ASSESSMENT OF AGREEMENT
AND SELECTION OF THE BEST INSTRUMENT
IN METHOD COMPARISON STUDIES

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Pankaj K. Choudhary, M.Sc.

* * * * *

The Ohio State University

2002

Dissertation Committee:

Prof. H. N. Nagaraja, Adviser
Prof. M. W. Browne
Prof. J. C. Hsu
Prof. T. J. Santner

Approved by

Adviser
Department of Statistics
ABSTRACT

We consider three problems that arise in studies comparing two or more methods that measure a continuous variable of interest.

First is the basic problem of how to assess the degree of agreement between two methods (or instruments). If there is sufficient agreement between them, one can use them interchangeably or prefer the one that is cheaper or is easier to use. We employ tests of hypotheses to assess if the data have evidence for satisfactory agreement. In particular, we consider three formulations of satisfactory agreement and provide tests for the associated hypotheses. The formulations are: (i) both the mean and the standard deviation of the difference of measurements from the two instruments are close to zero; (ii) the means of the two measurements are close, their standard deviations are close, and their correlation is high; and (iii) a large proportion of the difference lies in an interval close to zero. In all the cases, the thresholds determining closeness are supplied by the investigator. The first two formulations give additional information regarding the nature and extent of disagreement when the data do not have evidence for satisfactory agreement. For the last formulation, we discuss both the parametric and nonparametric tests.

Comparison of two instruments with a gold standard is the focus of the other two problems. We first consider how to select the instrument that agrees most with a gold standard in terms of mean squared deviation. This instrument is designated as
the best one. For this selection problem, we present two large sample single-stage procedures and a two-stage procedure using the multiple comparisons with the best approach. Questions like which parameterization works well for the comparison and what sample sizes are adequate are answered using asymptotic theory and simulation.

For the last problem, we determine, through a hypotheses test, whether the best instrument agrees sufficiently well with the gold standard before proceeding to its selection. We describe a two-stage procedure for this purpose and study its properties using the asymptotic distribution of the test statistic and simulation.
To my parents and teachers.
ACKNOWLEDGMENTS

I express my heartfelt gratitude and deepest appreciation to my guru, Prof. H. N. Nagaraja for his guidance, support and thoughtfulness throughout my graduate studies. His fatherly nurturing has greatly helped me grow both in my personal and professional lives.

I am thankful to my wonderful committee members, Professors Michael Browne, Jason Hsu and Thomas Santner, for their encouragement, comments and suggestions. I also thank Prof. Hsu for the several helpful discussions we had related to my dissertation.

I am indebted to my teachers from India, in particular, Professors Debashree Goswami, Debasis Kundu and Divakar Sharma, for giving a direction to my life.

I thank Alex, Babis, Bijula and Subharup for their treasured friendship and help.

I also thank the Department of Statistics for their financial support that afforded my graduate education, and for creating an excellent learning environment.

Finally, I offer my deepest appreciation to my family and Swati. Their endless love, encouragement and support nourished me all through my graduate education.
VITA

November 5, 1975 ....................... Born - Muzaffarpur, India

1996 ................................. B.Sc. Statistics, University of Delhi


1998-2002 ............................ Graduate Teaching Associate and Graduate Research Associate, The Ohio State University.

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Statistics
TABLE OF CONTENTS

Abstract ................................................................. ii
Dedication ................................................................. iv
Acknowledgments ......................................................... v
Vita .............................................................. vi
List of Symbols ........................................................... x
List of Tables ........................................................... xi
List of Figures ......................................................... xii
Chapters:

1. Introduction ......................................................... 1

2. Assessing agreement between two instruments: Parametric approaches . 6
   2.1 Introduction and notation ............................................ 6
   2.2 Literature review ...................................................... 8
      2.2.1 Limits of agreement .............................................. 8
      2.2.2 Intraclass correlation ............................................. 9
      2.2.3 Concordance correlation and related measures ............... 14
   2.3 Test based on the differences ....................................... 17
   2.4 Test based on the bivariate data ................................... 29
   2.5 Discussion .......................................................... 34
<table>
<thead>
<tr>
<th>3.</th>
<th>Assessing agreement using coverage probability</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2</td>
<td>Literature review</td>
<td>39</td>
</tr>
<tr>
<td>3.2.1</td>
<td>TDI approach</td>
<td>39</td>
</tr>
<tr>
<td>3.2.2</td>
<td>CP approach</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Distribution-free approach</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Normality based approach</td>
<td>45</td>
</tr>
<tr>
<td>3.5</td>
<td>Tests based on tolerance intervals</td>
<td>48</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Distribution-free tests</td>
<td>51</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Normality based tests</td>
<td>53</td>
</tr>
<tr>
<td>3.6</td>
<td>Connection with acceptance sampling</td>
<td>57</td>
</tr>
<tr>
<td>3.7</td>
<td>Discussion</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4.</th>
<th>Selecting the instrument that is closest to a gold standard</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction and notation</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Current approaches</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>Single-stage procedures</td>
<td>65</td>
</tr>
<tr>
<td>4.4</td>
<td>A two-stage procedure</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5.</th>
<th>Assessing agreement with a gold standard and selecting the instrument that is closest to it</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>83</td>
</tr>
<tr>
<td>5.2</td>
<td>A two-stage procedure</td>
<td>84</td>
</tr>
<tr>
<td>5.3</td>
<td>Small sample properties</td>
<td>89</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>98</td>
</tr>
</tbody>
</table>

| 6. | Summary and future work | 100 |

Appendices:

<table>
<thead>
<tr>
<th>A.</th>
<th>Technical Lemmas</th>
<th>102</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.</td>
<td>Datasets</td>
<td>109</td>
</tr>
</tbody>
</table>

