

On population-based measures of agreement

By

Kerrie P. Nelson†,

Department of Statistics, University of South Carolina, U.S.A.

and Don Edwards,

Department of Statistics, University of South Carolina, U.S.A.

and Teodor Gamishev,

University of South Carolina, U.S.A.

and Roumen Kozarev,

Palmetto GBA, South Carolina, U.S.A.

May 15, 2007

†*Address for correspondence:*

Department of Statistics,
University of South Carolina, 1523 Greene Street
Columbia,
SC 29208, U.S.A.

E-mail: kerrie@stat.sc.edu

Phone: (803) 777-7800

Fax: (803) 777-4048

Abstract

Measuring agreement between qualified experts is commonly used to determine the effectiveness of a diagnostic procedure. Many methods are available for assessing agreement, including Cohen's kappa, which is a very popular summary measure of agreement due to its appealingly simple usage and interpretation. However, it has been previously shown that a number of flaws exist in its usage, which can lead to biased estimates of agreement. Cohen's kappa seriously underestimates the level of agreement between experts classifying items for a rare disease. In this paper we compare the traditional Cohen's kappa with the true form of Cohen's kappa when the measurement of agreement between many raters and items in an underlying diagnostic procedure is of interest for binary classifications. We demonstrate how an alternative newly developed kappa statistic based upon the class of generalized linear mixed models, that is interpretable in the same manner as Cohen's kappa yet does not suffer from the same flaws can provide a more appropriate assessment of agreement in a population-based study. The form of the model-based statistic is examined for both the probit and logit scenarios. Simulation studies and an application to a cancer of the uterus study (Holmquist 1968) demonstrate the performance of the three different kappa statistics.

Key words: Cohen's kappa, model-based kappa, agreement, prevalence, generalized linear mixed model, crossed random effects, diagnostic procedure, reliability.

1. Introduction

In many biomedical settings, a subjective procedure is used to diagnose a patient's condition (Holmquist 1968, Beam 1996). For example, a qualified physician may examine a biopsy for the presence or absence of cancer, and a psychologist may label the behavior of children in the playground as aggressive or otherwise based upon some prespecified criteria. Strong agreement between raters in their classifications of such items could be considered a necessary prerequisite for the effectiveness of any diagnostic procedure. However, wide variability has been shown to exist between raters as far back as the 1950's (Yerushalmy 1956, Holmquist 1968, Elmore et al 1994, Beam 1996).

Inter-rater agreement studies generally involve the classification of a set of items by a small group of raters, however, the main interest often lies in how results extend to the general populations of such raters and items and the underlying diagnostic procedure, and not just the specific raters and items studied. Our focus in this paper is to examine the assessment of agreement over many raters and items, assumed to be randomly selected from their respective populations, and to make inference regarding the underlying diagnostic process of interest.

Many methods have been developed to assess the reliability between raters' classifications. These include summary statistics such as Cohen's kappa (Scott 1955, Cohen 1960), the intra-class correlation coefficient (Shrout and Fleiss 1979, Barlow 1996, Kraemer 1979, Bloch and Kraemer 1989), and others (for example, Chinchilli, Martel et al 1996). Model-based approaches include the class of log-linear models (Tanner and Young 1985, Agresti 1988, Goodman 1979, Schuster 2002), logistic regression models (Coughlin et al 1992, Graham 1995, Lipsitz et al 2003), latent class and latent trait models (Dawid and Skene 1979, Kraemer 1979, Uebersax and Grove 1990, Williamson and Manatunga 1997), and

others (Coull and Agresti 2003, Nelson and Pepe 2000). Many of these methods are designed for a small number of raters, and can become unwieldy for larger numbers of raters. In addition, these methods provide information about the agreement between a fixed set of raters under study, and do not extend to providing inference regarding the populations of such raters, or the underlying diagnostic procedure. Other methods, including a latent model approach developed by Williamson and Manutanga (1997), a generalized log-linear model approach with random effects (Coull and Agresti 2003), a population model for Cohen's kappa (Bloch and Kraemer 1989) consider the agreement scenario when many raters and items are involved, and are able to provide inference regarding the diagnostic procedure.

Originally developed as a chance-corrected measure of agreement of classifications made on a binary nominal scale between two raters, Cohen's kappa has since been extended to provide a summary measure of agreement for classifications made on ordinal and nominal scales, (Cohen 1968), for more than two raters (Fleiss 1971, Light 1971, Kraemer 1980), and various other extensions (Ghosh et al 1995, Landis and Koch 1977, Barlow 1996, Klar et al 2000). However, a number of issues have been raised in its usage, including a susceptibility to marginal and prevalence effects (Byrt et al, Maclure and Willett 1987, Nelson and Pepe 2000), Simpson's paradox (Thompson 2001), and sensitivity to the number of categories employed in the classification (Maclure and Willett 1987). Cohen's kappa is also rarely comparable across different studies (Feinstein and Cicchetti 1990), and some hypothesis tests have been developed for testing the homogeneity of Cohen's kappas across different populations (Donner and Klar 1996). However, while model-based approaches provide a more complete and informative description of agreement than the simpler summary statistics (Agresti 1988, Feinstein and Cicchetti 1990), Cohen's kappa remains a very popular tool in the assessment of agreement in many situations due to its

ease of calculation and simple interpretation.

In this paper we demonstrate that the true value of Cohen’s kappa, based upon population measures and over many raters and items, overcorrects for chance agreement (equivalently, prevalence), often resulting in severe underestimation of the agreement present between the raters’ classifications. We also describe how a simple model-based measure based upon the class of generalized linear mixed models (Nelson and Edwards 2007) provides a summary of the agreement in a setting involving many raters and items, while avoiding many of the issues raised in Cohen’s kappa usage, and still being simple to interpret in a similar manner to Cohen’s kappa. In addition, the described framework allows for inferences to be made to be made regarding the underlying diagnostic procedure.

The remainder of the paper is as follows: In section 2, we describe the true population-based measures of observed and chance agreement which are used to formulate the kappa measures of agreement. Section 3 provides details on the form of Cohen’s kappa when population-based inference is of interest. We describe the class of generalized linear mixed models and associated model-based measure of agreement and their estimation in section 4. Simulation studies are conducted in Section 5 to demonstrate the properties of the fore-mentioned summary statistics of agreement. In Section 6, an application to a breast cancer study dataset is described followed by discussion and concluding remarks in Section 7.

2. Population-based measures of agreement

When the primary focus is in making inference about an underlying diagnostic process, based upon raters and items randomly selected from their respective populations, the true values for the observed and chance agreement rates between the raters can be derived.

