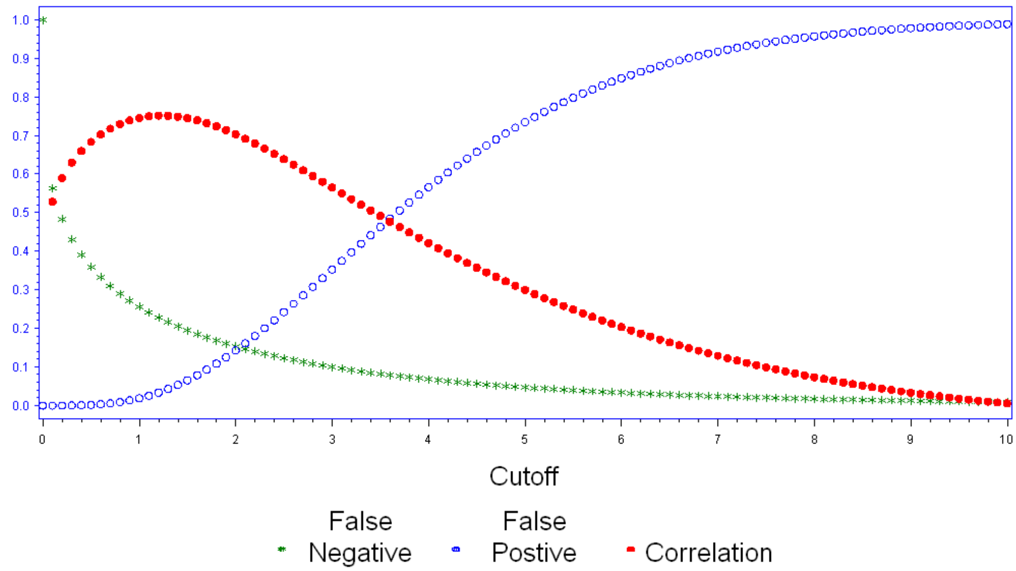

Figure 9 Correlation and false negative and false positive rate using different cutoffs based on the gamma distributions (α_1=0.25, β_1=4, α_2=4, β_2=1)

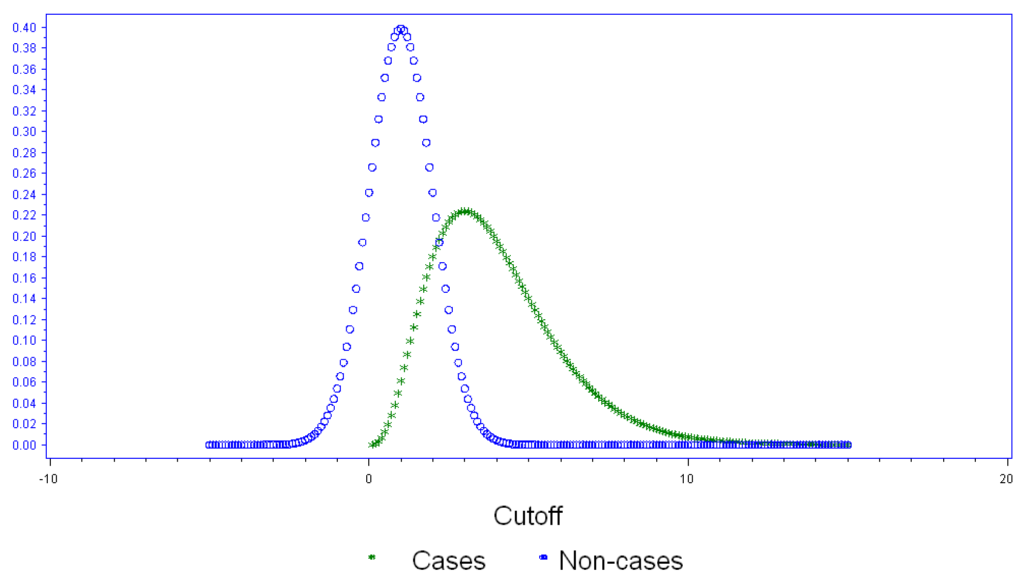

Figure 10 PDFs for case and non-case populations based on normal and gamma distributions (μ=1, σ=1, α=4, β=1)

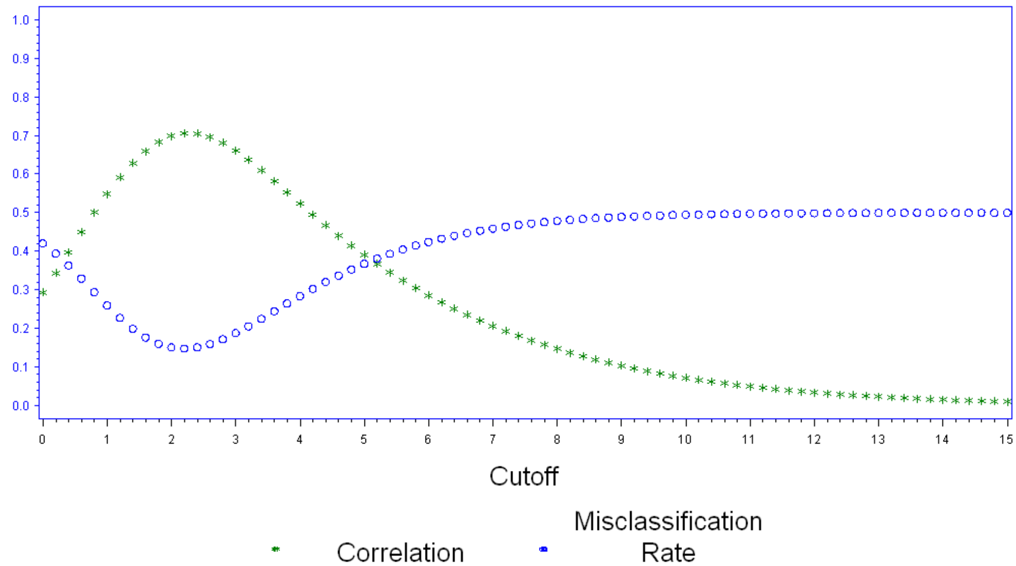

Figure 11 Correlation and misclassification rate using different cutoffs based on normal and gamma distributions (μ=1, σ=1, α=4, β=1)

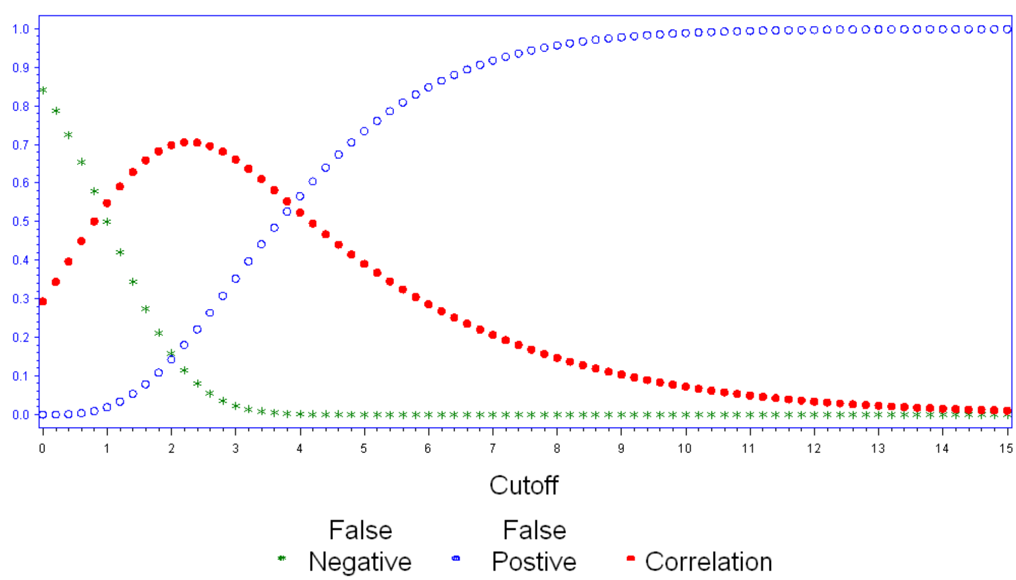

Figure 12 Correlation and false negative and false positive rate using different cutoffs based on the normal and gamma distributions (μ=1, σ=1, α=4, β=1)

3.6 Finding the optimal cutoff adjusting for covariates

In this section, we briefly discuss the extension of the proposed method to adjust for the covariates that could impact the cutoff. For example, in the varicella vaccine example discussed in Chapter 5, the protection is expected to be related to a child's age; the older the child is, the better the protection will be. In the extension, we model the FNR and the FPR by adjusting for covariates that could potentially affect the protection and/or the continuous variable using the generalized linear regression. This is accomplished by two separate models adjusting for covariates, one performed in the case population and the second one in the non-case population. Based on these two models, we can then calculate the adjusted FNR and the adjusted FPR. The adjusted FPR is estimated from the non-case population while the adjusted FNR is estimated from the case population. Specifically, if we use logit link (i.e., logistic models), the adjusted FPR and the adjusted FNR can be calculated as:

Adjusted FPR = P (Y=1 | Z=0, covariates x_is)

$$= \frac{\exp(\hat{\beta}_0 + \sum_{l=1}^{k} \hat{\beta}_l X_{li})}{1 + \exp(\hat{\beta}_0 + \sum_{l=1}^{k} \hat{\beta}_l X_{li})}$$

Adjusted FNR = P (Y=0 | Z=1, covariates x_js)

$$= \frac{1}{1 + \exp(\hat{\beta}_0^{'} + \sum_{l=1}^{k} \hat{\beta}_l^{'} X_{li})}.$$

Once the models are built, we can calculate the adjusted FPR and the adjusted FNR as the average of the predicted FPR and the predicted FNR across all subjects. We can then plug the adjusted FPR and the adjusted FNR into equation (1). The cutoff taking into account of the covariates will be the one that maximizes the correction function (1). We can also use the linear link to build the models. The similar steps will follow.

The above methods are to be evaluated through simulations. In order to generate the bivariate normal distribution, the following algorithm is to be used.

Assume that X and Y are marginally distributed as the normal distributions with means of μ_x and μ_y and the standard deviation of σ_x and σ_y, respectively. The correlation between X and Y is denoted by ρ. Therefore, in order to generate Y, a Z variate is created. Z is normally distributed as: mean of μ_z and the standard deviation of σ_z and is independently distributed of X such that $Y = aX + Z$; Therefore, given parameters for X, Y and a, Z can be derived. Specifically,

$$\mu_y = a\mu_x + \mu_z,$$

$$\sigma_y^2 = a^2\sigma_x^2 + \sigma_z^2,$$

$$\rho = \frac{EXY - EXEY}{\sqrt{Var(X)Var(Y)}} = \frac{E(X(X + aZ)) - EXE(X + aZ)}{\sqrt{Var(X)(a^2 Var(X) + Var(Z))}} = \frac{\sigma_x^2}{\sqrt{a^2\sigma_x^4 + \sigma_x^2\sigma_z^2}}.$$

Given the above equations, we can derive:

$$a = \frac{\rho\sigma_y}{\sigma_x},$$

$$\mu_z = \mu_y - a\mu_x,$$

$$\sigma_z = \frac{a\sigma_x\sqrt{1-\rho^2}}{\rho}.$$

The above process can be used to simulate the data to investigate the impact of the different combinations of correlations between X and Y within the case population and within the non-case population as well as different means and/or variances of X and Y within the case and the non-case population on the adjusted cutoff point.