Bibliography | 112 |
LIST OF SYMBOLS

Symbol | Meaning
---|---
$X_1, X_2$ | Measurement on a typical subject by first, second instrument
$\xi_1, \xi_2, \eta_1, \eta_2, \nu$ | $E(X_1), E(X_2), SD(X_1), SD(X_2), Corr(X_1, X_2)$
$D$ | $X_2 - X_1$
$\mu, \sigma$ | $E(D), SD(D)$
$N(\mu, \sigma)$ | Normal distribution with mean $\mu$ and standard deviation $\sigma$
$t_k, \chi^2_k$ | $t$-distribution, $\chi^2$-distribution with $k$ degrees of freedom
BVN | Bivariate normal distribution
i.i.d. | Independently and identically distributed
$n$ | Sample size
pdf | Probability density function
cdf | Cumulative distribution function
$\phi(\cdot), \Phi(\cdot)$ | Standard normal pdf, cdf
$\Phi_2(\cdot, \cdot; \cdot)$ | Bivariate standard normal cdf
$H, K$ | Null, alternative hypothesis
$z(\alpha), t_k(\alpha), \chi^2_k(\alpha)$ | Upper $\alpha$-th quantile of $N(\mu, \sigma)$, $t_k$, $\chi^2_k$ distribution
$\pi_l, \pi_u, \pi_l', \pi_u'$ | $Pr(D < l)$, $Pr(D > u)$, $\pi_l + \pi_u$, $1 - \pi_l$, $1 - \pi_u$, $1 - \pi$
$G$ | Measurement on a typical subject by the gold standard
$D_1, D_2$ | $G - X_1, G - X_2$
$\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ | $E(D_1), E(D_2), SD(D_1), SD(D_2), Corr(D_1, D_2)$
$\theta_1, \theta_2, \psi_1, \psi_2, \gamma$ | $E(D_1^2), E(D_2^2), SD(D_1^2), SD(D_2^2), Corr(D_1^2, D_2^2)$
$\lambda_1, \lambda_2, \tau_1, \tau_2$ | $ln(\theta_1), ln(\theta_2), \psi_1/\theta_1, \psi_2/\theta_2$
$\psi^2_d, \tau^2_d$ | $\psi^2_1 + \psi^2_2 - 2\gamma \psi_1 \psi_2, \tau^2_1 + \tau^2_2 - 2\gamma \tau_1 \tau_2$
$m, N_m$ | First-stage, two-stage sample size
$[1], [2]$ | Index of the true best, worst instrument
$(1), (2)$ | Index of the sample best, worst instrument
$Y_n \overset{B}{\rightarrow} Y, Y_n \overset{d}{\rightarrow} Y$ | $Y_n$ convergence to $Y$ in probability, in distribution
$a_n \approx b_n$ | $a_n$ and $b_n$ have the same order of convergence
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Salient features of various approaches.</td>
</tr>
<tr>
<td>2.2</td>
<td>Sample sizes required to achieve 80% power at ((\mu, \sigma) = (0, \delta_\sigma/2)) with 5% level test of hypotheses (2.10) for various ((\delta_\mu, \delta_\sigma)).</td>
</tr>
<tr>
<td>3.1</td>
<td>Sample sizes required to achieve 80% power with normality based 5% level test of hypotheses (3.2).</td>
</tr>
<tr>
<td>4.1</td>
<td>Two-stage sample sizes required by the procedure (\mathcal{D}(\delta, \alpha, 15)) to satisfy the probability requirement (4.5) approximately.</td>
</tr>
<tr>
<td>5.1</td>
<td>Five-number summaries of the empirical probabilities that the procedure (\mathcal{D}_0) rejects (H), given in (5.1), and simultaneously selects the first instrument, and the associated ASN’s for ((\mu_1, \mu_2) \in \Omega_2(\theta_1, \theta_2)).</td>
</tr>
<tr>
<td>5.2</td>
<td>Two-stage sample sizes required by the procedure (\mathcal{D}(0.05, \beta, \lambda_0, \delta, 15)) to satisfy the probability requirement (5.2) approximately.</td>
</tr>
<tr>
<td>B.1</td>
<td>Fractional area change data from Hutson et al. (1998). Here GSC, EXP1 and ABD refer to the fractional area changes computed using the measurements by the committee of experts, the first expert and the computer algorithm, respectively.</td>
</tr>
<tr>
<td>B.2</td>
<td>Oxygen saturation data from Bland and Altman (1986). Here PM and OM refer to the oxygen saturation measurements (in %) using the pulse oximeter saturation method and the oxygen saturation meter, respectively.</td>
</tr>
<tr>
<td>B.3</td>
<td>Plasma volume data from Bland and Altman (1999). Here NM and HM refer to the plasma volume measurements (as % of normal) using the sets of normal values due to Nadler and Hurley, respectively.</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Sample size versus expected coverage of the interval within 95% limits of agreement, assuming $D$ follows $N(\mu, \sigma)$ distribution.</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Contours of the power function $p_1(0.05, 1.0, 1.0, n, \mu, \sigma)$ for $n = 15, 30$.</td>
<td>20</td>
</tr>
<tr>
<td>2.3 Surface plot of the power function $p_1(0.05, 1.0, 1.0, 15, \mu, \sigma)$.</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Scatter plot for the fractional area change data (Example 2.1).</td>
<td>22</td>
</tr>
<tr>
<td>2.5 Normal probability plot for the differences, $D = X_2 - X_1$, in Example 2.1.</td>
<td>23</td>
</tr>
<tr>
<td>2.6 Scatter plot for the plasma volume data (Example 2.2).</td>
<td>25</td>
</tr>
<tr>
<td>2.7 Histogram and the normal probability plot for the differences, $D = X_2 - X_1$, in Example 2.2.</td>
<td>26</td>
</tr>
<tr>
<td>2.8 Plots of means versus differences for the data in Example 2.1 (top) and Example 2.2 (bottom).</td>
<td>28</td>
</tr>
<tr>
<td>3.1 Scatter plot (top) and the plot of means versus differences (bottom) for the Oxygen saturation data (Example 3.1).</td>
<td>43</td>
</tr>
<tr>
<td>3.2 Histogram for the differences, $D = X_2 - X_1$, in Example 3.1.</td>
<td>44</td>
</tr>
<tr>
<td>3.3 Contours of the power function $p(-1, 0.20, 0.05, n, \mu, \sigma)$ for $n = 15, 30$.</td>
<td>47</td>
</tr>
<tr>
<td>3.4 Surface plot of the power function $p(1, 0.20, 0.05, 15, \mu, \sigma)$.</td>
<td>48</td>
</tr>
<tr>
<td>4.1 Empirical coverage probability of 95% confidence interval for $\theta_1/(\theta_1+\theta_2)$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$, for $n = 20, 30$.</td>
<td>66</td>
</tr>
</tbody>
</table>
4.2 Average Shapiro-Wilk $p$-values for testing normality of studentized estimators of $\theta_1 - \theta_2$ and $\lambda_1 - \lambda_2$ for $n = 15$ and $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$. 67

4.3 Average Shapiro-Wilk $p$-values for testing normality of studentized estimator of $\theta_1/(\theta_1 + \theta_2)$ for $n = 15$ and $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$. 69

4.4 Empirical coverage probabilities of 95% unconstrained MCB confidence interval for $\lambda_1 - \lambda_2$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$ (top) and its 95% constrained version (bottom), for $n = 15$. 73