The observed probability of agreement p_0 is the proportion of time that two randomly selected raters agree in their classification of one randomly selected item. The true value of the observed or “raw” agreement rate p_0 in the underlying diagnostic process for binary classifications y_{ij} (1=diseased, 0=not diseased, for example), made by randomly selected raters j and j' , ($j \neq j'$) on a randomly selected item i :

$$\begin{aligned} p_0 &= pr\{(y_{ij} = 1) \cap (y_{ij'} = 1)\} + pr\{(y_{ij} = 0) \cap (y_{ij'} = 0)\} \\ &= 1 - 2pr\{(Y_{ij} = 1) \cap (Y_{ij'} = 0)\}. \end{aligned} \quad (1)$$

The classifications made by the j th and j' th raters on an item are interchangeable since the two raters are randomly selected from the population, and thus any pair of ratings has a distribution that is invariant under permutations of the raters (Bloch and Kraemer 1989, Banerjee et al 1999).

The true measure of chance agreement, p_c , is the probability that two randomly selected raters from a population make identical classifications on two different randomly chosen items. For raters j and j' and items i and i' ($i \neq i', j \neq j'$), the chance agreement rate takes the form:

$$\begin{aligned} p_c &= pr\{(y_{ij} = 1) \cap (y_{i'j'} = 1)\} + pr\{(y_{ij} = 0) \cap (y_{i'j'} = 0)\} \\ &= 1 - 2pr\{(y_{ij} = 1) \cap (y_{i'j'} = 0)\}. \end{aligned} \quad (2)$$

It is important to note that these quantities are a consequence of the setting (i.e. over many raters and items), and are not reliant upon any statistical model one may adopt for the agreement process. The following theorem demonstrates the relationship between p_0 , p_c , and the minimum values that these quantities can take in this longrun setting.

Theorem 2.1. Defining the prevalence p_1 as the probability (or expected value) of classifying a randomly selected item as positive (or a 1) by a single randomly chosen

rater, then over many raters and items, the rate of observed agreement is always equal to or greater than the rate of chance agreement, p_c , which in turn is always greater or equal to 0.5, i.e.

$$p_0 \geq p_c = 1 - 2p_1(1 - p_1) \geq \frac{1}{2}.$$

Proof. First, define $p_{ij} = pr(y_{ij} = 1)$ for a specific item i and randomly chosen rater j . We can define $p_{i\cdot} = pr(y_{ij} = 1 | y_{i1}, y_{i2}, \dots, y_{iN_J})$ where N_J is the total number of raters in the population. The unconditional prevalence (which assumes both a randomly selected item and rater) can be denoted as $p_1 = pr(y_{ij} = 1) = E_i(p_{i\cdot})$. Considering the probability of disagreement between two randomly selected raters j and j' ($j \neq j'$) classifying the same i th item, we observe that:

$$\begin{aligned} pr[(y_{ij} = 1) \cap (y_{ij'} = 0)] &= E[y_{ij}(1 - y_{ij'})] \\ &= E_i [E[y_{ij}(1 - y_{ij'}) | y_{i1}, y_{i2}, \dots, y_{iN_J}]] \\ &= E_i \left[\frac{N_J}{N_J - 1} p_{i\cdot} (1 - p_{i\cdot}) \right] \\ &\leq \frac{N_J}{N_J - 1} E_i(p_{i\cdot}) E_i(1 - p_{i\cdot}) \quad (\text{Jensen's inequality}) \\ &= \frac{N_J}{N_J - 1} p_1 (1 - p_1) \\ &\approx p_1 (1 - p_1) \quad \text{when } N_J \text{ is very large.} \end{aligned}$$

Thus, the rate of observed agreement satisfies:

$$\begin{aligned} p_0 &= 1 - 2pr[(y_{ij} = 1) \cap (y_{ij'} = 0)] \\ &\geq 1 - 2\left(\frac{N_J}{N_J - 1}\right)p_1(1 - p_1) \\ &\approx 1 - 2p_1(1 - p_1) = p_c \quad \text{as } N_J \rightarrow \infty. \end{aligned}$$

These relationships in the above theorem also hold as a consequence of the setting of many raters and items in the long run, and are not a consequence of a specific model framework. The minimum allowable value of 0.5 for both quantities p_0 and p_1 is also evidenced by work carried out by Bloch and Kraemer (1989). In a practical sense, the minimum value of p_0 will be achieved when the randomly selected items cannot be distinguished from each other when being rated, and the raters consequently classify each item in a random manner with a fifty-fifty chance as a success or failure.

3. Cohen’s Kappa

The true form of Cohen’s kappa (Scott 1955, Cohen 1960) based upon classifications made by a population of raters on a population of items, is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c},$$

where p_0 and p_c are defined in Section 2 above. Cohen’s kappa is commonly estimated for a sample of data as $\hat{\kappa} = (\hat{p}_0 - \hat{p}_c)/(1 - \hat{p}_c)$, where \hat{p}_0 and \hat{p}_c are consistent estimators of observed and chance agreement obtained from the sample data. This data-driven form of Cohen’s kappa can take on values between -1 and 1, where a large positive value is suggestive of strong agreement between raters, and values less than 0 means the agreement is less than what would be expected by chance. Cohen’s kappa is not model-driven (Brennan and Prediger 1981), which explains why there is often no distinction made between the estimated kappa, $\hat{\kappa}$, and its true value, κ . Cohen’s kappa is dependent upon the marginal distributions of the raters’ classifications, frequently leading to prevalence and bias effects, which can severely under- or over-estimate the resulting agreement between the raters. The prevalence effect is due to the effect of the relative proportions of the “yes” and “no” categories (for the same p_0 , Cohen’s kappa can be two or more times

greater), and the bias effect occurs when the marginal totals are unbalanced between the two raters. Various attempts to adjust for these effects, including the PABAK (prevalence and bias-adjusted kappa) (Byrt et al 1993, Feinstein and Cicchetti 1990), simply a scaled version of the observed rate of agreement and not chance-corrected, tend to suffer from the same issues as the original kappa.

4. Model-based measures of agreement

The class of generalized linear mixed models with a crossed random effects structure provides a flexible framework for examining agreement in an underlying diagnostic procedure for specified populations of raters and items, where any numbers of raters and items can be included, and missing data is allowed. These models incorporate the raters and items as random effects, allowing for inference to be made on the underlying populations of such raters and items, and thus providing consistent estimates of the true quantities used for measuring agreement as described in Section 2. Defining y_{ij} as the classification made by the j th randomly chosen rater on the i th randomly selected item, the associated probability of the item being classified as a success, p_{ij} , can be modelled using either a probit or logit link function for binary classifications. The general form of this model is as follows:

$$g(\cdot) = \eta + u_i + v_j \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$

where $g(\cdot)$ is the link function of choice, and η , the intercept term, adjusts for the internal prevalence (the long-run frequency of successes) in the data; a large value for η is associated with a high frequency of successes. The terms u_i and v_j refer to the random effects for the i th item and j th rater, which are assumed to be independent and normally distributed with mean 0 and variances σ_u^2 and σ_v^2 respectively. A positive value for u_i suggests that the i th item is more likely than other items to be classified as a success

(or positive) over many raters. A positive value for v_j suggests a rater who is relatively liberal in classifying an item as a success over his/her classification of many such items.