CHAPTER 4

4 ESTIMATING PARAMETERS BASED ON THE EM METHOD

4.1 Introduction

This chapter describes the parameter estimates based on different distributions of the continuous variable when cases are not completely identified. Recall in Chapter 3, it is assumed that we have completely observed the case and the non-case samples. However, this may not be true in practice. For example, in a clinical trial setting, the study will be terminated at some point of time. Thus, case status can only be observed up to the time of termination. Presumably, cases can still develop after the study termination. Under this scenario, we need to estimate the distribution parameters. In this section, we assume that we have observed a case sample and the other sample contains both the unobserved cases and the non-cases. The goal is to estimate the parameters for the case and the non-case populations. The parameters will include the disease case incidence rate (i.e., π in Section 3) and distribution parameters for both populations.

4.2 Parameter estimation

In this section, we provide the method to estimate the parameters on the basis of maximizing the likelihood function. Ordinarily, if we have observed two populations and are certain about the case status, it is fairly easy for us to obtain the maximum likelihood estimates for the parameters. However, it is most likely that by the time we review the data, we have not observed all the cases. Therefore, the problem at hand is that we have observed certain number of the cases (sample 1) and a mixture of the unobserved cases and the non-cases (sample 2). In this section, we will describe how to estimate the parameters in this situation. The distributions considered are the normal distribution, the log-normal distribution, the exponential distribution and a mix of the normal distribution and the exponential distribution. In all of the scenarios, we used the EM method to solve the problem (Dempster et al., 1977). For sample 2, we define the probability that the observation arises from the case population to be *w*. We further define an unobserved variable V to denote if the observation comes from the case population. Therefore, the unconditional expectation E(V) is *w*. Under this setup, the EM method can be implemented to estimate the parameters. After *w* is estimated, π can easily be obtained.

4.2.1 Normal distributions

We assume that the two populations follow different normal distributions as follows:

Population 1 (the case population) follows the normal distribution with mean of μ_1 and the standard deviation of σ and population 2 (the non-case population) follows the normal distribution with mean of μ_2 and the same standard deviation of σ with $\mu_1 < \mu_2$. Therefore, the density of the case population is:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_1 - \mu_1)^2}{2\sigma^2} \right],$$

and the density of the non-case population is:

$$\mathrm{f}(x_2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_2-\mu_2)^2}{2\sigma^2}\right].$$

Suppose that we have observed sample 1 of size m (x_1, x_2, … x_m) from population 1 and another sample of size n (x_{m+1}, x_{m+2}, … x_{m+n}) from a mixture of population 1 and population 2, the likelihood of the n + m subjects will be as follows:

$$L(\mu_1,\mu_{2,}\sigma,w,x) =$$

$$\Pi_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\Pi_{i=m+1}^{m+n}\left\{\begin{array}{l}\frac{w}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\\+\frac{1-w}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_2)^2}{2\sigma^2}\right]\end{array}\right\}.$$

Let Z_i = 1 denote that the observation comes from the case population and Z_i = 0 denote that the observation comes from the non-case population in Sample 2. The likelihood becomes:

$$L(\mu_1,\mu_{2,}\sigma,w,x) =$$

$$\Pi_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\Pi_{i=m+1}^{m+n}\left\{\begin{array}{l}\left\{\frac{w}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\right\}^{Z_i}\\\left\{\frac{1-w}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_2)^2}{2\sigma^2}\right]\right\}^{1-Z_i}\end{array}\right\}.$$

Therefore, the log-likelihood becomes:

$$l(\mu_1,\mu_{2,}\sigma,w,x) =$$

$$\sum_{i=1}^{m}\left[-\ln(\sqrt{2\pi}\sigma)-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]+$$

$$\sum_{i=m+1}^{m+n}\left\{\begin{array}{l}Z_i\left[\ln(w)-\ln(\sqrt{2\pi}\sigma)-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\\+(1-Z_i)\left[\ln(1-w)-\log(\sqrt{2\pi}\sigma)-\frac{(x_i-\mu_2)^2}{2\sigma^2}\right]\end{array}\right\}.$$

E-step at t-step:

$$Q(\mu_1, \mu_{2,} w, \sigma, x \mid \hat{\mu}_1^t, \hat{\mu}_2^t, \hat{w}^t, \hat{\sigma}^t) =$$
$$\sum_{i=1}^{m}\left[-\ln(\sqrt{2\pi}\sigma) - \frac{(x_i - \mu_1)^2}{2\sigma^2}\right] +$$
$$\sum_{i=m+1}^{m+n}\left\{\begin{array}{l} E(Z_i^t)\left[\ln w - \ln(\sqrt{2\pi}\sigma) - \dfrac{(x_i - \mu_1)^2}{2\sigma^2}\right] \\ + \left[1 - E(Z_i^t)\right]\left[\ln(1-w) - \ln(\sqrt{2\pi}\sigma) - \dfrac{(x_i - \mu_2)^2}{2\sigma^2}\right] \end{array}\right\},$$

Where,

$$E(Z_i^t) = \frac{\dfrac{\hat{w}^t}{\sqrt{2\pi}\hat{\sigma}^t}\exp\left[-\dfrac{(x_i - \hat{\mu}_1^t)^2}{2\hat{\sigma}^{t^2}}\right]}{\dfrac{\hat{w}^t}{\sqrt{2\pi}\hat{\sigma}^t}\exp\left[-\dfrac{(x_i - \hat{\mu}_1^t)^2}{2\hat{\sigma}^{t^2}}\right] + \dfrac{1-\hat{w}^t}{\sqrt{2\pi}\hat{\sigma}^t}\exp\left[-\dfrac{(x_i - \hat{\mu}_2^t)^2}{2\hat{\sigma}^{t^2}}\right]}.$$

M-step:

Maximizing the log-likelihood function, we get:

$$\hat{w}^{t+1} = \frac{\sum_{i=m+1}^{m+n} E(Z_i^t)}{n},$$

$$\hat{\mu}_1^{t+1} = \frac{\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{m+n} E(Z_i^t) x_i}{m + \sum_{i=m+1}^{m+n} E(Z_i^t)},$$

$$\hat{\mu}_2^{t+1} = \frac{\sum_{i=m+1}^{m+n}\left[1 - E(Z_i^t)\right] x_i}{\sum_{i=m+1}^{m+n}\left[1 - E(Z_i^t)\right]},$$

$$(\hat{\sigma}^{t+1})^2 = \frac{\sum_{i=1}^{m}(x_i - \hat{\mu}_1^{t+1})^2 + \sum_{i=m+1}^{m+n}\left\{E(Z_i^t)(x_i - \hat{\mu}_1^{t+1})^2 + \left[1 - E(Z_i^t)\right](x_i - \hat{\mu}_2^{t+1})^2\right\}}{m+n}.$$

By iterations between the E-step and the M-step, the maximum likelihood estimates for all the parameters can be obtained when the sum of the absolute differences between the parameters from two iteration steps < 0.00001, that is,

$$|\hat{\mu}_1^{t+1} - \hat{\mu}_1^{t}| + |\hat{\mu}_2^{t+1} - \hat{\mu}_2^{t}| + |\hat{w}^{t+1} - \hat{w}^{t}| + (\hat{\sigma}^{t+1})^2 - (\hat{\sigma}^{t})^2 | < 0.00001.$$

π can then be obtained as follows:

$$\hat{\pi} = \frac{m + n\hat{w}}{m + n}.$$

4.2.2 Log-normal distribution

We assume the two populations follow different log-normal distributions as follows. The log of Population 1 (the case population) follows the normal distribution with mean of μ_1 and the standard deviation of σ and the log of population 2 (non-case population) follows the normal distribution with mean of μ_2 and the same standard deviation of σ. Therefore,

The case population follows:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_1 - \mu_1)^2}{2\sigma^2}\right]}{x_1},$$

Similarly, the non-case population follows:

$$f(x_2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_2 - \mu_2)^2}{2\sigma^2}\right]}{x_2}.$$

Suppose that we have observed sample 1 of size m (x_1, x_2, … x_m) from population 1 and another sample of size n (x_{m+1}, x_{m+2}, … x_{m+n}) from a mixture of population 1 and population 2, the likelihood of the n+m subjects will be as follows:

$$L(\mu_1, \mu_{2,} \sigma, w, x) =$$

$$\Pi_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_i - \mu_1)^2}{2\sigma^2}\right]}{x_i} \Pi_{i=m+1}^{m+n} \left\{ \begin{array}{l} \frac{w}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_i - \mu_1)^2}{2\sigma^2}\right]}{x_i} \\ + \frac{1-w}{\sqrt{2\pi}\sigma} \frac{\exp\left[-\frac{(\ln x_i - \mu_2)^2}{2\sigma^2}\right]}{x_i} \end{array} \right\}.$$

Using the EM method as shown in Section 4.2.1 for the normal distributions, we get:

$$\hat{w}^{t+1} = \frac{\sum_{i=m+1}^{m+n} E(Z_i^t)}{n},$$

$$\hat{\mu}_1^{t+1} = \frac{\sum_{i=1}^{m} \ln(x_i) + \sum_{i=m+1}^{m+n} E(Z_i^t) \ln(x_i)}{m + \sum_{i=m+1}^{m+n} E(Z_i^t)},$$

$$\hat{\mu}_2^{t+1} = \frac{\sum_{i=m+1}^{m+n} \left[1 - E(Z_i^t)\right] \ln(x_i)}{\sum_{i=m+1}^{m+n} \left[1 - E(Z_i^t)\right]},$$

$$(\hat{\sigma}^{t+1})^2 = \frac{\sum_{i=1}^{m} (\ln x_i - \hat{\mu}_1^{t+1})^2 + \sum_{i=m+1}^{m+n} \left\{ E(Z_i^t)(\ln x_i - \hat{\mu}_1^{t+1})^2 + \left[1 - E(Z_i^t)\right](\ln x_i - \hat{\mu}_2^{t+1})^2 \right\}}{m+n}.$$

4.2.3 Exponential distribution

We assume the two populations follow different exponential distributions as follows. Population 1 (case population) follows the exponential distribution with mean of λ_1 and population 2 (non-case population) follows the exponential distribution with mean of λ_2. Therefore, the case population follows:

$$f(x_1) = \frac{1}{\lambda_1}\exp(-\frac{x_1}{\lambda_1}),$$

Similarly, the non-case population follows:

$$f(x_2) = \frac{1}{\lambda_2}\exp(-\frac{x_2}{\lambda_2}).$$

Suppose that we have observed sample 1 of size m ($x_1, x_2, \ldots x_m$) from population 1 and another sample of size n ($x_{m+1}, x_{m+2}, \ldots x_{m+n}$) from a mixture of population 1 and population 2, the likelihood of the n+m subjects will be as follows:

$$L(\lambda_1, \lambda_{2,} w, x) =$$
$$\Pi_{i=1}^{m}\frac{1}{\lambda_1}\exp(-\frac{x_1}{\lambda_1})\Pi_{i=m+1}^{m+n}\left\{\frac{w}{\lambda_1}\exp(-\frac{x_1}{\lambda_1}) + \frac{1-w}{\lambda_2}\exp(-\frac{x_2}{\lambda_2})\right\}.$$

