PART 1

Statistical Models and Methods

Chapter 1

Unidimensionality, Agreement and Concordance Probability

The evaluation and comparison of various methods often arise in medical research. For example, the evaluation of reproducibility of a new measurement technique often needs a comparison with the established technique, and image interpretation is often read by two or more observers. In this chapter, we provide a review of the measures of agreement and association, describe the statistical models underlying the Cronbach's alpha coefficient (CAC) and the backward reliability curve (BRC), the kappa coefficient, and present a general approach based on the concept of concordance probability. In particular, we illustrate the relationship between the concordance probability and various existing measures of agreement and association, namely Kendall's τ, Somer's D, area under receiver operating characteristic (ROC) curve and Harrell's c-index. In addition, we review the estimation of concordance probability and present its large sample properties. Recent developments in the analysis of right censored data are also presented.

1.1. Introduction

The evaluation and comparison of various methods often arise in medical research. For example, the evaluation of reproducibility of a measurement technique often needs a comparison with the established technique, and the interpretation of a computerized tomography (CT) or magnetic resonance imaging (MRI) scan is often read by two or more observers. There is considerable literature on the measure of

---

Chapter written by Zhezhen JIN and Mounir MESBAH.

agreement (see [CHO 04], [BAR 07], [WAT 10], [SHO 04] and [LIN 10]). The methods vary with different types of measurement, i.e. continuous or categorical measurements. When the response variable is continuous, there are several intuitive approaches, namely comparison of means, Cronbach's coefficient alpha (CAC), various correlation coefficients and the test of slope being 1 in a simple linear regression, as well as alternative methods, the limits of agreement [BLA 86, BLA 99], the concordance correlation coefficient [LIN 89], mean squared deviation and total deviation index [LIN 00], and coverage probability approach [LIN 02]. When the response variable is categorical, kappa statistic, Somer's D-statistic and logistic regression are commonly used. When one measure is binary and the other measure is continuous, the methods of the ROC curve and logistic regression approach are often applied. These methods are related to typical concordance correlation between repeated measurements through an underlying linear or nonlinear parametric model. Recently developed concordance probability is a non-parametric approach. The concordance probability is commonly used as a measure of discriminatory power and predictive accuracy of statistical models. We show that the concordance probability also provides a unified measure of agreement for different types of measurement.

In this chapter, we present a review of the statistical models underlying the CAC and the BRC in section 1.2, and the kappa coefficient in section 1.3. In section 1.4, we introduce the concordance probability and describe its relationship with Kendall's τ, Somer's D and area of ROC curve of sensitivity and 1–specificity for different cutoffs. In section 1.5, we review the estimation of concordance probability and present its large sample properties. In section 1.6, we present recent developments on how to use the concordance probability to assess the agreement among different measures. We present the extension of the approach to the right censored data in section 1.7 and conclude with some discussion in section 1.8.

1.2. From reliability to unidimensionality: CAC and curve

1.2.1. *Classical unidimensional models for measurement*

Latent variable models involve a set of observable variables $A = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k\}$ and a latent (unobservable) variable θ of dimension $d \leq k$. In such models, the dimensionality of A is captured by the dimension of θ, the value of d. When $d = 1$, the dimensionality of set A is called unidimensional.

In a health-related quality of life (HrQoL) study, measurements are taken with an instrument: the questionnaire, which consists of questions (or items). In such cases, the $\mathbf{X}_{ij}$ represents the random response of the jth question by the ith subject and the $\mathbf{X}_j$ denotes the random variable generating responses to the jth question.

The parallel model is a classical latent variable model describing the unidimensionality of a set $A = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k\}$ of quantitative observable

variables. Let $\mathbf{X}_{ij}$ be the measurement of subject i, given by a variable $\mathbf{X}_j$, $i = 1, \ldots, n, j = 1, \ldots, k$, then:

$$\mathbf{X}_{ij} = \tau_{ij} + \varepsilon_{ij}, \qquad [1.1]$$

where τ_{ij} is the unknown true measurement corresponding to the observed measurement $\mathbf{X}_{ij}$ and ε_{ij} a measurement error. The model is called a parallel model if the τ_{ij} can be divided as:

$$\tau_{ij} = \beta_j + \theta_i,$$

where β_j is an unknown fixed parameter (non-random) representing the effect of the jth variable, and θ_i is an unknown random parameter effect of the ith subject.

It is generally assumed that θ_i has zero mean and unknown standard deviation σ_θ. It should be noted that the zero-mean assumption is an arbitrary identifiability constraint with consequence on the interpretation of the parameter: its value must be interpreted comparatively to the mean population value. *In HrQoL setting, θ_i is the true latent HrQoL that the clinician or health scientist wants to measure and analyze*. It is a zero mean individual random part of all observed subject responses $\mathbf{X}_{ij}$, the same whatever the variable $\mathbf{X}_j$ (in practice, a question j of an HrQoL questionnaire). It is also generally assumed that ε_{ij} are independent random errors with zero mean and standard deviation σ corresponding to the additional measurement error. Moreover, the true measure and the error are assumed to be uncorrelated, i.e. $cov(\theta_i, \varepsilon_{ij}) = 0$. This model is known as the parallel model, because the regression lines relating any observed item $\mathbf{X}_j$, $j = 1, \ldots, k$, and the true unique latent measure θ_i are parallel.