4.1 Cohen's kappa for a probit generalized linear mixed model

A generalized linear mixed model with a probit link function for modelling binary classifications takes the form:

$$\Phi^{-1}(p_{ij}) = \eta + u_i + v_j \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (3)$$

Defining a new random variable W whose quantile function is the inverse cumulative normal distribution (more generally $g^{-1}(\cdot)$), where W is normally distributed with mean 0 and variance $\sigma_W^2 = 1$ is useful here. Under the probit model, the true value of the observed or "raw" agreement p_0 in the underlying diagnostic process, for randomly selected raters j and j' , ($j \neq j'$) classifying a randomly selected item i , is derived as follows (based upon equation 2 in Section 3.0):

$$\begin{aligned} p_0 &= 1 - 2pr\{(y_{ij} = 1) \cap (y_{ij'} = 0)\} \\ &= 1 - 2E\{g^{-1}(\eta + u_i + v_j) [1 - g^{-1}(\eta + u_i + v_{j'})]\} \\ &= 1 - 2E\{pr(W_1 \leq \eta + u_i + v_j) [1 - pr(W_2 \leq \eta + u_i + v_{j'})]\} \\ &= 1 - 2 \int_{-\infty}^{\infty} pr(W_1 - v_1 \leq \eta + u) [1 - pr(W_2 - v_2 \leq \eta + u)] \phi\left(\frac{u}{\sigma_u}\right) \frac{1}{\sigma_u} du \\ &= 1 - 2 \int_{-\infty}^{\infty} \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right) \left[1 - \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right)\right] \phi(z) dz, \quad 0 \leq p_0 \leq 1, \quad (4) \end{aligned}$$

where $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_v^2 + \sigma_w^2)$ and $\eta^* = \eta / \sqrt{\sigma_u^2 + \sigma_v^2 + \sigma_w^2}$ and the functions ϕ and Φ are the standard normal pdf and cdf respectively. The quantity ρ is itself an appealing measure of agreement in a similar manner to an intraclass correlation coefficient (Shrout and Fleiss

1979): large values of ρ occur when the variance of the items, σ_u^2 is large relative to the variability in the raters' attitudes and experience, σ_v^2 . It can be demonstrated that p_0 , given by equation (4), is a symmetric function of the standardized prevalence parameter η^* , and monotonically increases with $|\eta^*|$ and ρ , as shown in Figure 1(a). In practice, this means that as the items become more distinguishable from one another (larger values of σ_u^2) relative to the total variability, the rate of observed agreement between the raters increases, and at a faster rate for milder prevalence.

Under the probit model, the prevalence p_1 is calculated as:

$$p_1 = E\{\Phi(\eta + u_i + v_j)\} = pr\{W_{ij} - u_i - v_j \leq \eta\} = \Phi(\eta^*). \quad (5)$$

Thus the true measure of chance agreement, p_c , is a simple one-to-one function of the standardized internal prevalence η^* :

$$p_c = 1 - 2p_1(1 - p_1) = 1 - 2\Phi(\eta^*)[1 - \Phi(\eta^*)]. \quad (6)$$

These population-based quantities can be estimated by finding consistent estimators of the parameters η , σ_u^2 and σ_v^2 in the corresponding generalized linear mixed model through the use of maximum likelihood or other available methods.

The true form of Cohen's kappa under the probit model using the previously defined model-based measures of p_0 and p_c from equations (4) and (6) above is:

$$\kappa = \frac{p_0 - p_c}{1 - p_c} = 1 - \frac{\int_{-\infty}^{\infty} \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right) \left[1 - \Phi\left(\frac{z\sqrt{\rho} + \eta^*}{\sqrt{1-\rho}}\right)\right] \phi(z) dz}{\Phi(\eta^*)\{1 - \Phi(\eta^*)\}} \quad (7)$$

The true value of Cohen's kappa statistic, κ , is calculated based on population-based measures ρ and η^* obtained from the associated probit generalized linear mixed model.

Figure 2(a) displays a plot of Cohen’s kappa κ against ρ for increasing levels of the standardized internal prevalence η^* under the probit model. It is observed for a fixed value of ρ that the value of Cohen’s kappa decreases as η^* increases, indicating a reliance of Cohen’s kappa on the amount of chance agreement present (since chance agreement is a one-to-one function of prevalence). This is similar to the prevalence effect observed in the data-driven form of Cohen’s kappa. This relationship is also demonstrated in Figure 3, which displays the values of Cohen’s kappa against prevalence p_1 for differing values of ρ , where the influence of prevalence on the resulting value of Cohen’s kappa is clearly observed. However, a population-based measure of agreement, after an adjustment for chance-agreement has been made, should be robust to the level of prevalence (i.e. chance agreement), and influenced only by the item distinguishability and variability between the raters, components of the quantity ρ . The decrease in the value of κ associated with an increasing absolute value of prevalence indicates that Cohen’s kappa is overcorrecting for prevalence. The effect is relatively minor when the underlying prevalence is mild, say $-1 < \eta^* < 1$, but will severely bias the value of Cohen’s κ downwards when the prevalence is large. As a consequence of Theorem 2.1, we note that the true value of Cohen’s kappa cannot be less than zero when based upon population-based measures, since any rating process must have agreement at least as large as what would be observed when the items are indistinguishable.

Figure 2(a) also demonstrates that the value of Cohen’s kappa increases as the value of ρ increases - larger values of ρ suggest that the items are more distinguishable from each other, which leads to better agreement between raters. The increasing effect is slower for larger values of the standardized internal prevalence η^* .

