Lawrence Lin
A. S. Hedayat
Wenting Wu

Statistical Tools for Measuring Agreement

Springer

Statistical Tools for Measuring Agreement

Lawrence Lin • A.S. Hedayat • Wenting Wu

Statistical Tools for Measuring Agreement

Springer

Lawrence Lin
Baxter International Inc., WG3-2S
Rt. 120 and Wilson Rd.
Round Lake, IL 60073, USA
lawrence_lin@baxter.com

Wenting Wu
Mayo Clinic
200 First Street SW.
Rochester, MN 55905, USA
wu.wenting@mayo.edu

A.S. Hedayat
Department of Mathematics, Statistics
and Computer Science
University of Illinois, Chicago
851 S. Morgan St.
Chicago, IL 60607-7045, USA
hedayat@uic.edu

ISBN 978-1-4614-0561-0 e-ISBN 978-1-4614-0562-7
DOI 10.1007/978-1-4614-0562-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011935222

To

Sha-Li, Juintow, Buortau, and Shintau Lin
Batool, Leyla, and Yashar Hedayat
Xujian and MingEn Li

Preface

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 book seeks to blend theory and applications effectively and to present these two aspects with many practical examples.

The common theme in agreement assessment is to assess the agreement between observations of assay or rater (Y) and their 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 of 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 or known values, which will be discussed in the most basic model presented in Chapters 2 and 3.

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 random error. 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 unscaled (absolute) and scaled (relative) agreement statistics for both continuous and categorical variables. Unscale agreement statistics are independent of between-sample variation, while the scale agreement statistics are relative to the between-sample variance. For continuous variables with proportional error, we often can simply apply a log transformation to the data and would evaluate percent changes rather than absolute differences. In practically all estimation cases, the statistical inference for parameter estimates will be discussed.

This book should appeal to a broad range of statisticians, researchers, practitioners, and students, in areas such as biomedical devices, psychology, and medical research in which agreement assessment is needed. Knowledge of regression, correlation, the asymptotic delta method, U-statistics, generalized estimation equations (GEE), and the mixed-effect model would be helpful in understanding the material presented and discussed in this book.

In Chapter 1, we will discuss definitions of precision, accuracy, and agreement, and discuss the pitfalls of some misleading approaches for continuous data.

In Chapter 2, we will start with the basic scenario of assessing agreement of two assays or 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 unscaled (absolute) and scaled (relative) indices with constant or proportional error structure.

In Chapter 3, we will introduce traditional approaches for categorical data with the basic scenario for unscaled 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 in this chapter and Chapter 2 sets the stage for discussing unified approaches in Chapters 5 and 6. In both Chapters 2 and 3, there is available a wealth of references to the basic model of agreement assessment. We will provide brief tours of related publications in these two chapters.

In Chapter 4, we will discuss sample size and power calculations for the basic models for continuous data. We will also introduce a simplified approach that is applicable to continuous and categorical data. We will present many practical examples in which we know only the most basic historical information such as residual variance or coefficient of variation.

In Chapter 5, 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.

In Chapter 6, 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 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. Both Chapters 5 and 6 are applicable to continuous and categorical data.

In Chapter 7, we will present a workshop using a continuous data set, a categorical data set, and an individual bioequivalence data set as examples. We will then address the use of SAS and R macros and the interpretation of the outputs from the most basic cases to more comprehensive cases.

This book is concise and concentrates on topics primarily based on the authors' research. However, proofs that were omitted from our published articles will be

presented, and all other related tools will be well referenced. Many practical examples will be presented throughout the book in a wide variety of situations for continuous and categorical data.

A book such as this cannot have been written without substantial assistance from others. We are indebted to the many contributors who have developed the theory and practice discussed in this book. We also would like to acknowledge our appreciation of the students at the University of Illinois at Chicago (UIC) who helped us in many ways. Specifically, six PhD dissertations on agreement subjects have been produced by Robieson (1999), Zhong (2001), Yang (2002), Wu (2005), Lou (2006) and Tang (2010). Their contributions have been the major sources for this book. Most of the typing using MikTeX was performed by the UIC PhD student Mr. Yue Yu, who also double-checked the accuracy of all the formulas.

We would like to mention that we have found the research into theory and application performed by Professors Tanya King, of the Pennsylvania State Hershey College of Medicine; Vernon Chinchilli, of the Pennsylvania State University College of Medicine; and Huiman Barnhart, of the Duke Clinical Research Institute, are truly inspirational. Their work has influenced our direction for developing the materials of our book. We are also indebted to Professor Phillip Schluter, of the School of Public Health and Psychosocial Studies at AUT University, New Zealand, for his permission to use the data presented in Examples 5.9.3 and 6.7.2 prior to their publication.