Model [1.1] can be obtained in an alternative way through modeling the conditional moments of the observed responses. Specifically, the conditional mean of $\mathbf{X}_{ij}$ can be specified as:

$$E[\mathbf{X}_{ij}|\theta_i; \beta_j] = \beta_j + \theta_i, \qquad [1.2]$$

where β_j, $j = 1, \ldots, k$, are fixed effects and θ_i, $i = 1, \ldots, n$, are independent random effects with zero mean and standard deviation σ_θ. The conditional variance of $\mathbf{X}_{ij}$ is specified as:

$$Var[\mathbf{X}_{ij}|\theta_i; \beta_j] = Var(\varepsilon_{ij}) = \sigma^2. \qquad [1.3]$$

Assumptions [1.2] and [1.3] are classical in experimental design. The model defines relationships between different kinds of variable: the observed score $\mathbf{X}_{ij}$, the true score τ_{ij} and the measurement error ε_{ij}. It is significant to make some remarks about the assumptions underlying this model. The random part of the true measure

given by response by the ith individual does not vary with the question number j as the θ_i does not depend on j, $j = 1, \ldots, k$. The model is unidimensional in the sense that the random part of all observed variables (questions $\mathbf{X}_j$) is generated by the common unobserved variable (θ_i). More precisely, let $\mathbf{X}^*_{ij} = \mathbf{X}_{ij} - \beta_j$ be the calibrated version of the response to the jth item by the ith subject, then models [1.2] and [1.3] can be rewritten as:

$$E[X^*_{ij}|\theta_i; \beta_j] = \theta_i, \text{ for } \forall j, \quad [1.4]$$

along with the same assumptions on β and θ and the conditional variance model [1.3].

When both θ_i and ε_{ij} are normally distributed, then we have the so-called conditional independence property: whatever j and j', two observed items $\mathbf{X}_j$ and $\mathbf{X}_{j'}$ are independent conditional to the latent θ_i.

1.2.2. *Reliability of an instrument: CAC*

A measurement instrument yields values that we call the observed measure. The reliability ρ of an instrument is defined as the ratio of two variances of the true over the observed measure. Under the parallel model, we can show that the reliability of any variable $\mathbf{X}_j$ (as an instrument to measure the true value) is given by:

$$\rho = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2}. \quad [1.5]$$

This coefficient is also known as the intra-class coefficient. The reliability coefficient, ρ, can easily be interpreted as a correlation coefficient between the true measure and the observed measure. When the parallel model is assumed, the reliability of the sum of k variables is:

$$\tilde{\rho}_k = \frac{k\rho}{k\rho + (1 - \rho)}. \quad [1.6]$$

This formula is known as the Spearman–Brown formula [BRO 10, SPE 10].

The Spearman–Brown formula shows a simple relationship between $\tilde{\rho}_k$ and k, the number of variables. It is easy to see that $\tilde{\rho}_k$ is an increasing function of k.

The maximum likelihood estimator of $\tilde{\rho}_k$, under the parallel model with normal distribution assumptions, is known as CAC [CRO 51, BLA 97], which is denoted as α:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} S^2_j}{S^2_{tot}}\right), \quad [1.7]$$

where

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2$$

and

$$S_{tot}^2 = \frac{1}{nk-1}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \overline{X})^2.$$

Under the parallel model, the variance–covariance matrix of the observed items X_j and the latent trait θ is:

$$V_{X,\theta} = \begin{pmatrix} \sigma_\theta^2 + \sigma^2 & \sigma_\theta^2 & \cdots & \cdots \sigma_\theta^2 & \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma_\theta^2 + \sigma^2 & \sigma_\theta^2 & \cdots \sigma_\theta^2 & \sigma_\theta^2 \\ : & : & : & : & : \\ \sigma_\theta^2 & \cdots & \sigma_\theta^2 & \sigma_\theta^2 + \sigma^2 & \sigma_\theta^2 \\ \sigma_\theta^2 & \cdots & \cdots & \sigma_\theta^2 & \sigma_\theta^2 \end{pmatrix},$$

and the corresponding correlation matrix of the observed items X_j and the latent trait θ is:

$$R_{X,\theta} = \begin{pmatrix} 1 & \rho & \cdots & \cdots \rho & \sqrt{\rho} \\ \rho & 1 & \rho & \cdots \rho & \sqrt{\rho} \\ : & : & : & : & : \\ \rho & \cdots & \rho & 1 & \sqrt{\rho} \\ \sqrt{\rho} & \cdots & \cdots & \sqrt{\rho} & 1 \end{pmatrix}.$$