Using the EM method as shown in Section 4.2.1 for the normal distributions, we get:

$$\hat{w}^{t+1} = \frac{\sum_{i=m+1}^{m+n}E(Z_i^t)}{n},$$

$$\hat{\lambda}_1^{t+1} = \frac{\sum_{i=1}^{m}x_i + \sum_{i=m+1}^{m+n}E(Z_i^t)x_i}{m + \sum_{i=m+1}^{m+n}E(Z_i^t)},$$

$$\hat{\lambda}_2^{t+1} = \frac{\sum_{i=m+1}^{m+n}\left[1 - E(Z_i^t)\right]x_i}{\sum_{i=m+1}^{m+n}\left[1 - E(Z_i^t)\right]}.$$

4.2.4 Normal and exponential distribution

We assume the two populations follow different distributions as follows. Population 1 (case population) follows the normal distribution with mean of μ and the standard deviation of σ and population 2 (non-case population) follows the exponential distribution with mean of λ . Therefore, the case population follows:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_1-\mu\ )^2}{2\sigma^2}\right],$$

Similarly, the non-case population follows:

$$f\ (x_2) = \frac{1}{\lambda}\exp(-\frac{x_2}{\lambda}).$$

Suppose that we have observed sample 1 of size m ($x_1, x_2, \ldots x_m$) from population 1 and another sample of size n ($x_{m+1}, x_{m+2}, \ldots x_{m+n}$) from a mixture of population 1 and population 2, the likelihood of the n+m subjects will be as follows:

$$L(\mu\ ,\sigma,\lambda,w,x) =$$

$$\Pi_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu\ )^2}{2\sigma^2}\right]\Pi_{i=m+1}^{m+n}\left\{\begin{array}{l}\frac{w}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right]\\ +\frac{1-w}{\lambda}\exp\left[-\frac{x_i}{\lambda}\right]\end{array}\right\}.$$

Using the EM method as shown in Section 4.2.1 for the normal distributions, we get:

$$\hat{w}^{t+1} = \frac{\sum_{i=m+1}^{m+n}E(Z_i^t)}{n},$$

$$\hat{\mu}^{t+1} = \frac{\sum_{i=1}^{m}x_i + \sum_{i=m+1}^{m+n}E(Z_i^t)x_i}{m+\sum_{i=m+1}^{m+n}E(Z_i^t)},$$

$$\hat{\lambda}^{t+1} = \frac{\sum_{i=m+1}^{m+n} \left[1 - E(Z_i^t)\right] x_i}{\sum_{i=m+1}^{m+n} \left[1 - E(Z_i^t)\right]},$$

$$(\hat{\sigma}^{t+1})^2 = \frac{\sum_{i=1}^{m} (x_i - \hat{\mu}^{t+1})^2 + \sum_{i=m+1}^{m+n} \left\{E(Z_i^t)(x_i - \hat{\mu}^{t+1})^2\right\}}{m + \sum_{i=m+1}^{m+n} E(Z_i^t)}.$$

CHAPTER 5

5 NUMERICAL RESULTS

5.1 Using numerical algorithm to find the optimal cutoff

To evaluate the performance of the numerical algorithm described in Section 3.2 in finding the optimal cutoff, we compared the cutoffs obtained from the numerical algorithm to the true cutoffs in the following scenarios: 1) both populations follow the normal distribution (Table 2); 2) both populations follow the log-normal distribution (Table 3); 3) both populations follow the gamma distribution (Table 4); and 4) the case population follows the normal distribution and the non-case population follows the gamma distribution (Table 5). For each scenario, the true cutoffs that maximize the correlation were obtained by a grid search of the potential cutoff based on the continuous variable in an increment of 0.01. It is apparent that the cutoffs obtained based on the numerical algorithm are identical to the true cutoffs based on maximizing the correlation between the case status and the protection status defined by the cutoff. It is also noted that different values of π impact more on the cutoffs but minimally on correlations. In addition, we have evaluated the proposed method and numerical algorithm when the case and non-case populations are closer together with lower correlations (detailed data not shown). In the case of two normal distributions with means of 1 and 2 respectively (the same standard deviation of 1) for the case and non-case populations, the cutoffs obtained from the proposed method ranged from 2.0 to 2.5 for different incidence rates and the corresponding correlations ranged from 0.16 to 0.20. Similarly, for exponential distributions with means of 3 and 5 respectively, the obtained cutoffs ranged from 4.8 to 7.5 and the corresponding correlations ranged from 0.15 to 0.20. In all cases, the cutoffs obtained from the numerical algorithm are identical to the true cutoffs as expected. In all of these examples, the numerical algorithm converged very quickly to the optimal cutoffs, usually under 10 iterations.

Table 2 Cutoffs determined for different values of π; normal distributions with $\mu_1 = 2$, $\mu_2 = 5$ and $\sigma = 1$

π	Cutoff obtained from Maximizing correlation		ρ
	True Cutoff	Optimal cutoff from numerical algorithm	
0.2	3.09	3.09	0.84
0.3	3.25	3.25	0.86
0.4	3.38	3.38	0.86
0.5	3.50	3.50	0.87
0.6	3.62	3.62	0.86
0.7	3.75	3.75	0.86
0.8	3.91	3.91	0.84

Table 3 Cutoffs determined for different values of π; log-normal distributions with $\mu_1 = 1$, $\mu_2 = 4$ and $\sigma = 1$

π	Cutoff obtained from Maximizing the correlation		ρ
	True Cutoff	Optimal cutoff from numerical algorithm	
0.2	8.10	8.10	0.84
0.3	9.50	9.50	0.86
0.4	10.81	10.81	0.86
0.5	12.18	12.18	0.87
0.6	13.73	13.73	0.86
0.7	15.63	15.63	0.86
0.8	18.31	18.31	0.84

Table 4 Cutoffs determined for different values of π; gamma distributions with α_1=0.25 and β_1=4 and α_2=4 and β_2=1

π	Cutoff obtained from Maximizing the correlation		ρ
	True Cutoff	Optimal cutoff from numerical algorithm	
0.2	0.85	0.85	0.79
0.3	0.99	0.99	0.78
0.4	1.11	1.11	0.77
0.5	1.24	1.24	0.75
0.6	1.36	1.36	0.73
0.7	1.50	1.50	0.69
0.8	1.66	1.66	0.62

Table 5 Cutoffs determined for different values of π; normal distribution with μ = 1 and σ = 1 and gamma distribution with α=4 and β=1

π	Cutoff obtained from Maximizing the correlation		ρ
	True Cutoff	Optimal cutoff from numerical algorithm	
0.2	1.56	1.56	0.64
0.3	1.85	1.85	0.67
0.4	2.08	2.08	0.69
0.5	2.28	2.28	0.71
0.6	2.47	2.47	0.71
0.7	2.68	2.68	0.71
0.8	2.91	2.91	0.70

5.2 Comparisons of the Proposed Method with the Misclassification Method and the ROC-curve method

The proposed method with numerical algorithm was also compared with the misclassification method and the ROC-curve method assuming that the two populations follow the normal distributions with different parameters. Table 6 and Table 7 summarized the comparison results. The benefit associated with the true negatives and the benefit associated with the true positives were assumed to be 0 in the ROC-curve method. The cutoffs based on the misclassification method and the ROC-curve method were obtained by a grid search (in a 0.01 increment).

It showed that when the incidence rate is 0.5, the proposed method and the misclassification method yield the same cutoffs as expected. However, the two methods give different cutoffs when the incidence rate is not 0.5. In addition, the closer the two distributions (i.e., the difference between the two means of the normal distributions is smaller), the bigger the differences between the cutoffs obtained from the proposed method and the misclassification method are; the further away the incidence rate is from 0.5, the bigger the differences between the cutoffs obtained from the proposed method and the misclassification method are.

The ROC-curve method yields the same cutoffs as those obtained from the misclassification method when the cost is the same for the false negative and the false positive. However, in real-world situations, the cost associated with the false negative is in general greater than that with the false positive. In that situation, the cutoffs obtained from ROC-curve method are higher than those obtained from the other two methods since it puts more weight on minimizing the FNR. On the other hand, if the cost associated with the false negative is less than that with the false positive, the cutoffs obtained from the ROC-curve method are lower than those obtained from the other two methods since it puts more weight on minimizing the FPR. The further away the ratio of the cost associated with the false negative to that with the false positive is from 1, the bigger the differences of the cutoffs between the ROC-curve method and the other two methods are.

The misclassification rates from these three methods were also evaluated (Table 7). In general, the misclassification rates of the proposed method are almost the same as those obtained from the misclassification method. When the cost associated with the false negative is the same as that with the false positive, the misclassification rates from the ROC-curve method are the same as those obtained from the misclassification method. But when the ratio of the cost associated with the false negative to the cost with that of false positive is not 1, the misclassification rates from ROC-curve method are higher than those obtained from the proposed method and the misclassification method in our examples; the further away the ratio is from 1, the higher the misclassification rates are. For example, for two normal distributions with means of 2 and 5 respectively, the misclassification rates for both the proposed method and the misclassification method ranged from 5.0% to 6.7% for different incidence rates between 0.2 and 0.8. The corresponding misclassification rates for the ROC-curve method ranged from 6.7% to 9.5% when the ratio of the two costs is 5, and the misclassification rates increased to a range from 8.1% to 12.2% when the cost ratio increased to 10; and to a range from 9.8% to 16.0% when the cost ratio increased to 20. Lastly, the correlation for these three was evaluated (Table 8). In the examples evaluated, the correlations from the proposed method and the misclassification method are similar. When the cost associated with the false negative is the same as that with the false positive, the correlations from the ROC-curve method are the same as those obtained from the misclassification method. But when the ratio of the cost associated with the false negative to the cost with that of false positive is not 1, the correlations from ROC-curve method are lower than those obtained from the proposed method and the misclassification method in our examples; the further away the ratio is from 1, the lower the correlations are.