4.5 Scatter plot for the differences in the fractional area change data. 74

4.6 Marginal normal probability plots for the differences in the fractional area change data. 75

4.7 Empirical coverage probability of the confidence interval for $\lambda_1 - \lambda_2$ given by the procedure $\mathcal{D}(\ln(1.2), \alpha, 15)$ (top) and the associated ASN (bottom), over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$. 80

5.1 Empirical type-I error probability for testing hypotheses (5.1) using the procedure $\mathcal{D}_0$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$ (top) and the associated ASN (bottom). 92

5.2 Difference between the empirical type-I error probability for testing hypotheses (5.1) using the procedure $\mathcal{D}_0$ with $z(\alpha)$ in place of $b(\alpha)$ and 0.05 for $(\mu_1, \mu_2) \in \Omega_3(1.0,1.2)$. 93

5.3 Empirical probability that the procedure $\mathcal{D}_0$ rejects $H$, given in (5.1), and simultaneously selects the first instrument (top) and the associated ASN (bottom), over the grid $(\mu_1, \mu_2) \in \Omega_3(1/1.2,1.0)$. 95

5.4 Location of the points $\zeta_0, \cdots, \zeta_4$ in the region $\left\{ \lambda_0 - \lambda_1 \geq \delta_0, \lambda_2 - \lambda_1 \geq \delta_0 \right\}$ with $\lambda_0 = 0$ and $\delta_0 = \ln(1.2)$. 96
CHAPTER 1

INTRODUCTION

A method comparison study refers to the comparison of two or more methods of measurement of a variable with the major goal of assessing agreement among them. If there is sufficient agreement, then they can be used interchangeably. In this case one may also prefer a cheaper, easier to use or a less invasive method. Typically there is a gold standard, a method traditionally considered to be accurate, among the methods being compared that serves as a reference. The focus then is on assessing its agreement with the other methods.

We concentrate on the case when the measurements are continuous. The problem of assessing agreement in the categorical measurements case has been discussed extensively (see Banerjee et al., 1999 for a recent review). The literature, however, in the continuous case is more recent. Bland and Altman (1986), the paper that introduced the limits of agreement approach for assessing agreement between two methods is one of the earliest references. It is also the most popular one: it had 6101 hits in the citation database of the Institute for Scientific Information at the time of writing. It was pronounced a “citation classic” paper by the institute in 1992. We argue in Section 2.2 that the complexity of this agreement problem warrants a more formal
analysis than the mainly graphical approach of Bland and Altman (1986). For convenience, from here on we use the terms “method of measurement” and “instrument” interchangeably.

The large number of citations for the Bland and Altman paper is a good indicator of the wide applicability of this problem. In the experimental sciences, there is always a thrust towards finding instruments that produce measurements closer to the truth and are more convenient to use at the same time. The Center of Devices and Radiological Health, the wing of FDA that regulates the approval of medical devices in the US, deals with this problem routinely.

We assume that all the instruments under comparison have the same unit of measurement. We also assume that only one measurement per subject is available from every instrument. In this setup, two instruments are in perfect agreement if all the pairs of measurements lie on the 45° line through the origin. Assessing agreement, then, refers to assessing the deviation from this line of equality.

The goal of this dissertation is to develop statistical methodologies for various problems arising in method comparison studies. In particular, we address the following three problems:

(A) How to assess whether the agreement between two instruments is satisfactory?

(B) How to select the instrument that agrees most with a gold standard when two instruments are compared with it? We refer to this instrument as the best one.

(C) How to determine at the outset if the agreement between the best instrument and the gold standard is good enough for practical purposes? Only when it is, we proceed to its selection.
In Chapters 2 and 3 we focus on Problem $A$. Let $(X_1, X_2)$ be the pair of measurements by the two instruments on a typical subject. We assume that $(X_1, X_2)$ has continuous bivariate distribution with mean $(\xi_1, \xi_2)$, variance $(\eta_1^2, \eta_2^2)$ and correlation $\nu$. The following commonly used random effects model would provide a setup for these measurements:

$$X_i = T + b_i + \epsilon_i, \quad i = 1, 2,$$

where $T$ is the true unobservable measurement for the typical subject, and $b_i$ and $\epsilon_i$ are, respectively, the bias and the random error associated with the $i$-th instrument. Here (i) $T$ follows a continuous distribution with mean $\xi_T$ and variance $\eta_T^2$ over the population of subjects, (ii) $\epsilon_i$ follows a continuous distribution with mean zero and variance $\eta_{\epsilon_i}^2$, and (iii) $(T, \epsilon_1, \epsilon_2)$ are mutually independent. Thus $\xi_i = \xi_T + b_i$, $\eta_i^2 = \eta_T^2 + \eta_{\epsilon_i}^2$ and $\nu = \eta_T^2 / (\eta_T^2 + \eta_{\epsilon_1}^2)(\eta_T^2 + \eta_{\epsilon_2}^2))^{1/2}$. This correlation $\nu$ is the square-root of the product of reliabilities of the two instruments. Let $D$ denote the difference $X_2 - X_1$. It has a continuous distribution with mean $\mu = \xi_2 - \xi_1$ and variance $\sigma^2 = \eta_1^2 + \eta_2^2 - 2\nu \eta_1 \eta_2$.

In Chapter 2 we consider two approaches for defining satisfactory agreement. One is based on the difference $D$, and asks whether $\mu$ and $\sigma$ are practically equivalent to zero. The other is based on the bivariate data $(X_1, X_2)$. It asks whether $\xi_2 - \xi_1$ is practically equivalent to zero, and $\eta_2 / \eta_1$ and $\nu$ are practically equivalent to one.

In Chapter 3 we work with $D$ again but define agreement in terms of coverage probability of a given threshold interval around zero. Let $l (< 0)$ and $u (> 0)$ be specified limits such that a difference lying in the interval $[l, u]$ is believed to be practically equivalent to zero. The agreement is satisfactory when $Pr(l \leq D \leq u)$ is sufficiently high.
Chapters 4 and 5 deal with the Problems \( \mathcal{B} \) and \( \mathcal{C} \), respectively. Let the triplet \((G, X_1, X_2)\) denote the measurements on a typical subject by the gold standard, the first and the second instrument, respectively. A setup for these measurements is given by the model (1.1), which is extended to include

\[
G = T + b_G + \epsilon_G,
\]

where \( b_G \) and \( \epsilon_G \) are the bias and the random error of the gold standard. It is additionally assumed that \( \epsilon_G \) is a continuous random variable with mean zero and finite variance, and is mutually independent of \((T, \epsilon_1, \epsilon_2)\).