4.2 Cohen's kappa for a logit generalized linear mixed model

The results pertaining to the probit generalized linear mixed model above have the appeal of simplicity: the observed agreement p_0 and Cohen's kappa κ are functions only of the standardized internal prevalence $|\eta^*|$ and ρ . However, in practice, the logit link function enjoys far more widespread use than the probit link. In this section we show the true forms of observed and chance probabilities of agreement and Cohen's kappa when the logit link generalized linear mixed model is employed. The generalized linear mixed model considered here is:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta + u_i + v_j \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$

where the terms are as defined earlier in Section 4.1. The random variable W is again defined as the random variable whose quantile function is $g(\cdot)$; for the logit link function, W has a logistic distribution with density function

$$f_W(w) = \frac{e^{-w}}{(1+e^{-w})^2}, \quad -\infty < w < \infty,$$

with a mean of 0 and variance $\sigma_w^2 = \pi^2/3$. A new parameter defined as $\rho_2 = \sigma_v^2/(\sigma_v^2 + \sigma_w^2)$, which can take values between 0 and 1, measures the relative strength of the normally distributed rater random effect $v = v_j$ in the random variable $Q = (W - v)/\sigma_u$. When ρ_2 is near 1, the rater random effects v_j dominate the logistic component of Q , and the results will be very similar to those seen for the probit model. For example, consider the overall prevalence in the logistic model, with its exact form using the law of total

probability and algebra given by:

$$\begin{aligned}
p_1 &= pr\{W_{ij} - u_i - v_j \leq \eta\} \\
&= pr\{[(1 - \rho_2)(1 - \rho)]^{\frac{1}{2}}W^* + \rho^{\frac{1}{2}}Z_u + [\rho_2(1 - \rho)]^{\frac{1}{2}}Z_v \leq \eta^*\} \\
&= \int_{-\infty}^{\infty} \Phi \left\{ \frac{\eta^* - [(1 - \rho_2)(1 - \rho)]^{\frac{1}{2}}s}{[\rho + \rho_2(1 - \rho)]^{\frac{1}{2}}} \right\} f_{W^*}(s) ds,
\end{aligned}$$

for $W^* = W_{ij}/\sigma_W$, $z_u = -u_i/\sigma_u$, and $z_v = -v_j/\sigma_u$. As ρ_2 approaches 1, the influence of the standardized logistic term W^* in the above expression decreases, and p_1 converges to the same expression as observed under the probit model formulation:

$$\lim_{\rho_2 \rightarrow 1} p_1 = pr(\rho^{\frac{1}{2}}z_u + (1 - \rho)^{\frac{1}{2}}z_w \leq \eta^*) = \Phi(\eta^*).$$

Under the logit model in equation (3), the observed agreement rate p_0 is:

$$p_0 = 1 - 2 \int_{-\infty}^{\infty} F_Q \left(\frac{\eta^*}{\sqrt{\rho} + z} \right) \left[1 - F_Q \left(\frac{\eta^*}{\sqrt{\rho} + z} \right) \right] \phi(z) dz,$$

where η^* , ρ and ρ_2 are as defined earlier in equations (3), (4) and (12) respectively. Under the logit model when W has a logistic distribution and the rater random effects normally distributed, the cumulative distribution function of Q is:

$$F_Q(q) = pr(Q \leq q) = \int_{-\infty}^{\infty} \Phi \left\{ q \left(\frac{\rho}{\rho_2(1 - \rho)} \right)^{\frac{1}{2}} - s \left(\frac{1 - \rho_2}{\rho_2} \right)^{\frac{1}{2}} \right\} f_{W^*}(s) ds.$$

The true form of chance agreement is of a similar form as for the probit model, $p_c = 1 - 2p_1(1 - p_1)$. Figure 1(b) presents a plot of the true observed agreement rate against ρ for increasing amounts of standardized internal prevalence under the logit model, where the term ρ_2 is set at 0.50. We observe a similar trend as for the probit model displayed in Figure 1(a) (which corresponds to $\rho_2 \rightarrow 1$); as the distinguishability between the items improves relative to the total variability, the observed agreement rate p_0 increases.

Differing values of ρ_2 result in almost identical plots, with only subtle differences occurring at very low values of ρ .

Based upon the true forms of p_0 and p_c derived under the logit model, the analytical form of Cohen's kappa is:

$$\kappa_M = 1 - \left[\frac{\int_{-\infty}^{\infty} F_Q \left(\frac{\eta^*}{\sqrt{\rho+z}} \right) \left[1 - \left(\frac{\eta^*}{\sqrt{\rho+z}} \right) \right] \phi(z) dz}{\int_{-\infty}^{\infty} \Phi \left\{ \frac{\eta^* - [(1-\rho_2)(1-\rho)]^{\frac{1}{2}} s}{[\rho + \rho_2(1-\rho)]^{\frac{1}{2}}} \right\} f_{W^*}(s) ds \left(1 - \int_{-\infty}^{\infty} \Phi \left\{ \frac{\eta^* - [(1-\rho_2)(1-\rho)]^{\frac{1}{2}} s}{[\rho + \rho_2(1-\rho)]^{\frac{1}{2}}} \right\} f_{W^*}(s) ds \right)} \right],$$

where $F_Q(q)$ is as defined in equation (7). In Figure 2(b), the true values of Cohen's kappa under the logit model formulation are presented for increasing prevalence $|\eta^*|$ and for $\rho_2 = 0.5$. Again, for other values of ρ_2 the plots were virtually identical. The observed and chance agreement rates p_0 and p_c and Cohen's kappa are fairly flat as functions of ρ_2 , which explains why the choice of ρ_2 is of little importance. We again observe the similarity of the plot with those seen under the probit model - that Cohen's kappa overcorrects for prevalence/chance agreement, and more severely so for larger values of standardized internal prevalence.

Though the results of this section are mathematically more complex than for the probit link, it has been shown that the relationships between the true values of the observed agreement p_0 and $|\eta^*|$ and Cohen's kappa are for all intents and purposes equivalent to those found for the probit link - a fact that should not be surprising given the similarity of the standardized logistic and normal distributions.

5. A model-based measure of agreement for the probit generalized linear mixed model

Cohen's kappa was originally developed in the absence of any formal model for the data as a simple summary statistic of agreement with correction for chance agreement. For population-based studies where models such as the probit and logit generalized linear mixed models described in Section 3 can be employed, more appropriate corrections for chance-agreement correction can be developed.

In this section we describe a simple summary statistic κ_M based upon the probit generalized linear mixed model (Nelson and Edwards 2007) which estimates agreement between raters in an underlying diagnostic procedure. This statistic is based upon the classifications made by a set of randomly selected raters on a set of randomly selected items, while appropriately adjusting for chance agreement. The quantity κ_M is primarily based upon the observed agreement p_0 , and after adjusting for the chance agreement, is scaled to take values between 0 and 1 to allow for direct comparability and interpretability as Cohen's kappa.