Table 6 Comparisons of cutoff points from the three methods (Proposed correlation method, misclassification method and ROC-curve method) when the two populations follow normal distributions

Normal distributions and incidence	Proposed correlation method cutoff	Misclassification method cutoff	ROC-curve method cutoff				
			Ratio of cost of false negative to false positive				
			1:20	1:5	1:1	5:1	20:1
μ_1=2, μ_2=5, σ=1							
π = 0.2	3.09	3.04	2.04	2.50	3.04	3.57	4.04
π = 0.3	3.25	3.22	2.22	2.68	3.22	3.75	4.22
π = 0.4	3.38	3.36	2.37	2.83	3.36	3.90	4.36
π = 0.5	3.50	3.50	2.50	2.96	3.50	4.04	4.50
π = 0.6	3.62	3.64	2.64	3.10	3.64	4.17	4.63
π = 0.7	3.75	3.78	2.78	3.25	3.78	4.32	4.78
π = 0.8	3.91	3.96	2.96	3.43	3.96	4.50	4.96
μ_1=2, μ_2=3.5, σ=1							
π = 0.2	2.26	1.83	-0.17	0.75	1.83	2.90	3.82
π = 0.3	2.45	2.19	0.19	1.11	2.19	3.26	4.18
π = 0.4	2.61	2.48	0.48	1.41	2.48	3.55	4.48
π = 0.5	2.75	2.75	0.75	1.68	2.75	3.82	4.75
π = 0.6	2.89	3.02	1.02	1.95	3.02	4.09	5.02
π = 0.7	3.05	3.31	1.32	2.24	3.31	4.39	5.31
π = 0.8	3.24	3.67	1.68	2.60	3.67	4.75	5.67

Table 7 Comparisons of correlation from the three methods (Proposed correlation method, misclassification method and ROC-curve method) when two populations follow normal distributions

Normal distributions and incidence	Proposed correlation method	Misclassification method	ROC-curve method				
			Ratio of cost of false negative to false positive				
			1:20	1:5	1:1	5:1	20:1
μ_1=2, μ_2=5, σ=1							
π = 0.2	0.050	0.050	0.098	0.067	0.050	0.073	0.139
π = 0.3	0.060	0.060	0.126	0.082	0.060	0.086	0.156
π = 0.4	0.065	0.065	0.145	0.090	0.065	0.093	0.160
π = 0.5	0.067	0.067	0.157	0.095	0.067	0.095	0.157
π = 0.6	0.065	0.065	0.160	0.093	0.065	0.090	0.145
π = 0.7	0.060	0.060	0.156	0.086	0.060	0.082	0.126
π = 0.8	0.050	0.050	0.139	0.073	0.050	0.067	0.098
μ_1=2, μ_2=3.5, σ=1							
π = 0.2	0.165	0.151	0.197	0.181	0.151	0.256	0.507
π = 0.3	0.201	0.194	0.290	0.250	0.194	0.315	0.531
π = 0.4	0.220	0.219	0.375	0.300	0.219	0.336	0.505
π = 0.5	0.227	0.227	0.449	0.330	0.227	0.330	0.449
π = 0.6	0.220	0.219	0.505	0.336	0.219	0.300	0.375
π = 0.7	0.201	0.194	0.531	0.315	0.194	0.250	0.290
π = 0.8	0.165	0.151	0.507	0.256	0.151	0.181	0.197

Table 8 Comparisons of misclassification rates from the three methods (Proposed correlation method, misclassification method and ROC-curve method) when two populations follow normal distributions

Normal distributions and incidence	Proposed correlation method	Misclassification method	ROC-curve method				
			Ratio of cost of false negative to false positive				
			1:20	1:5	1:1	5:1	20:1
μ_1=2, μ_2=5, σ=1							
π = 0.2	0.84	0.84	0.67	0.78	0.84	0.80	0.69
π = 0.3	0.86	0.86	0.70	0.80	0.86	0.81	0.71
π = 0.4	0.86	0.86	0.71	0.82	0.86	0.82	0.72
π = 0.5	0.87	0.87	0.72	0.82	0.87	0.82	0.72
π = 0.6	0.86	0.86	0.72	0.82	0.86	0.82	0.71
π = 0.7	0.86	0.86	0.71	0.81	0.86	0.80	0.70
π = 0.8	0.84	0.84	0.69	0.80	0.84	0.78	0.67
μ_1=2, μ_2=3.5, σ=1							
π = 0.2	0.49	0.47	0.11	0.27	0.47	0.45	0.30
π = 0.3	0.52	0.52	0.15	0.34	0.52	0.45	0.28
π = 0.4	0.54	0.54	0.19	0.39	0.54	0.44	0.26
π = 0.5	0.55	0.55	0.23	0.42	0.55	0.42	0.23
π = 0.6	0.54	0.54	0.26	0.44	0.54	0.39	0.19
π = 0.7	0.52	0.52	0.28	0.45	0.52	0.34	0.15
π = 0.8	0.49	0.47	0.30	0.45	0.47	0.27	0.11

5.3 Finding optimal cutoff adjusting for covariates

In this section, within each of the case population and the non-case population, we have simulated some data for the continuous variable (Y) to be dichotomized and the covariate (X) to be adjusted for assuming bivariate normal distributions with different parameters (including different correlations, different means, and different incidence rates of the cases). All standard deviations are assumed to be 1 for simplicity. The process of generating these data is described in Section 3.6. Using the simulated data, we then identified the cutoffs adjusting for covariates as described in Section 4.

Table 9 presents the cutoffs based on the incidence of 0.5, the means of continuous variable Y for the cases and the non-cases are 1 and 5, respectively. The means of the covariates varied (either the same or different for the case population and the non-case population). The correlation between X and Y within the cases and non-cases also varied from high (0.9), median (0.5) and low (0.1). The unadjusted cutoff is 3.0 with a correlation of 0.95. Under different scenarios, the adjusted cutoffs are all around 3.0 with high correlations, similar to the unadjusted cutoffs. While keeping everything else the same as Table 9, Table 10 presented the results for a reduced incidence rate (0.2). In this situation, the unadjusted cutoff is 2.7 with a correlation of 0.95. The adjusted cutoff ranged from 3.2 to 3.4 with correlations ranged from 0.93 to 0.95. This shows that when incidence rate is low, there is some difference between the adjusted and unadjusted cutoffs while the correlation remains relatively stable. Table 11 presented the results for a higher incidence rate (0.8) while keeping everything else the same as Table 9 and Table 10. In this situation, similar to the case when the incidence is 0.2, there is some difference between the adjusted cutoffs and the unadjusted ones (unadjusted cutoff is 3.3 while the adjusted cutoffs ranged from 2.5 to 2.7). The correlations were around 0.94.

Table 12, Table 13 and Table 14 presented results when the case and the non-case populations were closer (means of the continuous variable Y are 1 and 2.5 respectively with the same standard deviation of 1). Similar findings were true; that is when the incidence rate is 0.5, there is little difference between the adjusted and the unadjusted cutoffs; when the incidence rate is 0.2 or 0.8, there is some difference between the adjusted and the unadjusted cutoffs. Again, in both the adjusted and the unadjusted analyses, the correlations were similar.

Table 9 Cutoffs adjusting for covariates based on different scenarios as compared with cutoff without adjustments

π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.5	1	1	3	1	0.9	5	1	8	1	0.9	3.1	0.95
0.5	1	1	3	1	0.9	5	1	8	1	0.5	3.0	0.95
0.5	1	1	3	1	0.9	5	1	8	1	0.1	3.0	0.94
0.5	1	1	3	1	0.5	5	1	8	1	0.5	3.1	0.95
0.5	1	1	3	1	0.1	5	1	8	1	0.9	3.1	0.96
0.5	1	1	3	1	0.1	5	1	8	1	0.1	3.1	0.94
0.5	1	1	3	1	0.9	5	1	3	1	0.9	3.1	0.95
0.5	1	1	3	1	0.5	5	1	3	1	0.5	3.1	0.95
0.5	1	1	3	1	0.1	5	1	3	1	0.1	3.1	0.94
Cutoff without adjusting for covariates											3.0	0.95

Table 10 Cutoffs adjusting for covariates based on different scenarios as compared to cutoff without adjustment

π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.2	1	1	3	1	0.9	5	1	8	1	0.9	3.4	0.94
0.2	1	1	3	1	0.9	5	1	8	1	0.5	3.4	0.95
0.2	1	1	3	1	0.9	5	1	8	1	0.1	3.4	0.94
0.2	1	1	3	1	0.5	5	1	8	1	0.5	3.3	0.93
0.2	1	1	3	1	0.1	5	1	8	1	0.9	3.3	0.93
0.2	1	1	3	1	0.1	5	1	8	1	0.1	3.2	0.94
0.2	1	1	3	1	0.9	5	1	3	1	0.9	3.4	0.94
0.2	1	1	3	1	0.5	5	1	3	1	0.5	3.3	0.93
0.2	1	1	3	1	0.1	5	1	3	1	0.1	3.2	0.94
Cutoff without adjusting for covariates											2.7	0.95

Table 11 Cutoffs adjusting for covariates based on different scenarios as compared to cutoff without adjustment

π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.8	1	1	3	1	0.9	5	1	8	1	0.9	2.6	0.93
0.8	1	1	3	1	0.9	5	1	8	1	0.5	2.5	0.93
0.8	1	1	3	1	0.9	5	1	8	1	0.1	2.7	0.94
0.8	1	1	3	1	0.5	5	1	8	1	0.5	2.5	0.92
0.8	1	1	3	1	0.1	5	1	8	1	0.9	2.7	0.93
0.8	1	1	3	1	0.1	5	1	8	1	0.1	2.7	0.94
0.8	1	1	3	1	0.9	5	1	3	1	0.9	2.6	0.93
0.8	1	1	3	1	0.5	5	1	3	1	0.5	2.5	0.92
0.8	1	1	3	1	0.1	5	1	3	1	0.1	2.7	0.94
Cutoff without adjusting for covariates											3.3	0.95

Table 12 Cutoffs adjusting for covariates based on different scenarios as compared to cutoff without adjustment

π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.5	1	1	3	1	0.9	2.5	1	8	1	0.9	1.6	0.53
0.5	1	1	3	1	0.9	2.5	1	8	1	0.5	1.9	0.53
0.5	1	1	3	1	0.9	2.5	1	8	1	0.1	1.9	0.51
0.5	1	1	3	1	0.5	2.5	1	8	1	0.5	1.7	0.54
0.5	1	1	3	1	0.1	2.5	1	8	1	0.9	1.7	0.55
0.5	1	1	3	1	0.1	2.5	1	8	1	0.1	1.7	0.54
0.5	1	1	3	1	0.9	2.5	1	3	1	0.9	1.6	0.53
0.5	1	1	3	1	0.5	2.5	1	3	1	0.5	1.7	0.54
0.5	1	1	3	1	0.1	2.5	1	3	1	0.1	1.7	0.54
Cutoff without adjusting for covariates											1.8	0.55

Table 13 Cutoffs adjusting for covariates based on different scenarios as compared to cutoff without adjustment

π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.2	1	1	3	1	0.9	2.5	1	8	1	0.9	2.4	0.47
0.2	1	1	3	1	0.9	2.5	1	8	1	0.5	2.5	0.47
0.2	1	1	3	1	0.9	2.5	1	8	1	0.1	2.4	0.47
0.2	1	1	3	1	0.5	2.5	1	8	1	0.5	2.1	0.47
0.2	1	1	3	1	0.1	2.5	1	8	1	0.1	2.1	0.49
0.2	1	1	3	1	0.1	2.5	1	8	1	0.9	1.9	0.48
0.2	1	1	3	1	0.9	2.5	1	3	1	0.9	2.4	0.47
0.2	1	1	3	1	0.5	2.5	1	3	1	0.5	2.1	0.47
0.2	1	1	3	1	0.1	2.5	1	3	1	0.1	2.1	0.49
Cutoff without adjusting for covariates											1.3	0.49

Table 14 Cutoffs adjusting for covariates based on different scenarios as compared to cutoff without adjustment

Π	Cases					Non-cases					Cutoff	ρ
	μ_{d_y}	σ_{d_y}	μ_{d_x}	σ_{d_x}	ρ_d	μ_{nd_y}	σ_{nd_y}	μ_{nd_x}	σ_{nd_x}	ρ_{nd}		
0.8	1	1	3	1	0.9	2.5	1	8	1	0.9	1.3	0.47
0.8	1	1	3	1	0.9	2.5	1	8	1	0.5	1.5	0.43
0.8	1	1	3	1	0.9	2.5	1	8	1	0.1	1.5	0.43
0.8	1	1	3	1	0.5	2.5	1	8	1	0.5	1.5	0.43
0.8	1	1	3	1	0.1	2.5	1	8	1	0.1	1.5	0.46
0.8	1	1	3	1	0.1	2.5	1	8	1	0.9	1.3	0.49
0.8	1	1	3	1	0.9	2.5	1	3	1	0.9	1.3	0.47
0.8	1	1	3	1	0.5	2.5	1	3	1	0.5	1.5	0.43
0.8	1	1	3	1	0.1	2.5	1	3	1	0.1	1.5	0.44
Cutoff without adjusting for covariates											2.2	0.49

5.4 Using E-M method to estimate the parameters with different distributions

In this section, we used simulation studies to evaluate the performance of EM method, where we first estimate the population parameters and then find the optimal cutoff using the proposed method with the numerical algorithm. In the simulations, we evaluated two scenarios for all three distributions. In the first scenario, we assume that 50 cases were observed and 250 observations were censored at the time of the study termination. We further assume that the 250 censored observations contained a mixture of 100 unobserved cases and 150 non-cases. Therefore, $\omega=0.4$ and $\pi=0.5$. In the second scenario, we assume that 25 cases were observed and 275 observations were censored at the time of the study termination. Out of the 275 observations, 35 were unobserved cases and 240 were non-cases. Therefore, $\omega=0.1$ and $\pi=0.2$.