Let \( D_i \) be the difference \( G - X_i \), \( i = 1, 2 \). We take \( \theta_i = E(D_i^2) \) as the measure of agreement between the \( i \)-th instrument and the gold standard. In terms of the \( \theta \)'s, the best instrument is the one that is associated with \( \min \{\theta_1, \theta_2\} \). The selection of this instrument is discussed in Chapter 4. Chapter 5 considers Problem \( \mathcal{C} \) where we first test whether \( \min \{\theta_1, \theta_2\} \) is practically equivalent to zero. Only when we have evidence in favor of this claim, we proceed to the selection of the best instrument.

Let us now introduce the datasets that we use to illustrate the various approaches introduced in this dissertation.

**Oxygen Saturation Data** (Bland and Altman, 1986): Here the percentage of oxygen saturation is measured in 72 subjects using the pulse oximeter saturation method and the oxygen saturation meter. The former is a non-invasive technique. Our interest lies in assessing agreement between the two methods so that they can be used interchangeably.

**Plasma Volume Data** (Bland and Altman, 1999): Plasma volume is the variable of interest here. It is expressed as a percentage of normal using sets of normal values.
due to Nadler and Hurley. We are concerned with evaluating the degree of agreement between the two methods. There are 99 subjects in the dataset.

**Fractional Area Change Data** (Hutson, Wilson and Geiser, 1998): Here the variable of interest is the fractional area change in endocardial surface, defined as the relative difference in endocardial areas at the diastole end and the systole end. The measurements are taken on 15 subjects using three different approaches. We discuss two issues related with these data. First, we compare the measurements by a committee of expert echocardiographers and the Autonomous Boundary Detection Algorithm. The committee of experts is treated as the gold standard and the interest lies in finding whether the computer algorithm agrees sufficiently well with the experts. Next, we compare their first expert echocardiographer and the Autonomous Boundary Detection Algorithm with the committee of expert echocardiographers, and ask whether the computer algorithm or the human expert is closer to the gold standard. The computer algorithm is an attractive alternative to the human expert as it requires no observer input.

We summarize our findings in Chapter 6 and conclude with directions for future work. Appendix A contains the lemmas that are used to prove various results. The three datasets introduced above are tabulated in Appendix B. We have collected the commonly used notations and abbreviations on page x for ready reference. We will introduce additional notations when found necessary. All the numerical computations in this dissertation have been done in FORTRAN 90, sometimes in conjunction with the IMSL Math and Stat Library.
CHAPTER 2

ASSESSING AGREEMENT BETWEEN TWO INSTRUMENTS: PARAMETRIC APPROACHES

2.1 Introduction and notation

In this chapter, we consider the problem of assessing agreement between two competing methods (or instruments) for measuring a continuous variable. The data in this problem consist of paired measurements from the instruments on a random sample of subjects. Assessing agreement refers to the evaluation of how far the measurements are from the 45° degree line through origin.

We consider a test of hypotheses approach for this problem. The hypotheses are of the general form:

\[ H : \ \text{The instruments lack satisfactory agreement, versus} \]

\[ K : \ \text{The instruments have satisfactory agreement.} \quad (2.1) \]

The advantage of this formulation is that we look for evidence in the data to claim satisfactory agreement. This way the probability of type-I error that the test would control, is actually the error of concluding satisfactory agreement when there is lack of it. The hypotheses testing formulation also enables the determination of sample sizes for planning method comparison studies.
In this chapter we focus on measuring/assessing agreement directly in terms of various parameters of the distribution of measurements. The chapter is organized as follows. Section 2.2 has a review of the literature and discusses the limitations of the current approaches. Two new approaches that address these limitations are introduced in Sections 2.3 and 2.4. Section 2.5 contains a discussion of the results.

The basic notation was introduced in Chapter 1. Let us now introduce some additional notation. Let \((X_{1j}, X_{2j})\) be the pair of measurements by the two instruments on the \(j\)-th subject in a random sample of \(n\) subjects. Thus \((X_{1j}, X_{2j})\)'s are i.i.d. copies of \((X_1, X_2)\), where \(X_i\) is the measurement on a typical subject by the \(i\)-th instrument, \(i = 1, 2\). The pair \((X_1, X_2)\) has a continuous bivariate distribution with parameters \((\xi_1, \xi_2, \eta_1, \eta_2, \nu)\), where \(\xi_i = E(X_i), \eta_i^2 = Var(X_i)\) and \(\nu = Corr(X_1, X_2)\).

Let \(D_j = X_{2j} - X_{1j}, j = 1, 2, \ldots, n\). These \(D_j\)'s are i.i.d. copies of the difference \(D = X_2 - X_1\), that has a continuous distribution with mean \(\mu = \xi_2 - \xi_1\) and variance \(\sigma^2 = \eta_1^2 + \eta_2^2 - 2\nu \eta_1 \eta_2\). Further, let \((\hat{\xi}_1, \hat{\xi}_2, \hat{\eta}_1^2, \hat{\eta}_2^2, \hat{\rho})\) and \((\hat{\mu}, \hat{\sigma}^2)\) be the usual unbiased estimators of \((\xi_1, \xi_2, \eta_1^2, \eta_2^2, \nu)\) and \((\mu, \sigma^2)\), respectively. Finally, let \(X = [(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})]\) and \(D = [D_1, D_2, \ldots, D_n]\).

We use \(N(\mu, \sigma)\) to denote a normal distribution with mean \(\mu\) and standard deviation \(\sigma\), \(Z\) for \(N(0,1)\) random variable, \(\Phi(\cdot)\) for its cdf and \(\phi(\cdot)\) for its pdf. In addition, let \(W\) and \(h_{n-1}(w)\) denote a \(\chi^2_{n-1}\) random variable and its pdf, respectively. Finally, let \(z(\alpha), t_k(\alpha)\) and \(\chi^2_k(\alpha)\) denote the upper \(\alpha\)-th quantiles of \(N(0,1)\), \(t_k\) and \(\chi^2_k\) distributions, respectively.
2.2 Literature review

2.2.1 Limits of agreement

The limits of agreement provide a simple and intuitive tool for informally assessing agreement between two instruments. It was proposed by Bland and Altman (1986). This is the “citation classic” paper mentioned in Chapter 1. The approach has two components: (a) 95\% limits of agreement, defined as \( \hat{\mu} \pm 1.96\hat{\sigma} \), and (b) the plot of mean, \( (X_1 + X_2)/2 \), versus difference, \( D \), popularly known as the Bland and Altman plot. JMP (Version 4) produces such a plot for matched pair data.