The *marginal* covariance V_X and correlation matrix R_X of the k observed variables X_j, under the parallel model, are:

$$V_X = \begin{pmatrix} \sigma_\theta^2 + \sigma^2 & \sigma_\theta^2 & \cdots & \cdots \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma_\theta^2 + \sigma^2 & \sigma_\theta^2 & \cdots \sigma_\theta^2 \\ : & : & : & : \\ \sigma_\theta^2 & \cdots & \sigma_\theta^2 & \sigma_\theta^2 + \sigma^2 \end{pmatrix}$$

and

$$R_X = \begin{pmatrix} 1 & \rho & \cdots & \cdots \rho \\ \rho & 1 & \rho & \cdots \rho \\ : & : & : & : \\ \rho & \cdots & \rho & 1 \end{pmatrix}.$$

This structure is known as a *compound symmetry*-type structure. It is easy to show that the reliability of the sum of k items given in [1.7] can be expressed as:

$$\tilde{\rho}_k = \frac{k}{k-1}\left[1 - \frac{trace(V_X)}{J'V_XJ}\right], \qquad [1.8]$$

with J a vector with all components being 1, and

$$\alpha = \frac{k}{k-1}\left[1 - \frac{trace(S_X)}{J'S_XJ}\right], \qquad [1.9]$$

where S_X is the observed variance, empirical estimation of S_X. There is, even in the recent literature, an understanable confusion between Cronbach's alpha as a population parameter (theoretical reliability of the sum of items) or its sample estimate.

In addition, it is easy to show a direct connection between the CAC and the percentage of variance of the first component in principal component analysis (PCA), which is often used to assess unidimensionality. The PCA is mainly based on the analysis of the latent roots of V_X or R_X (or, in practice their sample estimate). The matrix R_X has only two different latent roots, the greater root is $\lambda_1 = (k-1)\rho + 1$, and the other multiple roots are $\lambda_2 = \lambda_3 = \lambda_4 = \cdots = 1 - \rho = \frac{k-\lambda_1}{k-1}$. So, using the Spearman–Brown formula, we can express the reliability of the sum of the k variables as $\tilde{\rho}_k = \frac{k}{k-1}(1 - \frac{1}{\lambda_1})$.

This clearly indicates a monotonic relationship between $\tilde{\rho}_k$, which can be consistently estimated by the CAC, and the first latent root λ_{x}, which in practice is naturally estimated by the corresponding observed sample correlation matrix and thus the percentage of variance of the first principal component in a PCA. So, CAC can also be considered as a measure of unidimensionality.

Nevertheless such a measure is not very useful, because it is easy to show, using the Spearman–Brown formula [BRO 10, SPE 10], that under the parallel model assumption, the reliability of the total score is an increasing function of the number of variables.

Therefore, *if the parallel model is true*, increasing the number of items will increase the reliability of a questionnaire. Moreover, the coefficient lies between 0 and 1. Zero value indicates a totally unreliable scale, while unit value means that the scale is perfectly reliable. Of course, in practice, these two scenarios never occur.

The CAC is an estimate of the reliability of the raw-score (sum of item responses) of a person *if the model generating those responses is a parallel model*.

The result can be used as a criterion for checking the unidimensionality of such responses when those item responses are *generated by a parallel model*.

In the next section, we show how to build and to use a more operational and more valid criterion to measure the unidimensionality of a set of items: the BRC (the α-curve).

1.2.3. *Unidimensionality of an instrument: BRC*

Statistical validation of unidimensionality can be performed through a goodness-of-fit test of the parallel model or Rasch model. There is a vast literature on the subject, see [MES 12]. The goodness-of-fit tests generally do not have power because their null hypotheses do not focus on unidimensionality: this includes indirectly other additional assumptions (the normality for parallel models, local independence for Rasch models etc.) As a result, the departure from the null hypotheses is not necessarily an indication of the departure from a unidimensionality.

In the following, we describe a graphical tool, which is helpful for checking the unidimensionality of a set of variables. It draws a curve in a stepwise manner, using estimates of reliability of subscores (total of a subset included in the starting set).

In the first step, the CAC will be calculated with all the variables. Then, at every successive step, the CAC will be calculated by deleting one variable each time, and the variable whose deletion yields the maximum CAC value among those CAC values will be removed. This procedure is repeated until only two variables remain. If the parallel model is true, increasing the number of variables increases the reliability of the total score, which can be consistently estimated by Cronbach's alpha. The number of variables and the CAC values can be plotted, which would yield a curve. This procedure is called the backward reliability curve (BRC). If there is a decrease of such a curve after adding a variable, it would strongly indicate that the added variable does not constitute a unidimensional set with variables already in the curve.

Drawing the BRC of a set of *unidimensional* items is an essential tool in the validation process of an HrQoL questionnaire. When we develop an HrQoL questionnaire, the main goal is generally to measure some unidimensional latent subjective traits (such as sociability and mobility). The use of the BRC in empirical data is very helpful for detection of non-unidimensional subsets of items. When the BRC is not an increasing curve, we can remove one or more items to obtain an increasing curve [MES 13]. If the reduced set gives an increasing curve, it is in some sense *more valid in terms of unidimensionality* than the previous set.