In the simulation study for the first scenario, we generated 1000 samples of 50 known cases and 250 observations consisting of a mixture of 100 unobserved cases and 150 non-cases. For each sample, parameters were estimated using the EM method and the optimal cutoff was then obtained via the proposed method with numerical algorithm. Simulations for the second scenario were done similarly. The means of the parameter estimates and the optimal cutoffs from the 1000 runs were then presented in Table 15. It can be seen that the EM parameter estimates for the distributions are very close to the true parameters, and the incidence rate (π) of cases was also correctly estimated in every scenario. Subsequently, in every case, the cutoffs were very close to the true cutoffs.

It is argued that this method could be used even when no cases have been observed. We have investigated the performance of the EM method when no cases have been observed under a variety of scenarios with respect to different parameters. In the cases when the EM method does not work well, we investigated whether observing some cases would improve the performance in parameter estimation. Again, simulation studies were used for these evaluations. Similar to the simulations described above, in each of these simulations, 200 runs were generated and the means of the parameter estimates from all the 200 runs were presented to be compared with the true underlying parameters. In each run for all the scenarios, a sample size of 500 mixed cases and non-cases was used with varying number of observed cases.

Table 16 presents the simulation results of the EM method when no cases have been observed. In the simulations, we have a mix of 500 observations with incidence rates of 0.5, 0.4, 0.3 and 0.2 respectively. Within each incidence rate, we have assumed two normal distributions with different separations between the two populations as indicated by the differences of the means between the two populations (all with the same standard deviation of 1). It is noted from these studies that, when the incidence rate is 0.5, the performance is poorest even if the two populations are far apart (e.g., $\mu_1 = 1$ and $\mu_2 = 5$). However, when incidence rate decreases (e.g., 0.4), the performance improved a lot especially if the two populations were farther apart. Within each incidence rate, the performance became poorer when the distance between the two populations is closer. Tables 17, 18, 19 and 20 present the simulation results when there are some known cases. In these scenarios, the performance of the EM method was much improved, especially for the two populations that are closer together (e.g., the difference of the means between the two populations is ≤ 2). However, it does not appear that the performance further improved much with the increased number of observed cases. That argues for the importance of having observed some cases in a real-world setting before applying the EM method to estimate parameters. In advance, one may not know the assumptions about the true incidence rate and the distance between the two populations even if the EM method theoretically would work for the situations when no cases have been observed.

Table 15 Parameter estimates based on EM Method

Distributions	$\hat{\mu}_1$ (or $\hat{\lambda}_1$)	$\hat{\mu}_2$ (or $\hat{\lambda}_2$)	$\hat{\sigma}^2$	$\hat{\omega}$	$\hat{\pi}$	Optimal cutoff based on EM estimated parameters	True cutoff
50 known Cases and a mix of 100 unobserved cases and 150 non-cases (ω=0.4 and π=0.5)							
normal (μ_1=2, μ_2=5 and σ=1)	2.0	5.0	1.0	0.4	0.5	3.50	3.50
Lognormal (μ_1=1, μ_2=4 and σ=1)	1.0	4.0	0.99	0.4	0.5	12.21	12.18
exponential (λ_1=1 and λ_2=5)	1.0	5.0	-	0.4	0.5	2.46	2.45
25 known cases and a mix of 35 unobserved cases and 240 non-cases (ω=0.1 and π=0.2)							
normal (μ_1=2, μ_2=5 and σ=1)	2.0	5.0	0.99	0.1	0.2	3.10	3.09
lognormal (μ_1=1, μ_2=4 and σ=1)	1.0	4.0	0.99	0.1	0.2	8.22	8.10
exponential (λ_1=1 and λ_2=5)	1.0	5.0	-	0.1	0.2	1.77	1.76

Table 16 Parameter estimates based on 0 known cases

and 500 mixture of unobserved cases and non-cases

Parameters with two normal Distributions	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\pi}$
$\pi=0.5$, $\mu_1=1$, $\mu_2=5$, $\sigma=1$	2.0	4.0	3.04	0.3
$\pi=0.5$, $\mu_1=1$, $\mu_2=3$, $\sigma=1$	1.4	2.5	1.49	0.3
$\pi=0.5$, $\mu_1=1$, $\mu_2=2$, $\sigma=1$	0.9	1.7	1.08	0.2
$\pi=0.5$, $\mu_1=1$, $\mu_2=1.8$, $\sigma=1$	0.8	1.6	1.02	0.2
$\pi=0.5$, $\mu_1=1$, $\mu_2=1.5$, $\sigma=1$	0.7	1.4	0.93	0.2
$\pi=0.5$, $\mu_1=1$, $\mu_2=1.3$, $\sigma=1$	0.5	1.3	0.90	0.3
$\pi=0.4$, $\mu_1=1$, $\mu_2=5$, $\sigma=1$	1.0	5.0	1.00	0.4
$\pi=0.4$, $\mu_1=1$, $\mu_2=3$, $\sigma=1$	1.1	2.9	1.06	0.4
$\pi=0.4$, $\mu_1=1$, $\mu_2=2$, $\sigma=1$	0.8	1.8	1.04	0.2
$\pi=0.4$, $\mu_1=1$, $\mu_2=1.8$, $\sigma=1$	0.8	1.8	1.04	0.2
$\pi=0.4$, $\mu_1=1$, $\mu_2=1.5$, $\sigma=1$	0.7	1.7	1.00	0.2
$\pi=0.4$, $\mu_1=1$, $\mu_2=1.3$, $\sigma=1$	0.6	1.5	0.93	0.2
$\pi=0.3$, $\mu_1=1$, $\mu_2=5$, $\sigma=1$	1.0	5.0	0.99	0.3
$\pi=0.3$, $\mu_1=1$, $\mu_2=3$, $\sigma=1$	1.0	3.0	1.00	0.3
$\pi=0.3$, $\mu_1=1$, $\mu_2=2$, $\sigma=1$	0.8	1.9	1.02	0.2
$\pi=0.3$, $\mu_1=1$, $\mu_2=1.8$, $\sigma=1$	0.8	1.7	0.98	0.2
$\pi=0.3$, $\mu_1=1$, $\mu_2=1.5$, $\sigma=1$	0.7	1.5	0.93	0.2
$\pi=0.3$, $\mu_1=1$, $\mu_2=1.3$, $\sigma=1$	0.6	1.4	0.90	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=5$, $\sigma=1$	1.0	5.0	1.00	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=3$, $\sigma=1$	1.0	3.0	1.00	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=2$, $\sigma=1$	0.8	2.0	0.98	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=1.8$, $\sigma=1$	0.8	1.8	0.96	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=1.5$, $\sigma=1$	0.8	1.6	0.92	0.2
$\pi=0.2$, $\mu_1=1$, $\mu_2=1.3$, $\sigma=1$	0.6	1.4	0.89	0.2

Table 17 Parameter estimates based on 20 known cases

and 500 mixture of unobserved cases and non-cases

Parameters with two normal Distributions	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\pi}$
π =0.5, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.05	0.5
π =0.5, μ_1=1, μ_2=3 and σ=1	1.0	2.9	1.05	0.5
π =0.5, μ_1=1, μ_2=2 and σ=1	1.0	1.9	1.03	0.4
π =0.5, μ_1=1, μ_2=1.8 and σ=1	1.0	1.7	1.00	0.4
π =0.5, μ_1=1, μ_2=1.5 and σ=1	1.0	1.5	0.96	0.4
π =0.5, μ_1=1, μ_2=1.3 and σ=1	1.0	1.3	0.94	0.4
π =0.2, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.2
π =0.2, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.2
π =0.2, μ_1=1, μ_2=2 and σ=1	1.0	2.0	0.97	0.2
π =0.2, μ_1=1, μ_2=1.8 and σ=1	1.0	1.9	0.95	0.2
π =0.2, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.93	0.3
π =0.2, μ_1=1, μ_2=1.3 and σ=1	1.0	1.4	0.93	0.4

Table 18 Parameter estimates based on 50 known cases

and 500 mixture of unobserved cases and non-cases

Parameters with two normal Distributions	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\pi}$
π =0.5, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.5
π =0.5, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.02	0.5
π =0.5, μ_1=1, μ_2=2 and σ=1	1.0	2.0	1.02	0.4
π =0.5, μ_1=1, μ_2=1.8 and σ=1	1.0	1.8	1.00	0.4
π =0.5, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.98	0.4
π =0.5, μ_1=1, μ_2=1.3 and σ=1	1.0	1.4	0.96	0.4
π =0.2, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.2
π =0.2, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.2
π =0.2, μ_1=1, μ_2=2 and σ=1	1.0	2.0	0.98	0.2
π =0.2, μ_1=1, μ_2=1.8 and σ=1	1.0	1.9	0.96	0.2
π =0.2, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.95	0.3
π =0.2, μ_1=1, μ_2=1.3 and σ=1	1.0	1.5	0.95	0.4

Table 19 Parameter estimates based on 75 known cases

and 500 mixture of unobserved cases and non-cases

Parameters with two normal Distributions	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\pi}$
π =0.5, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.5
π =0.5, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.5
π =0.5, μ_1=1, μ_2=2 and σ=1	1.0	2.0	1.01	0.5
π =0.5, μ_1=1, μ_2=1.8 and σ=1	1.0	1.8	1.00	0.5
π =0.5, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.98	0.5
π =0.5, μ_1=1, μ_2=1.3 and σ=1	1.0	1.4	0.97	0.4
π =0.2, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.2
π =0.2, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.2
π =0.2, μ_1=1, μ_2=2 and σ=1	1.0	2.0	0.98	0.2
π =0.2, μ_1=1, μ_2=1.8 and σ=1	1.0	1.9	0.97	0.2
π =0.2, μ_1=1, μ_2=1.5 and σ=1	1.0	1.7	0.94	0.3
π =0.2, μ_1=1, μ_2=1.3 and σ=1	1.0	1.5	0.95	0.4

Table 20 Parameter estimates based on 200 known cases and 500 mixture of unobserved cases and non-cases (π =0.5) and 100 known cases and 500 mixture of unobserved cases and non-cases (π =0.2)