The limits of agreement essentially give the range in which most of the differences are expected to lie, so one could use the two instruments interchangeably if these limits are not too wide from a practical viewpoint. The mean versus difference plot also indicates whether \( \xi_1 \approx \xi_2 \) when this range is close to zero and \( \eta_1 \approx \eta_2 \) when the mean and the difference appear uncorrelated (see also Bland and Altman, 1995). A formal test of equality of variances based on mean and difference goes back to Pitman (1939).

A strength of Bland and Altman’s approach is that the level of agreement between instruments, as assessed by the limits of agreement, is not affected by the between-subject variation in the data. This approach was recently generalized by Bland and Altman (1999) to accommodate repeated measurements on each subject from every instrument.

Bland and Altman (1999) claim that if the differences are normally distributed, we would expect 95\% of the differences to lie between \( \hat{\mu} + 1.96\hat{\sigma} \) and \( \hat{\mu} - 1.96\hat{\sigma} \). This is roughly true only when we have a large sample even under the normal distribution.
assumption for $D$. Let $C(\hat{\mu}, \hat{\sigma}, k)$ be the coverage of the interval $(\hat{\mu} - k\hat{\sigma}, \hat{\mu} + k\hat{\sigma})$, i.e.,

$$C(\hat{\mu}, \hat{\sigma}, k) = \int_{\hat{\mu} - k\hat{\sigma}}^{\hat{\mu} + k\hat{\sigma}} f(t; \mu, \sigma)dt,$$

(2.2)

where $k > 0$ and $f(\cdot)$ is a normal pdf. If $k$ is such that $E(C(\hat{\mu}, \hat{\sigma}, k)) = p_n$, then the interval $(\hat{\mu} - k\hat{\sigma}, \hat{\mu} + k\hat{\sigma})$ is called a $p_n$-expectation tolerance interval. This interval can also be interpreted as a level-$p_n$ prediction interval for a future observation from a $N(\mu, \sigma)$ distribution (see Guttman, 1988). For normally distributed $D$, $p_n$ can be expressed as

$$p_n = Pr\left(-k/\sqrt{(n+1)/n} \leq T \leq k/\sqrt{(n+1)/n}\right),$$

(2.3)

where $T$ follows a $t_{n-1}$-distribution.

Thus by using $k = 1.96$ in (2.2), we can formally interpret the interval within 95% limits of agreement as a $p_n$-expectation tolerance interval where $p_n$ is computed by (2.3) with $k = 1.96$. It can be seen that $p_n$ is a strictly increasing function of $n$ and $\lim_{n \to \infty} p_n = 0.95$, so the expected coverage of $(\hat{\mu} - 1.96\hat{\sigma}, \hat{\mu} + 1.96\hat{\sigma})$ is always less than 0.95 in finite samples. Figure 2.1 has a plot of $n$ versus $p_n$ for $k = 1.96$. We note that this expectation is less than 0.94 for $n \leq 40$.

The protection provided by this approach in terms of expected coverage may not be enough for actually making a decision regarding satisfactory agreement between instruments. A hypotheses test type of formulation as in (2.1) is more appropriate in this respect.

2.2.2 Intraclass correlation

Chapter 1 of Fleiss (1986) has an overview of the intraclass correlation coefficient (ICC) as a measure of agreement. An ICC is defined under an ANOVA model. For
Figure 2.1: Sample size versus expected coverage of the interval within 95% limits of agreement, assuming $D$ follows $N(\mu, \sigma)$ distribution.
method comparison studies two such models may be appropriate. One model, say A, may be used for assessing agreement of an instrument with itself under identical experimental conditions, i.e., reproducibility. The other model, say B, can be used for summarizing agreement between \( k \) instruments. Let us first look at the Model A. **Model A:** Consider the following one-way random effects model for the measurements of an instrument:

\[
Y_{ij} = T_i + \epsilon_{ij}; \quad 1 \leq j \leq k_i, \ 1 \leq i \leq n. \quad (2.4)
\]

Here for the \( i \)-th subject, \( Y_{ij} = j \)-th replicate measurement by the instrument; \( T_i = \) the true unobservable measurement; and \( \epsilon_{ij} = \) measurement error. We assume: (a) \( T_i \)'s are i.i.d. \( N(\xi_T, \eta_T) \), (b) \( \epsilon_{ij} \)'s are i.i.d. \( N(0, \eta_E) \), (c) \( T_i \) and \( \epsilon_{ij} \) are mutually independent, and (d) the true measurement does not change during replication. The ICC for this model is

\[
\nu_A = \frac{\eta_T^2}{\eta_T^2 + \eta_E^2}.
\]

It represents the proportion of total variation attributed to subjects, and is also known as the reliability coefficient of the instrument. Note that \( \nu_A \in (0, 1) \) and the higher its value is, the more reproducible the instrument would be. For a fixed \( \eta_E^2 \), \( \nu_A \) increases with \( \eta_T^2 \), where \( \eta_T^2 \) is actually the between-subject variation. We have \( \nu_A = 1 \) iff \( \eta_E^2 = 0 \), indicating perfect reproducibility/reliability.

The ANOVA table for the above model is given in Fleiss (1986, p. 10). This approach summarizes the level of agreement in a single index. The lower confidence bound for ICC is also available, so one can perform tests of hypotheses such as (2.1), and do power and sample size computation for planning reproducibility studies. However, a major drawback of this index is that it is highly sensitive to the
between-subject variation. Even when the “real” agreement of an instrument with
itself as measured by $\eta_E^2$ remains fixed, the instrument can be made to appear highly
reproducible just by considering a very heterogeneous population of subjects. Bland
and Altman (1990), Atkinson and Nevill (1997), and Quan and Shih (1996) present
datasets to illustrate this phenomena. As a consequence the ICC computed from
different populations may not be comparable.

For this reason, Quan and Shih (1996) suggest using the within-subject coeffi-
cient of variation to assess reproducibility. The smaller this variation is, the more
reproducible the instrument would be. But, this coefficient is sensitive to a shift in
measurement. So, an instrument can be made to appear more reproducible by adding
a constant to all the measurements. Quan and Shih (1996) conclude by saying that
there is no absolute choice between these two indices. ICC may be preferred when
the population of subjects is well-identified. However, the coefficient of variation may
be preferable when the unit of measurement has absolute meaning and is not subject
to any shift.