Finally, all SAS and R macros and most data in the examples are provided at the web sites shown below:

1. http://www.uic.edu/~hedayat/
2. http://mayoresearch.mayo.edu/biostat/sasmacros.cfm

The U.S. National Science Foundation supported this project under Grants DMS-06-03761 and DMS- 09-04125.

Round Lake, IL, USA — Lawrence Lin
Chicago, IL, USA — Samad Hedayat
Rochester, MN, USA — Wenting Wu

Contents

Symbols Used and Abbreviations

In this book, we use a Greek letter (symbol) to represent a parameter to be estimated, and we use its respective English letter or the symbol with a hat to represent its sample counterpart or estimate. The exception is that we use $\overline{X}$ to represent the sample mean, due to the long history of that convention. When a transformation is performed, we use an uppercase letter to represent a transformed estimate. However, there are some complicated computational formulas in which we use uppercase letters to simplify the computation. In the sequel, we use Greek letters to represent parameters when the target value X is considered random. When the target value is considered fixed, we add $|X$ as a subscript to the corresponding parameter. For example, $\varepsilon^2_{|X}$ represents the mean squared deviation (MSD) when the target value X is assumed fixed. We use a boldface symbol or letter to represent a vector or matrix. Symbols and their corresponding definitions are listed below:

ε^2	Mean squared deviation
δ_{π_0}	Total deviation index
π_{δ_0}	Coverage probability
ρ_c	Concordance correlation coefficient
υ	Location shift
ϖ	Scale shift
χ_a	Accuracy coefficient
ρ	Precision coefficient
κ	Kappa
κ_w	Weighted kappa
ζ	Total-rater MSD to Intra-rater MSD ratio
ψ	Intra-rater MSD to Intra-rater MSD ratio
Δ	Relative bias squared

Abbreviations used in this book are (in alphabetical order):

CCC:	Concordance correlation coefficient
CDF:	Cumulative density function
CIA:	Coefficient of individual agreement
CL:	Confidence limit

CLIA: Clinical laboratory improvement amendments
CP: Coverage probability
GEE: Generalized estimation equations
GM: Geometric mean
ICC: Intraclass correlation coefficient
IIR: Intra rater MSD to Intra rater MSD ratio
ML: Maximum likelihood
MLE: Maximum likelihood estimate
MSD: Mean squared deviation
PT: Proficient testing
PTC: Proficient testing criterion
RBS: Relative bias squared
RML: Restricted maximum likelihood
RMLE: Restricted maximum likelihood estimate
SD: Standard deviation
TDI: Total deviation index based on absolute difference
TDI%: Total deviation index based on percent change
TIR: Total-rater MSD to Intra rater MSD ratio

Chapter 1
Introduction

Consider the problems of 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. The common theme is to assess the agreement between observations (Y) and their corresponding target values (X). Target values may be considered random or fixed. Commonly used random target values are the gold standard measurements, which are proven and widely acceptable. Commonly used fixed target values are the expected or known values. We might be interested in comparing two methods without a designated gold standard method. Sometimes, we may also be interested in comparing a newly developed assay that is alleged to be more precise and accurate than a designated gold standard assay. Within a method, we might be interested in comparing technicians/times/reagents.

For simplicity and the ease of reference, we will use the term *assays* and *raters* to represent assays, raters, instruments, methods, etc. Also, we will use the term *samples* to designate samples, patients, animals, or subjects. In the tradition of the subject matter, we use throughout this book the terms *index* and *coefficient* interchangeably.

Figure 1.1 presents a typical situation for assessing agreement. When we plot the observed values on the y-axis versus the corresponding target values over a desirable range on the x-axis, we would like to see agreement in the paired data so that the observations fall closely along the identity line, which is the straight line with zero intercept and unit slope. When there is evidence of disagreement, it is important to address the issue and search for the sources of that disagreement.

1.1 Precision, Accuracy, and Agreement

Generally, the common basic sources of disagreement come from within-sample variation (imprecision) and/or a shift in the marginal distributions (inaccuracy). Fixing imprecision is a within-sample variance reduction exercise in the medical-device

L. Lin et al., *Statistical Tools for Measuring Agreement*,
DOI 10.1007/978-1-4614-0562-7_1, © Springer Science+Business Media, LLC 2012