The model-based kappa takes the form:

$$\begin{aligned}
 \kappa_M &= 2p_0(\eta = 0, \sigma_u^2, \sigma_v^2) - 1 \\
 &= 2 \left[1 - 2 \int_{-\infty}^{\infty} \Phi \left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \left[1 - \Phi \left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \right] \phi(z) dz \right] - 1 \\
 &= 1 - 4 \int_{-\infty}^{\infty} \Phi \left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \left[1 - \Phi \left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}} \right) \right] \phi(z) dz. \tag{8}
 \end{aligned}$$

Chance agreement is adjusted for by setting the standardized internal prevalence term η^* to zero since chance agreement p_c is a monotone function of $|\eta^*|$ and is minimized when $\eta^* = 0$. The approximate variance associated with the model-based kappa κ_M in equation

(8) is derived using the asymptotic distributions of the estimated variances σ_u^2 and σ_v^2 and the multivariate delta theorem as follows:

$$\sqrt{n}((\hat{\sigma}_u^2, \hat{\sigma}_v^2) - (\sigma_u^2, \sigma_v^2)) \xrightarrow{d} N(0, \Sigma), \text{ where the matrix } \Sigma = \begin{pmatrix} 2\sigma_u^4 & 0 \\ 0 & 2\sigma_v^4 \end{pmatrix} \text{ and}$$

$$\text{var}(\hat{\kappa}_M) = \frac{16}{IJ} \left[\left\{ \int_{-\infty}^{\infty} \left(\frac{1}{2\hat{\rho}(1-\hat{\rho})} \left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \phi \left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \left[1 - 2\Phi \left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \right] \right) \phi(z) dz \right\}^2 \times \right. \\ \left. \left(\frac{2\hat{\sigma}_u^4}{(\hat{\sigma}_T^2)^4} [(\hat{\sigma}_v^2 + \hat{\sigma}_w^2)^2 + \hat{\sigma}_v^4] \right) \right]$$

where $\sigma_T^2 = \sigma_u^2 + \sigma_v^2 + \sigma_w^2$. Under the logit model described in Section 3.2, a value for Cohen's kappa can also be derived. However, since the prevalence p_1 and chance agreement p_c are still monotone functions of η and $|\eta^*|$ respectively, the expression for κ_M under the probit model is a viable choice for a chance-corrected measure of agreement under the logit link model.

Consistent estimators can be obtained for the parameters η , σ_u^2 and σ_v^2 in both the probit and logit generalized linear mixed model using a readily-available almost-exact maximum likelihood algorithm (McCulloch 1997). This algorithm is described in more detail below.

6. Simulation Studies

Simulation studies were carried out to compare the three summary statistics of agreement: the traditional form of Cohen's kappa κ , calculated on estimates of p_0 and p_c obtained directly from the data (Cohen 1968); and the true form of Cohen's kappa and the model-based kappa, based upon population measures of p_0 and p_c , calculated using estimated parameters from the associated generalized linear mixed model.

One hundred datasets were randomly generated according to the model

$$\Phi^{-1}(p_{ij}) = \eta + u_i + v_j \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$

based upon the true values of the parameters set at $\boldsymbol{\theta} = (\eta, \sigma_u^2, \sigma_v^2) = (4, 2, 2)$, for $J = 20$ raters and $I = 20$ items with random effects terms assumed normally distributed with mean 0. Other sets of simulations were also conducted to test the effects of increasing prevalence on the three kappa statistics, with other values of η set at $-4, 0$ and 1 . Almost-exact maximum likelihood estimates of the parameter vector $\boldsymbol{\theta} = (\eta, \sigma_u^2, \sigma_v^2)$ were obtained by implementing McCulloch's MCEM (Monte-Carlo expectation-maximization) algorithm (McCulloch 1997). This algorithm uses a Metropolis-Hastings step to sample sets of the (unobserved) random effects by drawing vectors $u = (u_1, u_2, \dots, u_I)$ and $V = (v_1, v_2, \dots, v_J)$ from the conditional distributions $f(u|y)$ and $f(v|y)$ respectively. At each iteration, m sets of random effects are drawn, and the value of $E(l(\boldsymbol{\theta}^t)|y)$ is estimated through the use of Monte-Carlo averaging, where $l(\boldsymbol{\theta}^t)|y$ is the log-likelihood function of the complete data (y, u, v) . The value of m ranged from 1000 for earlier iterations to 500,000 for the final iterations. Starting values were set at $\eta = 1$, $\sigma_u^2 = 1$, and $\sigma_v^2 = 1$. Other starting values were also tried and converged to the same parameter estimates. The parameter vector $\boldsymbol{\theta}$ was updated at each iteration using the one-step expectation-maximization algorithm. Convergence of the parameters was considered successful when the difference between the previous estimates and the updated estimates took a maximum value of 0.00001. More details on this algorithm can be found in McCulloch (1997). This algorithm was implemented by the authors using the C programming language.

Table 1 presents the mean value and observed standard errors of the estimated parameters obtained from fitting the probit generalized linear mixed model and the three estimated kappas for $\eta = 4$. It is observed that the almost-exact maximum likelihood estimates

yielded by McCulloch’s MCEM algorithm are unbiased. The mean value of Cohen’s kappa, both the data-driven and true forms are consistently smaller than the model-based kappas.

Table 1: Means and standard errors of the estimated parameters of the probit generalized linear mixed model and the estimated kappa statistics based on 100 simulated datasets where $\boldsymbol{\theta} = (\eta = 4, \sigma_u^2 = 2, \sigma_v^2 = 2)$, and $I, J = 20$.

Parameter	Mean	(Standard error)
$\eta (= 4)$	3.986	0.726
$\sigma_u^2 (= 2)$	2.281	1.363
$\sigma_v^2 (= 2)$	2.198	1.296
Cohen’s kappa (true) κ	0.156	0.102
Cohen’s kappa (data-driven) $\hat{\kappa}$	0.107	0.082
Model-based kappa κ_M	0.269	0.133

Figure 4 displays boxplots of the three different kappa statistics, the traditional Cohen’s kappa and population-based form of Cohen’s kappa, and the model-based kappa statistic, for increasing values of prevalence, $\eta = -4, 0, 1, 4$. Results are based on the one hundred simulated datasets for each value of η used. It is observed that as the absolute value of the prevalence term η increases, both the estimated Cohen’s kappa statistic κ and true Cohen’s kappa decrease on average, leading an overcorrection for prevalence, while the model-based kappa statistic κ_M remains constant, correctly adjusting for the increasing prevalence. The overcorrection in Cohen’s kappa is more noticeable as the true underlying prevalence increases. The data-driven form of Cohen’s kappa is more susceptible to this effect than the true form of Cohen’s kappa based upon population measures. The overcorrection leads to underestimating the level of agreement present between the raters.

7. Application to a cancer dataset

This section describes the application of the developed methodology and model-based kappa statistic, κ_M , to a biomedical dataset collected by Holmquist et al (1968). The dataset consists of classifications made by seven pathologists for the diagnosis of carcinoma in situ of the uterine cervix. Each pathologist rated 118 slides on a five-point scale, where 1 = negative, 2 = atypical squamous hyperplasia, 3=carcinoma in situ, 4=squamous carcinoma with early stromal invasion, 5=invasive carcinoma. For the current analyses, the classification scale has been dichotomized, so that 0 = absence of cancer combines the original categories 1 and 2, while 1 = presence of cancer combines the original categories 3, 4 and 5. Table 1 summarizes the pairwise agreement over all pairs of seven raters, where the classification of each item by a pair of raters, for example, raters 1 and 2 is included only once in the table.

