Agreement Statistics for Binary, Ordinal and Continuous Data

Lawrence Lin

Agreement assessments are widely used in assessing the acceptability of a new or generic process, methodology and/or formulation in areas of lab performance, instrument/assay validation or method comparisons, statistical process control, goodness-of-fit, and individual bioequivalence. Successful applications in these situations require a sound understanding of both the underlying theory and practical problems in real life. This workshop seeks to blend theory and applications effectively and to present these two aspects with many practical examples.

The common theme is to assess the agreement between observations of an assay or rater (*Y)* and its target (reference) counterpart values (*X)*. Target values may be considered random or fixed. Random target values are measured with random error. Common random target values are the gold-standard measurements, being both well established and widely acceptable. Sometimes we may also be interested in comparing two methods without a designated gold-standard method, or in comparing two technicians, times, reagents, or the like by the same method. Common fixed target values are the expected values and known values, which will be discussed in the most basic model.

When there is a disagreement between methods, we need to know whether the source of the disagreement is due to a systematic shift (bias or inaccuracy) or random error (imprecision). Specific coefficients of accuracy and precision will be introduced to characterize these sources. This is particularly important in the medical-device environment because a systematic shift usually can be easily fixed through calibration, while a random error usually is a more cumbersome variation-reduction exercise.

We will consider un-scaled (absolute) and scaled (relative to the between-sample variance) agreement statistics for both continuous and categorical variables. In practically all estimation cases, the statistical inference for parameter estimates will be discussed. Knowledge of regression, correlation, the asymptotic delta method, U-statistics, generalized estimation equations (GEE), and the linear mixed-effect model would be helpful in understanding the material presented and discussed.

We will discuss definitions of precision, accuracy, and agreement, and as well as the pitfalls of some misleading approaches with continuous data. For continuous data, we will start with the basic scenario of assessing agreement of two assays/raters, each with only one measurement for continuous data. In this basic scenario, we will consider the case of random or fixed target values for un-scaled (absolute) and scaled (relative) indices with constant or proportional error structure.

For categorical data, we will introduce traditional approaches for categorical data with the basic scenario for un-scaled and scaled indices. In terms of scaled agreement statistics, we will present the convergence of approaches for categorical and continuous data, and their association with a modified intraclass correlation coefficient. The information presented here and those presented earlier for continuous data sets the stage for discussing unified approaches for assessing agreement.

We will discuss sample size and power calculations for the basic models for continuous data. We will also introduce a simplified approach for the calculations in which we know only the most basic historical information such as residual variance or coefficient of variation. We will present many practical examples in which we know only the most basic historical information such as residual variance or coefficient of variation.

We will consider a unified approach to evaluating agreement among multiple *(k)* raters, each with multiple replicates *(m)* for both continuous and categorical data. Under this general setting, intrarater precision, interrater agreement based on the average of m readings, and total-rater agreement based on individual readings will be discussed.

We will consider a flexible and general setting in which where the agreement of certain cases can be compared relative to the agreement of a chosen case for both continuous and categorical data. For example, to assess individual bioequivalence, we are interested in assessing the agreement of test and reference compounds relative to the agreement of the within-reference compound. As another example, in the medical-device environment, we often want to know whether the within-assay agreement of a newly developed assay is better than that of an existing assay.

We will show how to use the validated web tools and the interpretation of the outputs from the most basic cases to more comprehensive cases. Many practical examples will be presented throughout the workshop in a wide variety of situations for continuous and categorical data. The materials presented in this workshop will largely based on the newly published book by Springer entitled '*Statistical Tools for Measuring Agreement*' by Lin, Hedayat and Wu (2012).