Parameters with two normal Distributions	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\pi}$
π =0.5, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.5
π =0.5, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.5
π =0.5, μ_1=1, μ_2=2 and σ=1	1.0	2.0	0.99	0.5
π =0.5, μ_1=1, μ_2=1.8 and σ=1	1.0	1.8	0.99	0.5
π =0.5, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.98	0.5
π =0.5, μ_1=1, μ_2=1.3 and σ=1	1.0	1.5	0.97	0.6
π =0.2, μ_1=1, μ_2=5 and σ=1	1.0	5.0	1.00	0.2
π =0.2, μ_1=1, μ_2=3 and σ=1	1.0	3.0	1.00	0.2
π =0.2, μ_1=1, μ_2=2 and σ=1	1.0	2.0	0.97	0.2
π =0.2, μ_1=1, μ_2=1.8 and σ=1	1.0	1.9	0.96	0.2
π =0.2, μ_1=1, μ_2=1.5 and σ=1	1.0	1.6	0.95	0.3
π =0.2, μ_1=1, μ_2=1.3 and σ=1	1.0	1.5	0.95	0.4

5.5 Application to a varicella vaccine example

We further applied the proposed methodologies to a varicella vaccine example. Varicella is a highly contagious disease caused by the varicella-zoster virus. It primarily affects preschoolers and school-aged children. A live attenuated varicella vaccine, VARIVAX™, was licensed in the United States in 1995 by Merck and Co. The data used in this report were reported elsewhere [Chan et al., 2002]. Briefly, these data were based on the studies of 7 years of follow-up in 1087 healthy children who were enrolled in two clinical studies during 1991 to 1992 in the United States. In these two studies, the children received one dose of varicella vaccine when they were between 1 and 12 years of age. Postvaccination antibody titers were measured at 6 weeks after vaccination. Subsequent breakthroughs were ascertained by asking the parents or guardians of each child to immediately report cases of chickenpox-like illness to the study site and to complete a questionnaire annually to capture any chickenpox cases that were not reported immediately. Out of 1087 children, 66 developed varicella during the follow-up. It is apparent that children who developed disease had lower antibody titers 6 weeks postvaccination (mean of log antibody titers and standard deviation of 1.8 ± 1.2) than those who did not develop disease (2.6 ± 1.1). Using the proposed method and numerical method in Section 2 with $\pi = 0.06$, $\mu_1 = 1.8$, $\mu_2 = 2.6$ and $\sigma = 1.1$, and the optimal cutoff was determined to be 1.6 with a 95% CI of [1.3, 2.0] (please note that the confidence interval was derived from the bootstrapping technique described in Section 3.4). This cutoff is essentially the same as what has been previously determined and used in clinical studies [Li et al., 2002].

Age affects the incidence of varicella, i.e., the older the child is; the less likely the child will contract the disease. Based on the data, in the case population, the mean age is around 2.7 years with a standard deviation of 1.2 and the correlation between age and the log antibody titer is around -0.23. In the non-case population, the mean age is around 4.3 years with a standard deviation of 2.6 and correlation between age and the log antibody titer is around -0.01. Adjusting for age, the cutoff is determined to be 1.8 which is slightly higher than the unadjusted cutoff of 1.6.

In addition, we have used the EM method to estimate parameters in different scenarios based on when we cut the data and obtained the disease case status (Table 21). For example, by the end of Year 1, only 2 children had developed disease and the rest was the mixture of unobserved cases and non-cases; the parameter estimates from the EM method are quite different from the parameter estimates obtained using the full data set. However, with increased information, the parameter estimates from EM method become closer to those from the full data. From Year 4 onward, the cutoff obtained using the EM method started to approach the cutoff obtained from the full data set. This observation is consistent with the results we have demonstrated in Section 5.4 though it was shown that EM method should have worked better even with fewer known cases (e.g., starting at Year 2 or Year 3). This may be due to the fact that the logarithm of antibody titers for cases and non-cases may not be exactly normally distributed while the simulated data in Section 5.4 were taken from theoretical normal distributions. Once the parameter estimates are obtained, the cutoff can be obtained using the proposed method with the numerical algorithm. It is reassuring to note that the EM methods yielded the same optimal cutoff as the numerical algorithm.

Table 21 Parameter estimate based on cases observed through different time points assuming normal distributions based on a real-world varicella example

Time Period	# of cases	# of mix	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$\hat{\omega}$	$\hat{\pi}$	cutoff
Year 1	2	1085	-0.2	2.6	1.1	1.7%	1.9%	-0.2
Year 2	13	1074	1.0	2.6	1.1	2.9%	4.1%	0.7
Year 3	26	1061	1.2	2.6	1.1	2.8%	5.1%	0.9
Year 4	49	1038	1.5	2.6	1.1	1.6%	6.0%	1.3
Year 5	60	1027	1.8	2.6	1.2	0.2%	5.7%	1.6
Year 6	63	1024	1.8	2.6	1.2	0.2%	6.0%	1.6
Year 7	66	1021	1.8	2.6	1.2	0.1%	6.2%	1.6

CHAPTER 6

6 CONCLUDING REMARKS AND FUTURE RESEARCH

It is argued that continuous variables capture more information than the binary variables based on a cutoff point. For example, if the continuous variable is normally distributed, the Pitman efficiency of the sign test compared with the t-test is about 64 percent ($\frac{2}{\pi}$). This is deemed to be the optimistic analogy to the loss in practice in dichotomizing when the median was used as the cutoff point. In addition, it is also sometimes criticized that the cutoff may be arbitrarily set (Senn 2003). Because of these reasons, determining a cutoff point has not been extensively studied. However, in some cases, a binary variable would constitute a more clinically meaningful and robust endpoint that is easy to communicate; thus, it is sometimes desirable in a clinical setting to dichotomize a continuous variable. In this dissertation, we proposed to define the cutoff point based on maximizing the correlation between the binary variable defined by population and the binary variable defined by the cutoff point. We have also developed a numerical algorithm to solve for the cutoff point. This method is straightforward and can be easily implemented. In addition, this approach would alleviate some of the concerns about the arbitrariness of the cutoff point since the cutoff aims to predict the disease occurrence. As demonstrated in the dissertation, the numerical algorithm correctly finds the optimal cutoff value that maximizes the correlation under a variety of distribution assumptions. In the varicella example, it is very reassuring to note that the cutoff based on the proposed method with the numerical algorithm was very close to what has been used in the clinical trials studying immunogenicity response in varicella vaccine. It should be noted that if the same study used to find the cutoff is also used to evaluate the performance of the cutoff, the assessment may be too optimistic (Linnet and Brandt, 1986; Leeflang et al., 2008). Therefore, it might be useful to validate the cutoff using a different study, if possible.

The proposed method was also compared with the ROC-curve method aimed at minimizing a specific cost function and the misclassification method aimed at minimizing the misclassification rate. In general, the three methods may give different results since they use different criteria to select the cutoff. ROC-curve method involves financial interpretation of the cutoff. The cost associated with the false positive and the false negative could be different for different diseases and also could be subjective. Therefore, the cutoff obtained from the ROC-curve method could vary depending on the choice of the cost function and at times be much different from those obtained from the other two methods. The cutoff obtained from the proposed method could be different and sometimes much different from that obtained from the ROC-curve method or the misclassification method. Since the goal is to identify a biomarker that correlates with disease protection, the proposed method yields a cutoff that may be the closest to the true correlate of protection as it maximizes the correlation between the disease status and the biomarker.

In defining the optimal cutoff using the numerical algorithm, we have assumed that all the cases have been observed. However, in real-world scenarios, the case-status may not be completely observed in the study. For this reason, we have applied the EM method to first estimate the population parameters and then use the proposed method with the numerical algorithm to find the optimal cutoff. The performance of the EM procedure has been confirmed through simulation studies. Furthermore, when applied to the varicella example, the optimal cutoff from the EM method is identical to that from the numerical algorithm. It is argued that using EM method, it is not necessary to have observed any case. However, as demonstrated in Section 5.3, the method performs poorly when the incidence rate is at 0.5 and/or the two populations are close together when no cases have been observed. The performance improved a lot even when only a small fraction of the cases have been observed, especially when the two populations are further apart. Furthermore, in practice, if we have not observed any case, it is difficult to know whether the prior assumptions about the incidence rate of disease or the distance about the two populations are correct. Therefore, using EM method to estimate parameters may provide very biased results when no case has been observed. Thus, it is strongly encouraged that we observe some cases to be comfortable about the parameter estimates obtained using EM method.

Finally, the method of defining cutoff based on maximizing the correlation can be extended to incorporate the covariates that are associated with the continuous variable and/or the case status. This is accomplished by modeling the FPR and the FNR adjusting for the covariates. Once the adjusted FPR and FNR are obtained, we can plug them into the correlation equation and find the cutoff point that maximizes the correlation. Through simulation studies, it appears that the correlations between the covariate and the continuous variable within case and within non-case do not impact the cutoff. In addition, when the incidence is 0.5, the adjusted cutoff does not differ much from the unadjusted one. However, when the incidence is different from 0.5, the adjusted cutoff point is more different from the unadjusted one.

The proposed method can be extended in many ways. First, instead of determining only one cutoff point, multiple cutoff points can be determined to categorize the continuous variable so that loss of power can be minimized. In addition, in assessing the proposed method with the misclassification method or the ROC-curve method, it is possible to derive the exact relationship between the three methods so that it can be certain under which condition(s) the three methods will be exactly the same.

REFERENCES

Anderson, P. (1984). The protective level of serum antibodies to the capsular polysaccharide of H. infuenzae type b. Journal of Infectious Disease, 140, 1034.

Chan, I.S.F., Li, S., Matthews, H., Chan, C.Y., Vessey, S.J.R., Sadoff, J.C., Heyse, J.F. (2002). Use of statistical models for evaluating antibody response as a correlate of protection against varicella. Statistics in Medicine, 21, 3411-3430.

Cochran W.G., Hopkins C.E. (1961). Some classification problems with multivariate qualitative data. Biometrics, 171, 10-32.

Crosnier J., Jungers P., Courouce A.M., Laplanche A., Benhamou E., Degos F., Lacour B., Prunet P., Cerisier Y., Guesry P. (1981). Randomized placebo-controlled trial of hepatitis B surface antigen vaccine in French haemodialysis units, I, Medical staff. Lancet, 454-459.

Crosnier J., Jungers P., Courouce A.M., Laplanche A., Benhamou E., Degos F., Lacour B., Prunet P., Cerisier Y., Guesry P. (1981). Randomized placebo-controlled trial of hepatitis B surface antigen vaccine in French haemodialysis units, II, Haemodialysis patients. Lancet, 797-800.

Damaraju, L. (2009). Screening Variables to Determine Binary Responses to Maximize the Asymptotic Relative Efficiency. J. P. J. Biostatistics, 3, 215-224.

Dempster A.P., Laird N.M., Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 39 (1), 1-38.

Kayhty H., Peltola H., Karanko V., Makela P.H. (1983). The protective level of serum antibodies to the capsular polysaccharide of haemophilus influenza type b. Journal of Infectious Disease, 147 (6), 1100.

Keith P. K., Hendrik J. K., John B. R., Nancy N. L. C. (1996). Immunogenicity, efficacy and serological correlate of protection of Salmonella typhi Vi capsular polysaccharide vaccine three years after immunization. Vaccine, 14 (8), 435-438.