Table 2: Summary of the pairwise agreement for the seven pathologists each classifying 118 slides for the presence (=1) or absence (=0) of cancer (Holmquist et al 1968).

		Rater 2		
		Category	1	0
Rater 1	1	475	148	1709
	0	294	1561	769
Total		1855	625	2478

An estimate of Cohen's kappa (equation (1)) based upon the observed and chance agreement calculated directly from the data was calculated for this dataset. As the dataset contains many raters, a weighted form of Cohen's kappa was implemented (Fleiss 1971).

The value of Cohen’s kappa based upon population-based measures of observed and chance agreement as estimated from a probit generalized linear mixed model fitted to the dataset, and the model-based kappa statistic, κ_M , were also calculated for the Holmquist data.

The probit link generalized linear mixed model in equation (3) was fitted to the data using the data for $I = 118$ slides and $J = 7$ raters. Almost-exact maximum likelihood estimates of the parameters in the vector $\theta = (\eta, \sigma_u^2, \sigma_v^2)$ were obtained by implementing McCulloch’s MCEM (Monte-Carlo expectation-maximization) algorithm (McCulloch 1997) in a similar manner described in the previous section.

The estimated parameters from the fitted probit generalized linear mixed model, and the three versions of Cohen’s kappa are presented in Table 1. We observe that the estimated prevalence η is negative, reflecting the fact that less than half of all the slides were rated as diseased by the pathologists. A relatively high value of σ_u^2 estimated for these classifications suggests that there is considerable variability between the items, making them reasonably distinguishable from each other. In contrast, the rater-to-rater variability is small, suggesting relative consistency of the pathologists in their classifications of slides such as these.

Based on the combined data from the seven raters, the data-driven value of Cohen’s kappa as originally defined by Cohen (1960) and extended to a weighted kappa (Fleiss, 1971) to allow for multiple raters is 0.512. This suggests moderate agreement between the raters in their classifications of the items. The estimate of the model-based kappa in equation (7) is 0.537, slightly higher in value than Cohen’s kappa, due to the higher prevalence η . The differences between the three kappa statistics are not overly dramatic in this example, since prevalence is mild.

8. Discussion

In many instances, the assessment of agreement between qualified raters in a diagnostic procedure is helpful in determining its effectiveness. Concern regarding the wide variability often observed between raters in commonly used diagnostic procedures has led to the development of many different methods for assessing agreement. When the focus is on the characteristics of agreement in the underlying diagnostic procedure over many raters and items, the class of generalized linear mixed models provides an appealing framework for the measurement of agreement, where inference can be made regarding the populations of raters classifying the items of interest. Our focus here is on classifications made on a dichotomous scale, for example, an item is rated as diseased versus not diseased.

In this paper, we have shown that the true value of Cohen's kappa statistic, based upon population measures of observed and chance agreement, overcorrects for prevalence, leading to overly conservative estimates of agreement. In situations where the prevalence is more extreme, i.e. close to 0 or 1, Cohen's kappa will be seriously biased. For any rare disease, Cohen's kappa would seriously underestimate the amount of agreement present between the raters. However, an alternative summary statistic based upon a generalized linear mixed model is shown to more appropriately adjust for prevalence (and chance agreement), while remaining appealingly simple to interpret in a similar manner to Cohen's kappa. Further details on the application of this statistic can be found in Nelson and Edwards (2007).

More extensive work on optimal estimation and formal inference on the parameters and quantities of interest, including the model-based kappa statistic κ_M under the generalized linear mixed model framework are issues deserving further study and are topics for future research.

REFERENCES

- AGRESTI, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, **44**, 539-548.
- AGRESTI, A. and GHOSH, A. (1995). Raking kappa: describing potential impact of marginal distributions on measures of agreement. *Biometrical Journal*, **7**, 811-820.
- BANERJEE, M., CAPOZZOLI, M., McSWEENEY, L. and SINHA, D. (1999). Beyond kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics*, **27** (1), 3-23.
- BARLOW, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, **52**, 695-702.
- BEAM, C.A., LAYDE, P.M., and SULLIVAN, D.C. (1996). Variability in the interpretation of screening mammograms by US radiologists. *Archives of Internal Medicine*, **156**, 209-213.
- BLOCH, D.A. and KRAEMER, H.C. (1989). 2×2 kappa coefficients: Measurements of agreement and association. *Biometrics*, **45**, 269-287.
- BYRT, T., BISHOP, J., and CARLIN, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46** (5), 423-429.
- CHINCILLI, V.M., MARTEL, J.K., KUMANYIKA, S. and LLOYD, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, **52**, 341-353.
- COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

- COHEN, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.
- CONGER, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, **88**, 322-328.
- COUGHLIN, S.S., PICKLE, L.W., GOODMAN, M.T. and WILKENS, L.R. (1992). The logistic modeling of interobserver agreement. *Journal of Clinical Epidemiology*, **45 (11)**, 1237-1241.
- COULL, B.A. and AGRESTI, A. (2003). Generalized log-linear models with random effects, with application to smoothing contingency tables. *Statistical Modelling*, **3**, 251-271.
- DAWID, A.P. and SKENE, A.M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, **28**, 20-28.
- DONNER, A. and KLAR, N. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology*, **43**, 543-548.
- NELSON, K.P. and EDWARDS, D. (2007). A model and measure of agreement for population-based studies. *Technical Report*, Department of Statistics, University of South Carolina.
- ELMORE, J.G., WELLS, C.K., LEE, C.H., HOWARD, D.H. and FEINSTEIN, A.R. (1994). Variability in radiologists' interpretations of mammograms. *The New England Journal of Medicine*, **331 (22)**, 1493-1499.
- FEINSTEIN, A.R. and CICHETTI, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, **43 (6)**, 543-549.