1.3. Agreement between binary outcomes: the kappa coefficient

1.3.1. *The kappa model*

The CAC in the previous section is a natural estimate of a monotonic function of the intra-class coefficient. In the case of multiple items, the parallel model leads to equal variances among marginal distributions of the items and equal covariances between any pairs of items. It can be shown that the formula of an intra-class coefficient is a correlation coefficient under such a constraint [MAK 88]. We illustrate this with the simple case of two binary items X and Y. Let us assume that $E(X) = E(Y) = \pi$. Under such an assumption, it is easy to derive the correlation coefficient between X and Y: $\rho_{X,Y} = \frac{E(XY)-\pi^2}{\pi(1-\pi)}$. The $\rho_{X,Y}$ is often denoted as the coefficient κ. It follows that:

$$\begin{aligned} p_{11} &= Prob(X = 1, Y = 1) = E(XY) = \pi^2 + \pi(1-\pi)\kappa, \\ p_{10} &= Prob(X = 1, Y = 0) = \pi(1-\pi) - \pi(1-\pi)\kappa, \\ p_{01} &= Prob(X = 0, Y = 1) = \pi(1-\pi)(1-\kappa), \end{aligned}$$

and

$$p_{00} = Prob(X = 0, Y = 0) = (1-\pi)^2 + \pi(1-\pi)\kappa.$$

1.3.2. *The kappa coefficient*

The probability of concordance between the two items is:

$$p_c = p_{11} + p_{00} = 2\pi^2 + 1 - 2\pi + 2\pi(1-\pi)\kappa.$$

The probability of concordance due to chance, i.e. when the two items are independent ($\kappa = 0$), is:

$$p_h = 1 - 2\pi(1-\pi).$$

Consequently:

$$\kappa = \frac{p_c - p_h}{1 - p_h}.$$

1.3.3. *Estimation of the kappa coefficient*

When we observe a 2×2 contingency table $\{n_{ij} : i, j = 0, 1\}$, a natural estimation of p_c is:

$$\hat{p_c} = \frac{n_{11} + n_{00}}{n}.$$

A natural estimation of p_h is:

$$\hat{p_h} = \frac{n_{1.}n_{.1} + n_{0.}n_{.0}}{n}.$$

Cohen's kappa coefficient [COH 60] is:

$$\hat{\kappa} = \frac{\hat{p_c} - \hat{p_h}}{1 - \hat{p_h}} = \frac{n_{11} + n_{00} - n_{1.}n_{.1} - n_{0.}n_{.0}}{n - n_{1.}n_{.1} - n_{0.}n_{.0}},$$

which is a natural estimator of the κ coefficient.

The κ can also be estimated by the maximum likelihood estimation method. Specifically, the likelihood of the observations is:

$$L = (p_{11})_{11}^n \times (p_{10})_{10}^n \times (p_{01})_{01}^n \times (p_{00})_{00}^n.$$

So, we easily find:

$$\begin{aligned} L(\kappa) &= (\pi^2 + \pi(1-\pi)\kappa)_{11}^n \times (\pi(1-\pi) - \pi(1-\pi)\kappa)_{10}^n \\ &\quad \times (\pi(1-\pi)(1-\kappa))_{01}^n \times ((1-\pi)^2 + \pi(1-\pi)\kappa)_{00}^n. \end{aligned}$$

The maximum likelihood estimator of κ can be obtained by maximizing $L(\kappa)$ over κ.

1.4. Concordance probability

For a pair of bivariate observations (X_1, Y_1) and (X_2, Y_2), the concordance probability is defined as:

$$C_{X,Y} = P\{Y_2 > Y_1 | X_2 > X_1\}.$$

As it is defined as a conditional probability, the concordance probability can be used to assess the relationship between two variables that have a natural ordering. In particular, it is useful for the assessment of monotonic correlation between two variables. The concordance probability is invariant to rank-preserving transformation on either X or Y. It takes values between 0 and 1. If $C_{X,Y} = 0$, then it means that X and Y are inversely related. If $C_{X,Y} = 1$, then it means that X and Y are related. If X and Y are independent, then $C_{X,Y} = 0.5$. In addition, if Y is continuous, then it is easy to see that:

$$1 - C_{X,Y} = P\{Y_1 > Y_2 | X_2 > X_1\} = C_{-X,Y},$$

as $P(Y_1 = Y_2) = 0$.

1.4.1. *Relationship with Kendall's τ measure*

Kendall's τ measure has been used for the assessment of the relationship between two random variables. The concordance probability is closely related to Kendall's τ measure.