Klugman, K.P., Koornhof, H.J., Robbins, J.B., Le Cam, N.N. (1996). Immunogenicity, efficacy and serological correlate of protection of Salmonella typhi Vi capsular polysaccharide vaccine three years after immunization. Vaccine, 14(5), 435-438.

Kohberger R., Xie F. (1998). A theoretical framework for the relationship of correlates of protection, population immunogenicity, and vaccine efficacy. Paper presented at the Joint Statistical Meetings, Dallas, TX.

Kuter, B.J., Weibel, R.E., Guess, H.A., Matthews, H., Morton, D.H., Neff, B.J., Provost, P.J., Watson, B.A., Starr, S.E., Plotkin, S.A. (1991). Oka/Merck varicella vaccine in healthy children, Final report of a 2-year efficacy study and 7-year follow-up studies. Vaccine, 9(9), 643-647.

Lakshmi, D. (1995). Inferences in the Presence of Misclassification Errors. Unpublished Ph. D. Dissertation, Temple University.

Leeflang M.M.G., Moons K.G.M., Reitsma J.B., Zwinderman A.H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values, mechanisms, magnitude, and solutions. Clinical Chemistry, 54(4), 729-737.

Lewis J.A. (2004). In defense of the dichotomy. Pharmaceutical Statistics, 3, 77-79.

Li, S., Chan, I.S.F., Matthews, H., Heyse, J.F., Chan, C.Y., Kuter, B.J., Kaplan, K.M., Vessey, S.J.R., Sadoff, J.C. (2002). Childhood vaccination against varicella, Inverse relationship between 6-week postvaccination varicella antibody response and likelihood of long-term breakthrough infection. Pediatric Infectious Disease Journal, 21,337-42.

Linnet, K., Brandt, E. (1986). Assessing diagnostic tests once an optimal cutoff point has been selected. Clinical Chemistry, 32, 1341-1346.

Metz, C.E. (1978). Basic principles of ROC analysis, Seminars in nuclear medicine, Vol. VIII, No.4.

Metz, C.E (1986). ROC methodology in radiologic imaging. Investigational Radiology, 21,720-33.

Padgett, L. (2011). Practical Statistical Methods, A SAS Programming Approach. Chapman&Hall/CRC Press.

Senn, S. (2003). Disappointing dichotomies. Pharmaceutical Statistics, 2, 239-240.

Senn, S, Julious, S. (2009). Measurement in clinical trials, a neglected issue for statisticians? Statistics in Medicine, 28, 3189-3209.

Stein, R. M. (2005), The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing, Journal of Banking & Finance, 29, 1213-1236.

Vessey, S.J.R., Chan, C.Y., Kuter, B.J., Kaplan, K.M., Waters, M., Kutzler, D.P., Carfagno, P.A., Sadoff, J.C., Heyes, J.R., Matthews, H., Li, S., Chan, I.S.F. (2001). Childhood vaccinatgion against varicella: Persistence of antibody duration of protection, and vaccine efficacy. Journal of Pediatrics, 139 (2), 297-304.

Weibel, R.E., Neff, B.J., Kuter, B.J., Guess, H.A., Rothenberger, C.A., Fitzgerald, A.J., Connor, K.A., Mclean, A.A., Hilleman, M.R., Buynak, E.B., Scolnick, E.M. (1984). Live attenuated varicella virus vaccine. Efficacy trial in healthy children. New England Journal of Medicine, 310 (22), 1409-1415.

White, C.J., Kuter, B.J., Ngai, A., Hildebrand, C.S., Isganitis, K.L., Patterson, C.M., Capra, A., Miller, W.J., Krah, D.L., Provost, P.J. (1992). Modified cases of chickenpox after vaccination: Correlation of protection with antibody response. Pediatric Infectious Disease Journal, 11, 19-23.

Yue S., Xie M. (2010). A note on dichotomization of continuous response variable in the presence of contamination and model misspecification. Statistics in Medicine, 29, 2200-2214.

Zhou X.H., Obuchowski N.A., Mcclish D.K. (2002). Statistical methods in diagnostic medicine. New York: John Wiley & Sons, Inc.