- FLEISS, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378-382.
- GOODMAN, L.A. (1979). Simple models for the analysis of association in cross classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537-552.
- GRAHAM, P. (1995). Modeling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine*, **14**, 299-310.
- HOLMQUIST, N.D., McMAHAN, C.A. and WILLIAMS, O.D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, **84**, 334-345.
- KLAR, N, LIPSITZ, S.R. and IBRAHIM, J.G. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal*, **1**, 45-58.
- KRAEMER, H.C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, **44** (4), 461-472.
- KRAEMER, H.C. (1980). Extension of the kappa coefficient. *Biometrics*, **36**, 207-216.
- LANDIS, J.R. and KOCH, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, **33**, 363-374.
- LANDIS, J.R. and KOCH, G.G. (1977). A one-way components of variance model for categorical data. *Biometrics*, **33**, 671-679.
- LANDIS, J.R. and KOCH, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

- LIGHT, R.J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, **76**, 365-377.
- LINVER, M.N., PASTER, S.B., ROSENBERG, R.D., KEY, C.R., STIDLEY, C.A., KING, W.V. (1992). Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology*, **184**, 39-43.
- LIPSITZ, S.R., PARZEN, M., FITZMAURICE, G.M. and KLAR, N. . (2003). A two-stage logistic regression model for analyzing inter-rater agreement. *Psychometrika*, **68** (2), 289-298.
- MACLURE, M. and WILLETT, W.C. (1987). Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology*, **126** (2), 161-169.
- McCULLOCH, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **437**, 162-170.
- MIGLIORETTI, D.L. and HEAGERTY, P.J. (2007). Marginal modeling of nonnested multilevel data using standard software. *American Journal of Epidemiology*, **165**, 453-463.
- NELSON, K.P. and EDWARDS, D. (2007). A model and measure of agreement for population-based studies. *Technical Report*, Department of Statistics, University of South Carolina.
- NELSON, J.C. and PEPE, M.S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research*, **9**, 475-496.

- SCHUSTER, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology*, **55**, 289-303.
- SCOTT, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, **19**, 321-325.
- SHROUT, P.E. and FLEISS, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, **2**, 420-428.
- TANNER, M.A. and YOUNG, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, **80 (389)**, 175-180.
- THOMPSON, J.R. (2001). Estimating equations for kappa statistics. *Statistics in Medicine*, **20**, 2895-2906.
- UEBERSAX, J.S. and GROVE, W.M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, **9**, 559-572.
- WILLIAMSON, J.M. and MANATUNGA, A.K. (1997). Assessing interrater agreement from dependent data. *Biometrics*, **54**, 707-714.
- YERULSHALMY, J. (1956). The importance of observer error in the interpretation of photofluorograms and the value of multiple readings. *Radiology and Mass Radiography*, , 110-124.

Acknowledgements

The authors are grateful for the support provided by the following grant from the United States' Institutes of Health R03CA114783-01A1. We thank Craig Beam for kindly provid-

ing us with the breast cancer dataset. We are grateful for helpful advice and suggestions from Charles McCulloch, Patrick Graham and Robert Best.

Table 3: Parameter estimates for the Holmquist cancer dataset.

Parameter	Estimate	(Standard error)
η	-0.408	0.602
σ_u^2	8.491	2.213
σ_v^2	1.874	1.083
Cohen's kappa (true) κ	0.536	
Cohen's kappa (data-driven) $\hat{\kappa}$	0.512	
Model-based kappa κ_M	0.537	0.011

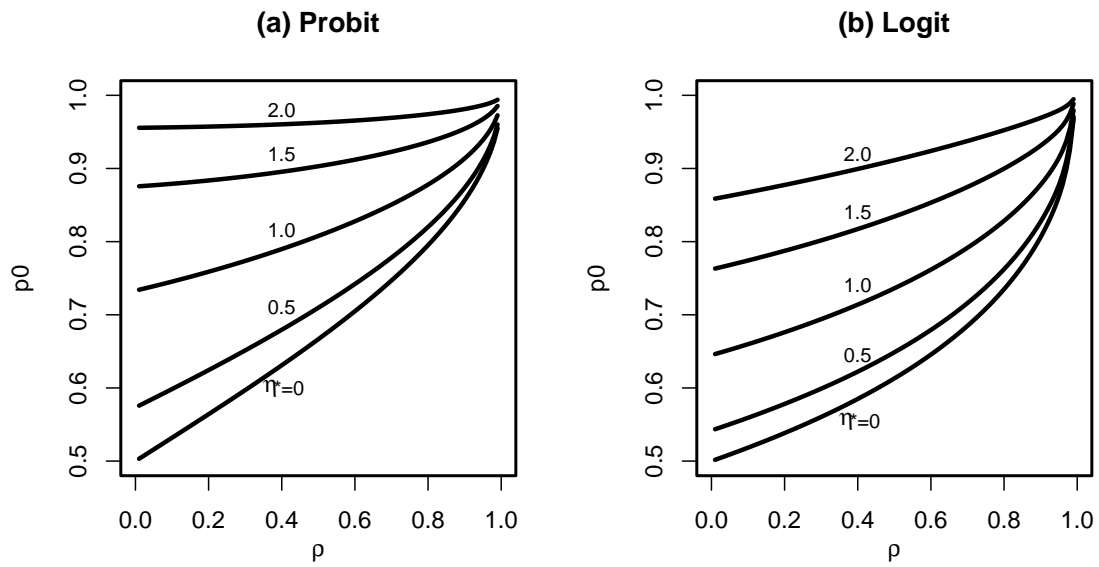


Figure 1: Plots of η^* for increasing ρ and p_0 under (a) the probit model and (b) the logit model.

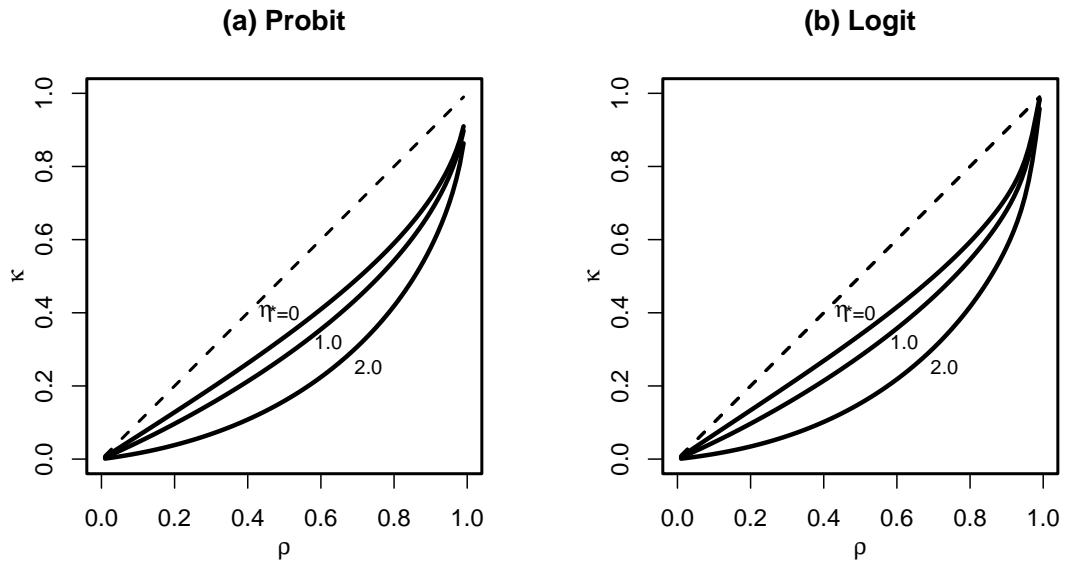


Figure 2: Plot of the true value of Cohen's kappa κ versus ρ for increasing standardized internal prevalence η^* under (a) the probit model and (b) the logit model.