For a pair of bivariate observations (X_1, Y_1) and (X_2, Y_2), Kendall's τ is defined as:

$$\tau_{X,Y} = E[\text{sign}(X_2 - X_1)\text{sign}(Y_2 - Y_1)].$$

It is easy to see that $\tau_{X,Y} = \tau_{Y,X}$, which shows that Kendall's τ is a symmetric measure. In addition:

$$\begin{aligned} \tau_{X,Y} &= P(X_2 > X_1, Y_2 > Y_1) + P(X_2 < X_1, Y_2 < Y_1) \\ &\quad -P(X_2 > X_1, Y_2 < Y_1) - P(X_2 < X_1, Y_2 > Y_1), \end{aligned}$$

which means that Kendall's τ is the difference between two probabilities: the probability of concordance and the probability of discordance. Equivalently, it follows that:

$$\tau_{X,Y} = 4P(X_2 > X_1, Y_2 > Y_1) - 2P(X_2 > X_1) - 2P(Y_2 > Y_1) + 1.$$

If (X_1, Y_1) and (X_2, Y_2) are independent and identically distributed, then:

$$\tau_{X,Y} = 2P(Y_2 > Y_1 | X_2 > X_1) - 1 = 2C_{X,Y} - 1,$$

which shows that there is 1–1 correspondence between Kendall's τ and the concordance probability, namely a linear relationship.

1.4.2. *Relationship with Somer's D measure*

Somer's D is defined as:

$$D_{X,Y} = E[\text{sign}(X_2 - X_1)\text{sign}(Y_2 - Y_1) | X_2 \neq X_1].$$

It is the difference between two conditional probabilities: the conditional probability of concordance and the conditional probability of discordance given that two X values are not equal. In light of the relationship between the concordance probability and Kendall's τ, it follows that $D_{X,Y} = \frac{2C_{X,Y}-1}{P(X_2 \neq X_1)}$.

Somer's D is often used as a measure of association for ordinal variables. For ordinal X, it is clear that $D_{X,Y} \neq D_{Y,X}$, which shows that Somer's D is an asymmetric measure.

If X is a continuous variable, then $P(X_2 \neq X_1) = 1$ and Somer's D is the same as Kendall's τ, which implies that $D_{X,Y} = 2C_{X,Y} - 1$.

If X is not a continuous variable, then $P(X_2 \neq X_1) < 1$ and Somer's D is larger than Kendall's τ, and $D_{X,Y} > 2C_{X,Y} - 1$.

1.4.3. *Relationship with ROC curve*

The area under the curve (AUC) is defined as the area under the ROC curve of 1-specificity against sensitivity. If X is binary with possible values 0 and 1, and Y is continuous or ordinal, let $TPP(c) = P(Y \geq c|X = 1)$ be the true positive proportion or sensitivity, and $FPP(c) = P(Y \geq c|X = 0)$ the false positive proportion or 1-specificity, then the AUC will be the area under the set of points $\{(FPP(c), TPP(c)) : c \in (-\infty, \infty)\}$.

If Y is continuous, then the AUC is the $C_{X,Y}$.

PROOF.– By definition $C_{X,Y} = P\{Y_2 > Y_1|X_2 > X_1\}$, therefore:

$$\begin{aligned} C_{X,Y} &= \int_{-\infty}^{\infty} P\{Y_2 > c|X_2 > X_1\} dP\{Y_1 \leq c|X_2 > X_1\} \\ &= \int_{-\infty}^{\infty} P\{Y_2 \geq c|X_2 = 1\} dP\{Y_1 < c|X_1 = 0\} \\ &= \int_{-\infty}^{\infty} TPP(c) d(1 - FPP(c)) \\ &= \int_{0}^{1} TPP(FPP^{-1}(u)) du. \end{aligned}$$

Here, $X_2 = 1$ and $X_1 = 0$ because both X_1 and X_2 are binary with values 0 and 1 and $X_2 > X_1$.

If Y is not continuous, then the AUC is $C_{X,Y} + P(Y_2 = Y_1|X_2 = 1, X_1 = 0)/2$.

1.5. Estimation and inference

Suppose $(X_1, Y_1), \cdots, (X_n, Y_n)$ are a sample of n-independent and identically distributed bivariate random vectors. The concordance probability can be estimated by:

$$\hat{C}_{X,Y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I\{X_i > X_j\} I\{Y_i > Y_j\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\{X_i > X_j\}},$$

where $I(\cdot)$ is an indicator function taking values of 0 or 1.

The estimator is of the form of a U-statistic. Using the large sample theory of a U-statistic, it can be shown that $\hat{C}_{X,Y} \to C_{X,Y}$ as $n \to \infty$ and $\sqrt{n}(\hat{C}_{X,Y} - C_{X,Y}) \to N(0, \sigma_1^2)$ as $n \to \infty$. The variance σ_1^2 can be estimated with a plug-in method with corresponding empirical estimators.

1.6. Measure of agreement

The concordance probability can be used as a measure of agreement. Suppose X is a latent variable, and Y and Z are two measures of the latent variable, we may want to assess if Y or Z is in agreement of measuring X. Liu *et al.* [LIU 12] suggested that this can be done by a comparison of concordance probabilities $C_{X,Y}$ and $C_{X,Z}$ with their difference, namely:

$$d_{YZ|X} = P(Y_2 > Y_1 | X_2 > X_1) - P(Z_2 > Z_1 | X_2 > X_1).$$

If the difference is close to 0, it means that Y and Z are in agreement. If the difference is different from 0, it means that Y and Z are not in agreement measuring the latent variable X: a positive value means that Y is better than Z in agreement with X, and a negative value means that Z is better than Y in agreement with X.