APPENDIX: SAS PROGRAMS

```
***************************************************************;
* FUNCTION:    This program is to graph the distribution of the case and non-case; graph
the correlation and misclassification rate and false negative and false positive rate based on
different cutoffs
***************************************************************;

options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro pdf (mu1=5, /* mean of non-case normal distribution */
            mu2=2, /* mean of case normal distribution    */
            std=2); /* std of the two normal distributions  *

data a;
  do i=-5 to 15 by 0.1;
    pdf2 = 1/(sqrt(2*3.1415926)*&std)*exp(-(i-&mu1)**2/(2*&std**2));
         pdf1 = 1/(sqrt(2*3.1415926)*&std)*exp(-(i-&mu2)**2/(2*&std**2));
    output;
  end;
run;

symbol1 v=star   color=green;
symbol2 v=circle  color=blue;
axis1 offset=(1,1)
     color=blue
     label=none
;
axis2 label=('Cutoff');
legend1 label=(position=(top left)
             ' ')
       shape=symbol(0.5,0.5)
       value=(tick=1 justify=c 'Diseased'
            tick=2 justify=c 'Non-diseased');
proc gplot data=a;
  plot pdf1*i pdf2*i / overlay vaxis=axis1 haxis=axis2 legend=legend1;
title1 'PDF for diseased and nondiseased for normal distribution';
title2 'mu1=5, mu2=2, std=2';
run;

%mend;

%pdf;
run;

%macro corr(incid=, /* incidence */
        mu1=5,  /* mean of non-case normal distribution */
        mu2=2,  /* mean of case normal distribution */
        std=2   /* std of the two normal distribution */);
data b;
  pai = &incid;
```

```
  do cutoff=-5 to 15 by 0.2;
  fp = cdf("normal", cutoff, &mu1, &std);
  fn = 1 - cdf("normal", cutoff, &mu2, &std);
  Exy = pai*(1-fn);
  Ex=pai*(1-fp-fn)+fp;
  Ey=pai;
  var1=pai-pai**2;
  var2=ex-ex**2;
  corr=(Exy-(Ex)*EY)/sqrt(var1*var2);
  MR = pai*fn+(1-pai)*fp;
  output;
  end;
run;

symbol1 v=star    color=green;
symbol2 v=circle  color=blue;
symbol3 v=dot     color=red;
axis1 order=(0 to 1.0 by 0.1)
      offset=(1,1)
      color=blue
      label=none
;
axis2 label=('Cutoff');
legend1 label=(position=(top left)
              ' ')
        shape=symbol(0.5,0.5)
        value=(tick=1 justify=c 'Correlation'
               tick=2 justify=c 'Misclassification'
              justify=c 'Rate');
legend2 label=(position=(top left)
              ' ')
        shape=symbol(0.5,0.5)
        value=(tick=1 justify=c 'False'
              justify=c 'Negative'
       tick=2 justify=c 'False'
              justify=c 'Postive'
       tick=3 justify=c 'Correlation');

proc gplot data=b;
  plot (corr mr)*cutoff / overlay vaxis=axis1 haxis=axis2 legend=legend1;
  title "normal: mu1=5, mu2=2, std=2, incidence of disease &incid";
run;

proc gplot data=b;
  plot (fn fp corr)*cutoff / overlay vaxis=axis1 haxis=axis2 legend=legend2;
run;

%mend;

%corr(incid=0.5);
```

```
*************************************************************;
* FUNCTION:     This program is to find the true cutoff
*************************************************************;

*************************************************************;
*---define input library ---*;
*************************************************************;
options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro corr(incid=, /* incidence */
                 mu1=5,  /* mean of non-case normal distribution */
                mu2=2,  /* mean of case normal distribution */
                std=2);   /* std of the two normal distribution */

data b;
  pai = &incid;
  retain corr1 0;
  do cutoff=1 to 5 by 0.01;
  fp = cdf("normal", cutoff, &mu1, &std);
  fn = 1 - cdf("normal", cutoff, &mu2, &std);
  Exy = pai*(1-fn);
  Ex=pai*(1-fp-fn)+fp;
  Ey=pai;
  var1=pai-pai**2;
  var2=ex-ex**2;
  corr=(Exy-(Ex)*EY)/sqrt(var1*var2);
  mis_cls=pai*fn+(1-pai)*fp;
  if corr>corr1 then corr1=corr;
  output;
  end;
run;

data b;
  set b;
  if corr1=corr;
run;

proc sort data=b; by pai cutoff;
data b;
  set b;
  by pai cutoff;
  if last.pai;
run;

data temp;
  set temp b;
run;

%mend;
```

```
data temp;
  set _null_;
run;

%corr(incid=0.2);
```

```
***********************************************************;
* FUNCTION:     This program is to find the cutoff based on numerical algorithm
***********************************************************;
options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro num(incid=, /* incidence */
               mu1=,   /* mean for case normal distribution */
              mu2=,   /* mean for non-case normal distribution */
              std=2); /* std of the two normal distributions */


data a;
  pai = &incid;
  mu1 = &mu1;
  mu2 = &mu2;
  std=  &std;
  x=0.1;
  cutoff=2.6;
  diff = abs(cutoff-x);
  count=0;

  do until (diff<0.00001);
  count=count+1;
  x=cutoff;
  a = probnorm((x-mu1)/std) - probnorm((x-mu2)/std);
  c = pai*a+probnorm((x-mu2)/std);
  b = sqrt(c*(1-c));
  da = 1/(sqrt(2*3.1416)*std)*exp(-(x-mu1)**2/(2*std**2))
    - 1/(sqrt(2*3.1416)*std)*exp(-(x-mu2)**2/(2*std**2));
  dc = pai*da+1/(sqrt(2*3.1416)*std)*exp(-(x-mu2)**2/(2*std**2));
  db = (1-2*c)/(2*b)*dc;
  daa = (x-mu2)/(sqrt(2*3.1416)*std**3)*exp(-(x-mu2)**2/(2*std**2))
      -(x-mu1)/(sqrt(2*3.1416)*std**3)*exp(-(x-mu1)**2/(2*std**2));
  dcc = pai*daa-(x-mu2)/(sqrt(2*3.1416)*std**3)*exp(-(x-mu2)**2/(2*std**2));
  dbb = 1/(2*b**2)*(-2*dc*b-(1-2*c)*db)*dc+(1-2*c)/(2*b)*dcc;
  first = (b*da - a*db)/(b**2);
  second = (b**2*daa - a*b*dbb-2*b*da*db+2*a*db**2)/(b**3);
  cutoff = (x*second - first)/second;
  diff = abs(cutoff-x);
  a1 = probnorm((cutoff-mu1)/std) - probnorm((cutoff-mu2)/std);
  c1 = pai*a1+probnorm((cutoff-mu2)/std);
  b1 = sqrt(c1*(1-c1));
  corr = sqrt(pai*(1-pai))*a1/b1;
  fn = 1-probnorm((cutoff-mu1)/std);
  fp = probnorm((cutoff-mu2)/std);
  mis_rate = pai*fn+(1-pai)*fp;
  log_nor = exp(cutoff);
  output;
  end;
run;

%mend;
%num(incid=0.2, mu1=2, mu2=5);
```

```
****************************************************************;
* FUNCTION:     Estimate the parameters using EM method

****************************************************************;

libname in 'h:\personal\dissertation';

options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro em (n_know=, n_mixed=, pai=, mu1=, mu2=, std=);

/*** create known disease ***/
 data dis_know;
  do i=1 to &n_know;
   x = &mu1 + &std*rannor(1000000);
        x2 = x**2;
  output;
  end;
 run;

/*** create a mixed sample of disease and non-disease with an incidence of pai ***/;

 data mixed;
  do i=1 to &n_mixed;
   a = RANBIN(200,1,&pai);
        if a=1 then x = &mu1 + &std*rannor(2000000);
        else if a=0 then x = &mu2 + &std*rannor(3000000);
        output;
  end;
run;

 proc means data=dis_know noprint;
  var x x2;
       output out=a n=m sum=sumxdis sumx2dis;
 run;

 data mixed;
  set mixed;
  keep x;
 run;

 proc transpose data=mixed out=mixed2 prefix=x;
 run;

 data mixed2;
   merge a mixed2;
        drop _type_ _freq_ _name_;
 run;

 data mixed2;
```

```
   set mixed2;

         /*** initialize the parameters ***/

   sigma_t=0.01;
         mu1_t=0.5;
         mu2_t=1;
         pai_t=0.1;
         diff=1;

  do until (diff<0.00001);
         array z1 {*} z1_1-z1_&n_mixed;
         array z0 {*} z0_1-z0_&n_mixed;
         array x  {*} x1-x&n_mixed;
         array xz1 {*} xz1_1-xz1_&n_mixed;
         array xz0 {*} xz0_1-xz0_&n_mixed;
         array x2z1 {*} x2z1_1-x2z1_&n_mixed;
         array x2z0 {*} x2z0_1-x2z0_&n_mixed;

         do i=1 to &n_mixed;

         z1 {i} = (exp(-(x{i}-mu1_t)**2/(2*sigma_t))*pai_t)/
        (exp(-(x{i}-mu1_t)**2/(2*sigma_t))*pai_t
        +exp(-(x{i}-mu2_t)**2/(2*sigma_t))*(1-pai_t));
         z0 {i} = (exp(-(x{i}-mu2_t)**2/(2*sigma_t))*(1-pai_t))/
        (exp(-(x{i}-mu1_t)**2/(2*sigma_t))*pai_t
        +exp(-(x{i}-mu2_t)**2/(2*sigma_t))*(1-pai_t));

   xz1{i} = z1{i}*x{i};
   xz0{i} = z0{i}*x{i};
         x2z1{i} = x{i}**2*z1{i};
         x2z0{i} = x{i}**2*z0{i};

   end;

         pai_t1 = sum(of z1_1 - z1_&n_mixed)/(sum(of z1_1 - z1_&n_mixed)+ sum(of z0_1 -
z0_&n_mixed));
         mu1_t1 = (sumxdis + sum(of xz1_1 - xz1_&n_mixed))/(m + sum(of z1_1 -
z1_&n_mixed));
         mu2_t1 = sum(of xz0_1 - xz0_&n_mixed)/sum(of z0_1 - z0_&n_mixed);
         sigma_t1 = (sumx2dis - 2*mu1_t1*sumxdis + m*mu1_t1**2 +
             sum(of x2z1_1 - x2z1_&n_mixed) - 2*mu1_t1*sum(of xz1_1 -
xz1_&n_mixed)+ mu1_t1**2*sum(of z1_1-z1_&n_mixed)+
        sum(of x2z0_1 - x2z0_&n_mixed) - 2*mu2_t1*sum(of xz0_1 - xz0_&n_mixed)+
mu2_t1**2*sum(of z0_1-z0_&n_mixed))/(&n_know+&n_mixed);

         diff=abs(mu1_t1-mu1_t)+abs(mu2_t1-mu2_t)+abs(sigma_t1-sigma_t)+abs(pai_t1-
pai_t);
         pai_t=pai_t1; mu1_t=mu1_t1; mu2_t=mu2_t1; sigma_t=sigma_t1;
         pai_f = (m+&n_mixed*pai_t)/(m+&n_mixed);
         output;
  end;
```

```
  run;

  data out;
    set mixed2 end=eof;
            if eof then output;
  run;

  proc print data=out;
  var pai_t pai_f mu1_t mu2_t sigma_t;
  title "normal distributions: n_know=&n_know, n_mixed=&n_mixed, pai=&pai;
mu1=&mu1; mu2=&mu2; std=&std";
  run;
%mend;

%em (n_know=50, n_mixed=250, pai=0.4, mu1=2, mu2=5, std=1);
```

```
*****************************************************************;
* FUNCTION:     This program is to find the cutoff based on minimizing cost function
*****************************************************************;

libname in 'h:\personal\dissertation';

options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro ROC(incid=,/*** incidence ***/
        mu1=5, /*** means of normal distribution of non-cases***/
                mu2=2, /*** mean of normal distribution of cases ***/
                std=1, /*** std of normal distribution distributions ***/
                c_fn=10); /*** cost of false negative to false positive ***/
data b;
  pai = &incid;
  c_fn=&c_fn;
  mu1=&mu1;
  mu2=&mu2;
  std=&std;
  retain cost1 1;
  do cutoff=-5 to 15 by 0.01;
  fp = cdf("normal", cutoff, &mu1, &std);
  fn = 1 - cdf("normal", cutoff, &mu2, &std);
  Exy = pai*(1-fn);
  Ex=pai*(1-fp-fn)+fp;
  Ey=pai;
  var1=pai-pai**2;
  var2=ex-ex**2;
  corr=(Exy-(Ex)*EY)/sqrt(var1*var2);
  mis_cls = pai*fn+(1-pai)*fp;
  cost = pai*&c_fn*fn+(1-pai)*fp;
  if cost<cost1 then cost1=cost;
  output;
  end;
run;

data b;
  set b;
  if cost=cost1;
run;

proc sort data=b; by pai cutoff;
data b;
  set b;
  by pai cutoff;
  if last.pai;
run;

data temp;
  set temp b;
run;
```

```
%mend;
data temp;
  set _null_;

%ROC(incid=0.2, c_fn=10);
```

```
******************************************************************;
* FUNCTION:     This program is to find the cutoff based on minimizing misclassification
rate
******************************************************************;

libname in 'h:\personal\dissertation';

options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro mis(incid=, /** incidence **/
          mu1=5,  /** mean of normal distribution of non-cases **/
                   mu2=2,  /** mean of normal distribution of cases
                   std=1);   /** std of normal distribution **/
data b;
  pai = &incid;
  mu1=&mu1;
  mu2=&mu2;
  std=&std;
  retain mis_cls1 1;
  do cutoff=-5 to 15 by 0.01;
  fp = cdf("normal", cutoff, &mu1, &std);
  fn = 1 - cdf("normal", cutoff, &mu2, &std);
  Exy = pai*(1-fn);
  Ex=pai*(1-fp-fn)+fp;
  Ey=pai;
  var1=pai-pai**2;
  var2=ex-ex**2;
  corr=(Exy-(Ex)*EY)/sqrt(var1*var2);
  mis_cls = pai*fn+(1-pai)*fp;
  if mis_cls<mis_cls1 then mis_cls1=mis_cls;
  output;
  end;
run;

data b;
  set b;
  if mis_cls1=mis_cls;
run;

proc sort data=b; by pai cutoff;
data b;
  set b;
  by pai cutoff;
  if last.pai;
run;

data temp;
  set temp b;
run;

%mend;
```

```
data temp;
  set _null_;
run;

%mis(incid=0.2);
```

```
***************************************************************;
* FUNCTION:     Find the cutoff by adjusting for covariate
***************************************************************;

options ps=42 ls=140 source nomprint nomlogic nosymbolgen missing=' ' nonumber nodate;

%macro data (mu1_d=, std1_d=, pai=, rho_d=, mu2_d=, std2_d=,
            mu1_nd=, std1_nd=, rho_nd=, mu2_nd=, std2_nd=);
data disease;
   do i=1 to 1000;
            a1 = &rho_d*&std1_d/&std2_d;
            muz_d = &mu1_d - a1*&mu2_d;
            stdz_d = a1*&std2_d*sqrt(1-&rho_d**2)/&rho_d;
   age = &mu2_d + &std2_d*rannor(1000000);
            z = muz_d + stdz_d*rannor(2000000);
            y = a1*age+z;
            pevent=1;
            keep pevent age y;
   output;
  end;
run;
data nodis;
  do j=1 to 1000*(1-&pai)/&pai;
            a1 = &rho_nd*&std1_nd/&std2_nd;
            muz_nd = &mu1_nd - a1*&mu2_nd;
            stdz_nd = a1*&std2_nd*sqrt(1-&rho_nd**2)/&rho_nd;
   age = &mu2_nd + &std2_nd*rannor(3000000);
            z = muz_nd + stdz_nd*rannor(4000000);
            y = a1*age+z;
            pevent=0;
            keep pevent age y;
   output;
  end;

  proc corr data=disease;
  var pevent y;
  with age;
run;

%mend;

%macro logistic (cutoff=);

data indata;
  set disease nodis;
  if y>=&cutoff then event2=0;
  else if y ne . then event2=1;
run;

data disease nodis;
  set indata;
  if pevent=1 then output disease;
```

```
  else if pevent=0 then output nodis;
run;

/* model the false negative rate in the case population */;

proc logistic data=disease;
  model event2 = age;
  ODS output ParameterEstimates=PE1;
run;

data pe1;
  set pe1;
  retain incercep1 age1;
  if variable='Intercept' then incercep1=estimate;
  else if variable='age' then age1=estimate;
  if _n_=2 then output;
  keep incercep1 age1;
run;

/* model false positive rate in the non-case population */;

proc logistic data=nodis descending;
  model event2 = age;
  ODS output ParameterEstimates=PE2;
run;

data pe2;
  set pe2;
  retain incercep2 age2;
  if variable='Intercept' then incercep2=estimate;
  else if variable='age' then age2=estimate;
  if _n_=2 then output;
  keep incercep2 age2;
run;

data a;
  set disease;
  if _n_=1 then set pe1;
  xb = incercep1+age*age1;
  predict1 = exp(xb)/(1+exp(xb));
  keep predict1;
run;

data b;
  set nodis;
  if _n_=1 then set pe2;
  xb = incercep2+age*age2;
  predict2 = exp(xb)/(1+exp(xb));
  keep predict2;
run;

proc means data=indata noprint;
```

```
  var pevent;
  output out=pai (keep=pai) mean=pai;
run;

proc means data=a noprint;
  var predict1;
  output out=fn (keep=fn) mean=fn;
run;

proc means data=b noprint;
  var predict2;
  output out=fp (keep=fp) mean=fp;
run;

data c;
  merge pai fn fp;
  Exy = pai*(1-fp);
  Ex=pai*(1-fp-fn)+fn;
  Ey=pai;
  var1=pai-pai**2;
  var2=ex-ex**2;
  corr=(Exy-(Ex)*EY)/sqrt(var1*var2);
  cufoff=&cutoff;
run;

data temp1;
  set temp1 c;
run;

%mend;

%data (pai=0.5, mu1_d=1, std1_d=1, rho_d=0.1, mu2_d=3, std2_d=1,
            mu1_nd=2.5, std1_nd=1, rho_nd=0.9, mu2_nd=3, std2_nd=1);

data temp1;
  set _null_;
run;

%logistic (cutoff=1.2);
```