St. Laurent (1998) considers a model similar to (2.4) and suggests the associated
ICC for assessing agreement of an instrument with a gold standard. His model is:

$$Y_i = G_i + \epsilon_i; \quad 1 \leq i \leq n,$$

(2.5)

where $Y_i =$ measurement on $i$-th subject by the instrument, $G_i =$ measurement on
$i$-th subject by the gold standard, and $\epsilon_i =$ measurement error in the instrument.

This model postulates that the true regression line of $Y_i$ on $G_i$ is the $45^\circ$ degree line
through the origin. Let $G_i$’s ($\epsilon_i$’s) be i.i.d. with mean $\xi_G$ (0) and variance $\eta_G^2$ ($\eta_E^2$). It
is also assumed that $G_i$ and $\epsilon_i$ are independently distributed. Although, the model
(2.5) resembles the classical test theory model, e.g. (2.4), see Lord and Novick, (1968), but here the $G_i$'s are observable.

The intraclass correlation associated with the model (2.5) is

$$
\nu_G = \frac{\eta_G^2}{\eta_G^2 + \eta_E^2},
$$

which is also the squared correlation between $Y_i$ and $G_i$. Clearly, $0 < \nu_G < 1$ and it equals one iff $\eta_E^2 = 0$. St. Laurent (1998) uses $\nu_G$ as an index of agreement between the two instruments. He suggests the ML estimator of $\nu_G$ for inference when $G_i$ and $\epsilon_i$ are normally distributed. He also develops a large sample theory for the non-normal case. However, being an ICC, this coefficient also suffers from the drawback of being sensitive to the between-subject variation. Harris, Burch and St. Laurent (2001) have developed a family of estimators of $\nu_G$ that includes the estimator proposed by St. Laurent (1998). They indicate that at times some other member of the family may be preferable over St. Laurent’s estimator in terms of bias and mean-squared error.

We now consider the second model.

**Model B:** Let $Y_{ij}$ be the measurement on the $i$-th subject by the $j$-th instrument. The following two-way mixed effects model is used for $Y_{ij}$.

$$
Y_{ij} = T_i + b_j + \epsilon_{ij}; \quad 1 \leq i \leq n, \quad 1 \leq j \leq k \tag{2.6}
$$

where $b_j$ is the fixed bias of the $j$-th instrument. The rest of the notation and assumptions are the same as those in model A (2.4). It is additionally assumed that $\sum_{j=1}^{k} b_j = 0$. The ICC from this model may be used for summarizing agreement among $k$ instruments with a-priori assumption of equal error variance for all the $k$
instruments. It is defined as:

\[ \nu_B = \frac{\eta^2_2}{\eta^2_T + \theta^2_\mu + \eta^2_E}, \]  

where \( \theta^2_\mu = \sum_{j=1}^{k} b^2_j / (k-1) \). This ICC is also known as the inter-instrument reliability. Note that \( \nu_B \) is not a proportion of total variation. But still \( 0 < \nu_B < 1 \). It equals one iff \( \eta^2_E = 0 \) and \( b_j = 0 \ \forall j \), indicating perfect agreement among all the \( k \) instruments.

The ANOVA table for this model can be found in Fleiss (1986, p. 19). This approach also summarizes the level of agreement in a single index and a standard theory is available for inference purposes. This model can accommodate \( k > 2 \) instruments, but, the assumption of equal error variance for all the instruments is generally not reasonable. Further, \( \nu_B \) is non-negative but its estimates may be negative, and it is also sensitive to the between-subject variation. If \( \nu_B \) is low, it is unclear whether the lack of agreement is due to low between-subject variation and/or high error variation and/or location bias. See also Müller and Büttner (1994) for a critique.

### 2.2.3 Concordance correlation and related measures

Lin (1989) proposed concordance correlation coefficient (CCC) as an index for measuring agreement between two instruments. He developed the CCC, defined as,

\[ \nu_L = \frac{2\nu \eta_1 \eta_2}{\eta^2_1 + \eta^2_2 + (\xi_1 - \xi_2)^2}, \]  

by scaling the expected squared distance of a point \((X_1, X_2)\) from the 45° line through the origin, to lie between \([-1, 1]\]. Thus \( \nu_L \) measures how close the (paired) observations are to the 45° line. There is perfect agreement if \( \nu_L = 1 \), no agreement if it is zero, and perfect negative agreement if it equals \(-1\). It can also be written as \( \nu C_b \), where \( C_b = 2/(v + 1/v + u^2) \), \( v = \eta_1/\eta_2 \), and \( u = (\xi_1 - \xi_2)/\sqrt{\eta_1 \eta_2} \). So, \( \nu_L \)
has two components: (a) \( \nu \), the “precision” component that measures how close the observations are to the best fit line, and (b) \( C_b \in (0, 1] \), the “accuracy” component that measures how close the best fit line is to the 45° line.

It may be noted that \( |\nu_L| \leq |\nu| \leq 1; \nu_L = 0 \) iff \( \nu = 0 \); and \( \nu_L = \pm 1 \) iff \( \nu = \pm 1 \), \( \eta_1 = \eta_2 \), and \( \xi_1 = \xi_2 \). The estimate of \( \nu_L \) is:

\[
\hat{\nu}_L = \frac{2\hat{\mu} \hat{\eta}_1 \hat{\eta}_2}{\hat{\eta}_1^2 + \hat{\eta}_2^2 + (\xi_1 - \xi_2)^2}.
\] (2.9)

Under bivariate normality for \((X_1, X_2)\), Lin (1989) showed that \( \hat{\nu}_L \) is asymptotically normal with mean \( \nu_L \). He suggests using the transformation \( \tanh^{-1}(\hat{\nu}_L) \) for faster convergence. Another advantage of this transformation is that the asymptotic confidence interval for \( \nu_L \) is constrained to lie within \([-1, 1]\). Lin (1992) discussed the hypotheses of the form (2.1) and gave a sample size formula based on allowable losses in “precision” and “accuracy”. Chinchilli et al. (1996) extended \( \nu_L \) to handle repeated measurements data. Further, King and Chinchilli (2001) generalized \( \nu_L \) to incorporate distance functions, other than the squared error, to construct more robust forms of \( \nu_L \). They also demonstrated the relationship between \( \nu_L \) and the kappa statistic (a measure of agreement for nominal/ordinal measurements) and introduced additional extensions. Additionally, Barnhart and Williamson (2001) suggested a generalized estimating equations approach to model \( \nu_L \).