For independent and identically distributed $(X_i, Y_i, Z_i), i = 1, \cdots, n$, the difference can be estimated by:

$$\hat{d}_{YZ|X} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} [I\{Y_i > Y_j, X_i > X_j\} - I\{Z_i > Z_j, X_i > X_j\}]}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\{X_i > X_j\}},$$

which is a ratio of two U-statistics. By using the large sample theory of a U-statistic, it follows that $\hat{d}_{YZ|X} \to d_{YZ|X}$ and $\sqrt{n}(\hat{d}_{YZ|X} - d_{YZ|X}) \to N(0, \sigma_2^2)$.

The asymptotic results can be used to construct a statistical test for the hypotheses $H_0 : C_{X,Y} = C_{X,Z}$ versus $H_1 : C_{X,Y} \neq C_{X,Z}$, which is equivalent to the hypotheses $H_0 : d_{YZ|X} = 0$ versus $H_1 : d_{YZ|X} \neq 0$.

Liu *et al*. [LIU 12] noted that the denominator $\sum_{i=1}^{n}\sum_{j=1}^{n} I\{X_i > X_j\}$ in $\hat{d}_{YZ|X}$ does not play a significant role in the test and evaluation of the agreement, and the difference:

$$\hat{\Delta}_{YZ|X} = \sum_{i=1}^{n}\sum_{j=1}^{n}[I\{Y_i > Y_j, X_i > X_j\} - I\{Z_i > Z_j, X_i > X_j\}]$$

can be used. Obviously, $\hat{\Delta}_{YZ|X}$ is a U-statistic with kernel U_{ij} where:

$$\begin{aligned} U_{ij} &= 0.5[I\{Y_i > Y_j, X_i > X_j\} + I\{Y_i < Y_j, X_i < X_j\} \\ &\quad -I\{Z_i > Z_j, X_i > X_j\} - I\{Z_i < Z_j, X_i < X_j\}] \\ &= 0.5\{\text{sign}^{+}[(X_i - X_j)(Y_i - Y_j)] - \text{sign}^{+}[(X_i - X_j)(Z_i - Z_j)]\}. \end{aligned}$$

Consequently,

$$Var(\hat{\Delta}_{YZ|X}) = \frac{n(n-1)}{2}Var(U_{12}) + \frac{n}{4}Cov(U_{12}, U_{13}),$$

which can be estimated by replacing $Var(U_{12})$ and $Cov(U_{12}, U_{13})$ by their corresponding moment estimators.

The inference can be carried out by examining the $1-\alpha$ confidence interval of $\Delta_{YZ|X}$ or constructing a Wald-type test statistic $\frac{\hat{\Delta}_{YZ|X}}{\sqrt{\widehat{Var}(\hat{\Delta}_{YZ|X})}}$ for testing null hypothesis $H_0 : d_{YZ|X} = 0$.

1.7. Extension to survival data

In survival analysis, the outcome variable is the time to event occurrence, where the event could be the initial diagnosis of a disease or death. It is likely that the time variable is right-censored due to dropout or study termination. Let T be the length of time to event occurrence, Q be the censoring time and X be a predictor variable. Then, the observed data consist of (Y, δ, X), where $Y = \min(T, Q)$ and $\delta = I\{T \leq Q\}$. It is of significance to estimate the concordance probability $C_{T,X} = P(X_2 > X_1 | T_2 > T_1)$ with the n-independent observations $(X_i,\ Y_i,\ \delta_i),\ i = 1, \cdots,\ n$.

1.7.1. *Harrell's c-index*

Harrell *et al*. [HAR 82, HAR 96] developed Harrell's c-index to estimate the concordance probability for survival data. Harrell's c-index is defined as:

$$c_{Y,X} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[\delta_i I\{Y_i < Y_j\}I\{X_i < X_j\} + \delta_j I\{Y_j < Y_i\}I\{X_j < X_i\}]}{\sum_{i=1}^{n}\sum_{j=1}^{n}[\delta_i I\{Y_i < Y_j\} + \delta_j I\{Y_j < Y_i\}]}.$$

Nam and D'Agostino [NAM 02] developed a method to estimate the standard error of the estimator. Pencina and D'Agostino [PEN 04] derived alternative formulas for standard error estimation using the relationship between the c-index and the modified Kendall's τ for bivariate correlation and investigated how to construct its confidence intervals.

If there is no censoring, i.e. $\delta_i = 1$ for $i = 1, \cdots, n$, then by the large sample theory of a U-statistic, $c_{Y,X} = c_{T,X} \rightarrow C_{T,X} = C_{Y,X}$, where $C_{T,X}$ is the concordance probability $P(X_2 > X_1 | T_2 > T_1)$.