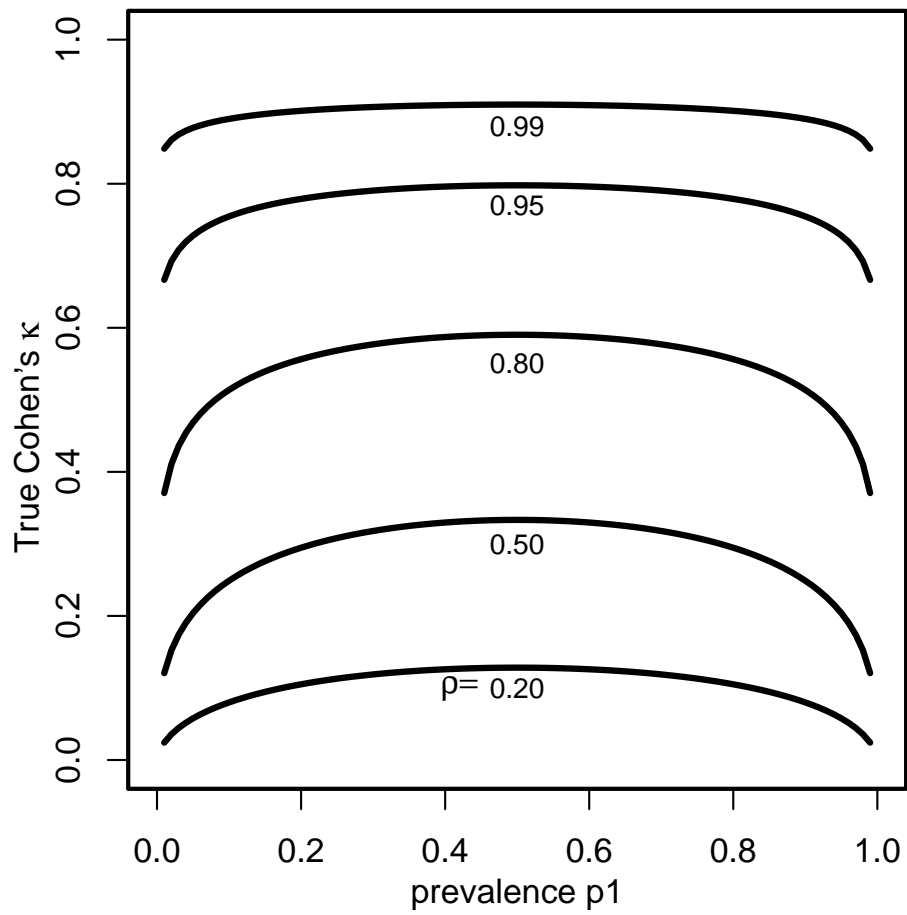


Figure 3: Plot of ρ for increasing values of the true Cohen's kappa versus prevalence p_1 under a population-based model.

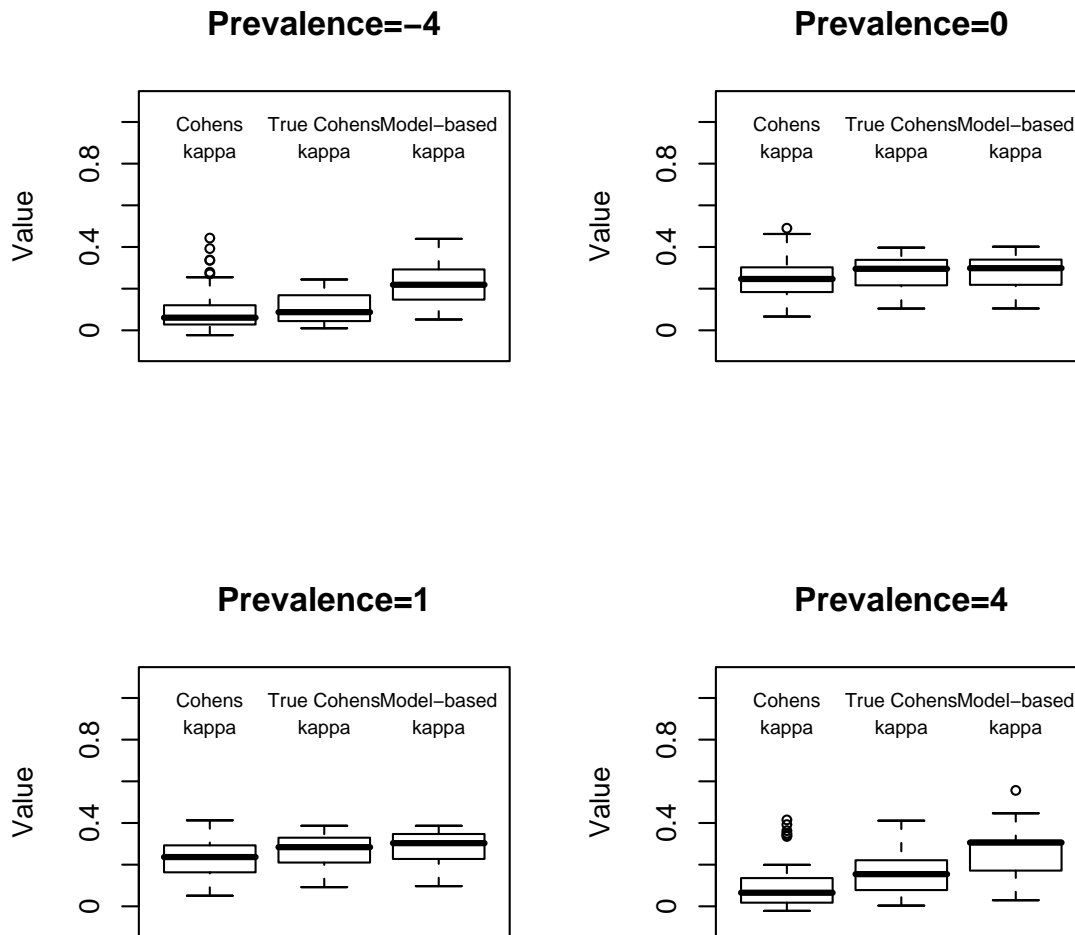


Figure 4: Boxplots of the three kappa statistics: Cohen's kappa, the true value of Cohen's kappa under a population-based model, and the model-based kappa. Results based upon simulation studies of one hundred datasets for different values of η , where $\sigma_u^2 = 2$ and $\sigma_v^2 = 2$.