Nickerson (1997) explained that \( \hat{\nu}_B \), the estimate of the ICC \( \nu_B \) defined in (2.7), and \( \hat{\nu}_L \) in (2.9) tend to be similar. When \( k = 2 \), \( \hat{\nu}_B \) becomes

\[
\hat{\nu}_B = \frac{2\hat{\mu} \hat{\eta}_1 \hat{\eta}_2}{\hat{\eta}_1^2 + \hat{\eta}_2^2 + (\xi_1 - \xi_2)^2 + \hat{\sigma}^2/n}.
\]

Thus \( \hat{\nu}_L \) and \( \hat{\nu}_B \) are identical except for the \( \hat{\sigma}^2/n \) term in the denominator. She cites several published datasets and argues that \( \hat{\sigma}^2 \) tends to be small in method comparison
studies, so even with fairly small $n$ the difference between $\hat{\nu}_L$ and $\hat{\nu}_B$ is likely to be quite small.

CCC and its various forms are also sensitive to the between-subject variation like the ICC’s. Recently Lin et al. (2002) compared $\nu_L$ with competing approaches based on coverage probability, and concluded that the latter ones have better power for inference than $\nu_L$.

Liao and Lewis (2000) allude to some downsides of CCC and suggest a similar index. They extend their index to handle situations when parameters cannot be assumed to remain fixed over the entire range for measurement. For example, in certain chemical applications, the variation is smaller at low concentrations but is larger at higher concentrations. For these situations, Liao and Lewis suggest a curve, where in a small local window, the observations are treated to have a distribution with fixed parameters. They choose Gaussian kernel methods to assign weights and select bandwidth that minimize the sum of the confidence interval length at all observed points.

**Remark 2.1.** The total deviation index of Lin (2000) and the “coverage probability” approach of Lin et al. (2002) are two other methods for assessing agreement. They are based on the coverage probability of a given interval, and are described in detail in Chapter 3.

To summarize, the ICC, CCC, and their various forms quantify agreement in a single index. The major weakness of these indices is their sensitivity to the between-subject variation. So, if the lack of satisfactory agreement is concluded, it may be unclear whether it is due to low between-subject variation or location/scale bias (see Atkinson and Nevill, 1997, and Lin and Chinchilli, 1997). Further, these indices
combine the differences in means and variabilities, and the correlation. Thus if a lack of satisfactory agreement is concluded, they would not help in suggesting corrective actions without additional statistical investigation. They may also fail to detect location/scale bias if the correlation $\nu$ is close to one, and hence may lead to wrong conclusions. We demonstrate this scenario in Section 2.3 with the plasma volume data from Bland and Altman (1999). Another drawback of the existing approaches is that none of them uses the exact distribution theory. They are not applicable if normality assumption for differences of the (paired) data does not hold or if a large sample is not available.

In Sections 2.3 and 2.4 we introduce two approaches for testing the hypotheses of satisfactory agreement, defined in (2.1), that address various limitations described above. The model introduced in (1.1) provides the setup for the measurements in the rest of this chapter and the next one.

### 2.3 Test based on the differences

In this section we define satisfactory agreement in terms of parameters of the differences, $(\mu, \sigma)$. Suppose that the user can specify two positive thresholds, $\delta_\mu$ and $\delta_\sigma$, such that the two instruments will be considered to have satisfactory agreement if $\mu$ is within $\pm \delta_\mu$ and $\sigma$ is less than $\delta_\sigma$. The thresholds $\delta_\mu$ and $\delta_\sigma$ are like thresholds for practical equivalence of $\mu$ and $\sigma$ with zero. The hypotheses (2.1) then become

$$H_1 : \{ |\mu| \geq \delta_\mu \} \cup \{ \sigma \geq \delta_\sigma \}, \text{ versus } K_1 : \{ |\mu| < \delta_\mu \} \cap \{ \sigma < \delta_\sigma \}.$$  \hspace{1cm} (2.10)

Here the rejection of $H_1$ would mean evidence in support of satisfactory agreement. The region under $K_1$ is an open rectangle in the $(\mu, \sigma)$-plane with co-ordinates $(-\delta_\mu, 0), (\delta_\mu, 0), (\delta_\mu, \delta_\sigma)$ and $(-\delta_\mu, \delta_\sigma)$.
Let us now assume that $D$ has a $N(\mu, \sigma)$ distribution. We can test (2.10) by separately testing hypotheses on $\mu$ and $\sigma$ components and combining the results using the intersection-union principle of Berger (1982). Here the component hypotheses are:

$$H_\mu : |\mu| \geq \delta_\mu, \text{ versus } K_\mu : |\mu| < \delta_\mu, \quad (2.11)$$

and

$$H_\sigma : \sigma \geq \delta_\sigma, \text{ versus } K_\sigma : \sigma < \delta_\sigma. \quad (2.12)$$

The critical regions

$$C_\mu = \left\{ D : (\hat{\mu} - n^{-1/2} t_{n-1}(\alpha) \hat{\sigma}, \hat{\mu} + n^{-1/2} t_{n-1}(\alpha) \hat{\sigma}) \subset (-\delta_\mu, \delta_\mu) \right\} \quad (2.13)$$

and

$$C_\sigma = \left\{ D : (n - 1) \hat{\sigma}^2 < \delta_\sigma^2 \chi^2_{n-1}(1 - \alpha) \right\} \quad (2.14)$$

have size $\alpha$ for testing (2.11) (see Berger and Hsu, 1996) and (2.12), respectively. Hence the level $\alpha$ intersection-union test of (2.10) has the critical region,

$$C_1 = C_\mu \cap C_\sigma. \quad (2.15)$$

Using Theorem 2 in Berger and Hsu (1996) it is easy to see that $C_1$ has actually size $\alpha$ for testing (2.10). The $100(1 - \alpha)\%$ confidence interval for $\mu$ associated with the hypotheses (2.11) is

$$\left[ \min \left\{ \hat{\mu} - t_{n-1}(\alpha)n^{-1/2}\hat{\sigma}, 0 \right\}, \max \left\{ \hat{\mu} + t_{n-1}(\alpha)n^{-1/2}\hat{\sigma}, 0 \right\} \right], \quad (2.16)$$

and, for $\sigma$ associated with the hypotheses (2.12), it is

$$\left( 0, (n - 1)^{1/2}\hat{\sigma}/\sqrt{\chi^2_{n-1}(1 - \alpha)} \right]. \quad (2.17)$$

Berger and Hsu (1996) derived the former confidence interval in the context of average bioequivalence. Note that the interval is constrained to contain zero. But, this is not a problem as the inference whether $\mu = 0$ is not of interest here. One may also assume
that $\mu \neq 0$, which would be reasonable from practical point of view as it says that the two instruments do not have exactly the same means. The confidence interval for $\sigma$ is easy to derive.