If there is censoring, and censoring random variable Q is independent of T, $c_{Y,X}$ will converge to a quantity that will depend on censoring distribution [LIU 09] and the consistency to a concordance probability will become questionable. Specifically, for continuous X, the estimator $c_{Y,X}$ will converge to $P(X_2 > X_1 | T_2 > T_1, T_1 < Q_1, T_1 < Q_2)$ [LIU 09].

To estimate the concordance probability $C_{T,X} = P(X_2 > X_1 | T_2 > T_1)$ properly with the n-independent observations $(X_i,\ Y_i,\ \delta_i),\ i = 1, \cdots,\ n$, Liu and Jin [LIN 09] proposed a modified estimator using the idea of inverse probability weighting:

$$\hat{c}^*_{Y,X} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(X_i < X_j) I(Y_i < Y_j) \delta_i / G^2(Y_i)}{\sum_{i=1}^{n} \sum_{j=1}^{n} I(Y_i < Y_j) \delta_i / G^2(Y_i)},$$

where $G(t) = P(t < Q)$ for $t > 0$. If $G(t)$ is unknown, a consistent estimator $\hat{G}(t)$, constructed by the Kaplan–Meier product limit method, may be used. For $Y_{(n)} = \max_{1 \leq i \leq n} Y_i$, it is required to set $b_{(n)} = 0$ if $\delta_{(n)} = 0$ and $\hat{G}(Y_{(n)}) = 0$. The estimator is consistent to the concordance probability $C_{T,X}$ if censoring variable Q is completely random and independent of T.

1.7.2. *Measure of discriminatory power*

Suppose that variables X and Z are quantitative predictors of the time to event variable T. To examine whether X and Z have the same discrimination accuracy for T, we may examine the difference between their concordance probabilities as $d_{XZ|T} = C_{T,X} - C_{T,Z}$, or, equivalently, the difference between the bivariate probabilities as:

$$\Delta_{XZ} = P(X_2 > X_1, T_2 > T_1) - P(Z_2 > Z_1, T_2 > T_1).$$

If there is no censoring, note that the difference between the bivariate probabilities Δ_{XZ} can be estimated by a U-statistic:

$$\begin{aligned}\hat{\Delta}_{XZ} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\text{sign}^{+}[(X_i - X_j)(T_i - T_j)] \\ &\quad -\text{sign}^{+}[(Z_i - Z_j)(T_i - T_j)]\} \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}U_{ij}.\end{aligned}$$

By the central limit theorem of U-statistics, under some regularity conditions, $\hat{\Delta}_{XZ}$ converges to Δ_{XZ} in distribution that $\sqrt{n}(\hat{\Delta}_{XZ} - \Delta_{XZ}) \rightarrow N(0, V)$ as $n \rightarrow \infty$. The variance V can be estimated empirically.

With random censoring, Liu *et al*. [LIU 12] have developed a modified estimator using the idea of inverse probability weighting:

$$\hat{\Delta}^{*}_{XZ} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i I(Y_i < Y_j)}{\hat{G}^2(Y_i)}[I(X_i < X_j) - I(Z_i < Z_j)],$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier product limit estimator of censoring distribution $G(\cdot)$. Under regularity conditions, Liu *et al*. [LIU 12] showed that $\hat{\Delta}^{*}_{XZ}$ is consistent and $\sqrt{n}(\hat{\Delta}^{*}_{XZ} - \Delta_{XZ}) \rightarrow N(0, V^{*})$ as $n \rightarrow \infty$. The variance V^{*} can also be estimated empirically. As a result, the inference can be carried out by examining the $1-\alpha$ confidence interval of $\hat{\Delta}^{*}_{XZ}$ or constructing a Wald-type test statistic $\frac{\hat{\Delta}^{*}_{XZ}}{\sqrt{\hat{V}^{*}}}$ for testing null hypothesis $H_0 : \Delta_{XZ} = 0$.

1.8. Discussion

In this chapter, we have reviewed various concordance coefficient measures. In particular, we presented non-parametric concordance coefficient and concordance probability, and described its relationship with commonly used measures of association, Kendall's τ, Somer's D, AUC and Harrell's c-index. In addition, we have illustrated how to use the concordance probability to assess agreement and to evaluate the discriminatory power of two different predictors.

For more than two covariates, Gönen and Hellner [GON 05] proposed model-based estimation of concordance probability using the Cox proportional hazards model for right-censored data. However, it is challenging to estimate the

asymptotic variance of the resulting estimator of concordance probability as it involves a non-smooth and non-continuous function of unknown parameters. It is of significance to develop an alternative method to estimate the asymptotic variance efficiently.

It is also noted that the concordance probability is invariant to the monotone transformation as its estimator only depends on ordering. However, the difference between two concordance probabilities might not be sensitive for detecting small differences in two competitive measures; caution should be taken in applications.

1.9. Bibliography

[BAR 07] BARNHART H.X., HABER M.J., LIN L.I., "An overview on assessing agreement with continuous measurement", *Journal of Biopharmaceutical Statistics*, vol. 17, pp. 529–569, 2007.

[BLA 86] BLAND J.M., ALTMAN D.G., "Statistical methods for assessing agreement between two methods of clinical measurement", *Lancet*, vol. i, pp. 307–310, 1986.

[BLA 89] BLAND J.M., ALTMAN D.G., "Measuring agreement in method comparison studies", *Statistical Methods in Medical Research*, vol. 8, pp. 135–160, 1989.

[BLA 97] BLAND J.M., ALTMAN D.G., "Statistics notes: Cronbach's alpha", *British Medical Journal*, vol. 314, p. 572, 1997.

[BRO 10] BROWN W., "Some experimental results in the correlation of mental abilities", *British Journal of Psychology*, vol. 3, pp. 296–322, 1910.

[CHO 04] CHOUDHARY P.K., NAGARAJA H.N., "Measuring agreement in method comparison studies-a review", in BALAKRISHNAN N., KANNAN N., NAGARAJA H.N. (eds), *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, Birkhauser, Boston, MA, pp. 215–244, 2004.

[COH 60] COHEN J., "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[CRO 51] CRONBACH L.J., "Coefficient alpha and the internal structure of tests", *Psychometrika*, vol. 16, pp. 297–334, 1951.

[GON 05] GONON M., HELLER G., "Concordance probability and discriminatory power in proportional hazards regression", *Biometrika*, vol. 92, pp. 965–970, 2005.

[HAR 84] HARRELL F.E., LEE K.L., CALIFE R.M., *et al*., "Regression modeling strategies for improved prognostic prediction", *Statistics in Medicine*, vol. 3, no. 2, pp. 143–152, 1984.

[HAR 96] HARRELL F.E., LEE K.L., MARK D.B., "Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors", *Statistics in Medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[KEN 38] KENDALL M.G., "A new measure of rank correlation", *Biometrika*, vol. 30, pp. 81–93, 1938.

[KEN 49] KENDALL M.G., "Rank and product-moment correlation", *Biometrika*, vol. 36, pp. 177–193, 1949.

[LIN 89] LIN L.I., "A concordance correlation coefficient to evaluate reproducibility", *Biometrics*, vol. 45, pp. 255–268, 1989.

[LIN 00] LIN L.I., "Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence", *Statistics in Medicine*, vol. 19, pp. 255–270, 2000.

[LIN 02] LIN L.I., HEDAYAT A.S., SINHA B., *et al* "Statistical methods in assessing agreement: models, issues and tools", *Journal of American Statistical Association*, vol. 97, pp. 257–270, 2002.

[LIN 10] LIN L.I., HEDAYAT A.S., WU W., *Statistical Tools for Measuring Agreement*, Springer, New York, 2010.

[LIU 09] LIU X., JIN Z., "A non-parametric approach to scale reduction for uni-dimensional screening scales", *International Journal of Biostatistics*, vol. 5, no. 1, pp. 1–22, 2009.

[LIU 12] LIU X., JIN Z., GRAZIANO J.H., "Comparing paired biomarkers in predicting quantitative health outcome subject to random censoring", *Statistical Methods in Medical Research*, 2012.

[MAK 88] MAK T.K., "Analysing intraclass correlation for dichotomous variables", *Applied Statistics*, vol. 37, pp. 344–352, 1988.

[MES 12] MESBAH M., "Measurement and analysis of quality of life in epidemiology", in CHAKRABORTY R., RAO C.R., SEN P.K. (eds), *Handbook of Statistics*, *BioInformatics in Human Health and Heredity*, vol. 2, Chapter 15, Amsterdam, North Holland, 2012.

[MES 13] MESBAH M., "From measurement to analysis", in CHRISTENSEN K.B., KREINER S., MESBAH M. (eds), *Rasch Models in Health*, Chapter 13, ISTE, London, John Wiley & Sons, New York, 2013.

[NAM 02] NAM B.H., D'AGOSTINO R.B., *Discrimination Index, the Area under the ROC Curve, in Goodness-of-Fit tests and Model Validity*, Chapter 20, Birkhauser, Boston, MA, 2002.

[PEN 04] PENCINA M.J., D'AGOSTINO R.B., "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation", *Statistics in Medicine*, vol. 23, pp. 2109–2123, 2004.

[SHO 04] SHOUKRI M.M., *Measures of Interobserver Agreement*, Chapman & Hall/CRC, New York, 2004.

[SOM 62] SOMERS R.H., "A new asymmetric measure of association for ordinal variables", *American Sociological Review*, vol. 27, pp. 799–811, 1962.

[SPE 10] SPEARMAN C., "Correlation calculated from faulty data", *British Journal of Psychology*, vol. 3, pp. 271–295, 1910.

[WAT 10] WATSON P.F., PETRIE A., "Method agreement analysis: a review of correct methodology", *Theriogenology*, vol. 73, pp. 1167–1179, 2010.