**Proposition 2.1.** The critical region $C_1$ for testing (2.10) has the power function

$$p_1(\alpha, \delta_\mu, \delta_\sigma, n, \mu, \sigma) = \int_0^a \left( \Phi \left( n^{1/2} \sigma^{-1} (\delta_\mu - \mu) - t_{n-1}(\alpha) (n - 1)^{-1/2} w^{1/2} \right) - \Phi \left( - n^{1/2} \sigma^{-1} (\delta_\mu + \mu) + t_{n-1}(\alpha) (n - 1)^{-1/2} w^{1/2} \right) \right) \times h_{n-1}(w) \, dw,$$

where $a = \min \left\{ \chi^2_{n-1}(1 - \alpha) \delta^2_\sigma \sigma^{-2}, (t_{n-1}(\alpha) \sigma)^{-2} n (n - 1) \delta^2_\mu \right\}$, and $h_{n-1}(\cdot)$ is the density function of $\chi^2_{n-1}$-distribution.

**Proof:** From the definition of power function,

$$p_1(\alpha, \delta_\mu, \delta_\sigma, n, \mu, \sigma) = Pr(D \in C_1)$$

$$= Pr \left( \left( \hat{\mu} - n^{-1/2} t_{n-1}(\alpha) \hat{\sigma}, \hat{\mu} + n^{-1/2} t_{n-1}(\alpha) \hat{\sigma} \right) \subset \left( - \delta_\mu, \delta_\mu \right) \right.$$ 

$$(n - 1) \hat{\sigma}^2 < \delta^2_\sigma \chi^2_{n-1}(1 - \alpha) \right)$$

$$= Pr \left( - n^{1/2} \sigma^{-1} (\delta_\mu + \mu) + t_{n-1}(\alpha) (n - 1)^{-1/2} W^{1/2} < Z < n^{1/2} \sigma^{-1} (\delta_\mu - \mu) - t_{n-1}(\alpha) (n - 1)^{-1/2} W^{1/2}, \right.$$ 

$$W < \chi^2_{n-1}(1 - \alpha) \delta^2_\sigma \sigma^{-2})$$

where $Z = n^{1/2}(\hat{\mu} - \mu)/\sigma$ and $W = (n-1) \hat{\sigma}^2 / \sigma^2$. Here $Z$ follows a $N(0, 1)$ distribution, $W$ follows a $\chi^2_{n-1}$-distribution and they are independent. The result is now established by conditioning on $W$. \hfill \square

The power function depends on $(\delta_\mu, \delta_\sigma, \mu, \sigma)$ only through $(\delta_\mu / \sigma, \delta_\sigma / \sigma, \mu / \sigma)$, and is symmetric in $\mu$ about zero. Figure 2.2 presents the contours $p_1(0.05, 1.0, 1.0, n, \mu, \sigma) = 19$
Figure 2.2: Contours of the power function $p_1(0.05, 1.0, 1.0, n, \mu, \sigma)$ for $n = 15, 30$.

0.05, 0.90 in $(\mu, \sigma)$-plane for $n = 15, 30$. Due to the symmetry of $p_1$ we only consider the region with $\mu \geq 0$. We see that the type-I error rate of 0.05 is achieved on the boundary that divides the null and alternative regions. This verifies that the test has size 0.05. The surface plot of $p_1(0.05, 1.0, 1.0, 15, \mu, \sigma)$ is given in Figure 2.3. As expected, the power function increases as $\sigma$ and $|\mu|$ decrease. It is close to one when $\sigma$ is close to zero. The expression for the power function is also useful in determining sample sizes for planning method comparison studies.

**Example 2.1.** We now illustrate this procedure with the fractional area change data from Hutson et al. (1998). The data were introduced in Chapter 1. We denote their committee of expert echocardiographers by $X_1$ and the Autonomous Boundary
Figure 2.3: Surface plot of the power function $p_1(0.05, 1.0, 1.0, 15, \mu, \sigma)$. 
Detection Algorithm by $X_2$. The interest lies in evaluating the agreement between the computer algorithm and the experts. Figure 2.4, also given in Hutson et al. (1998), has the scatter plot of the 15 pairs of measurements in these data. The agreement between the two instruments does not appear to be strong. Their correlation is fairly weak and it seems that $\xi_1 > \xi_2$. The parameter estimates are: $(\hat{\xi}_1, \hat{\xi}_2, \hat{\eta}_1, \hat{\eta}_2, \hat{\nu}) = (37.50, 32.83, 3.66, 5.83, 0.13)$ and $(\hat{\mu}, \hat{\sigma}) = (-4.68, 6.47)$.

Figure 2.5 has the normal probability plot for the differences. Apparently the normality assumption for $D$ looks fine. Further, the Shapiro-Wilk test of normality gives a $p$-value of 0.70. Thus, it may be reasonable to assume the differences to
Figure 2.5: Normal probability plot for the differences, $D = X_2 - X_1$, in Example 2.1.

have come from a normal distribution. Using (2.16) and (2.17) the 95\% confidence intervals for $\mu$ and $\sigma$ are, respectively, $[-7.62, 0]$ and $(0, 9.45]$.

Thus for given $\delta_\mu$ and $\delta_\sigma$, we accept a component null hypothesis iff the associated confidence interval intersects the corresponding null hypothesis region. We can then combine the results to get a level 0.05 test of (2.10). Here the data would support the claim of satisfactory agreement between the two instruments only when $\delta_\mu > 7.62$ units and $\delta_\sigma > 9.45$ units. In this sense the confidence intervals indicate the nature and extent of disagreement. These thresholds appear to be too high taking into consideration the magnitude of measurements. So, overall the data does not support the claim of satisfactory agreement.
**Example 2.2.** These data, taken from Bland and Altman (1999), were also introduced in Chapter 1. Here the plasma volume, expressed as a percentage of normal, is measured in 99 subjects using two sets of normal values: Nadler’s ($X_1$) and Hurley’s ($X_2$). Figure 2.6 has their scatter plot, which is also given in Bland and Altman (1999). The two methods appear to be highly correlated. But the measurements from the Nadler method are always higher than those from the Hurley method. Further, it seems that their differences increase as the measurements increase, with about 9.5 units in the middle part of the data. The parameter estimates are: $(\hat{\xi}_1, \hat{\xi}_2, \hat{\eta}_1, \hat{\eta}_2, \hat{\nu}) = (98.50, 89.24, 15.18, 13.89, 0.99)$ and $(\hat{\mu}, \hat{\sigma}) = (-9.26, 2.40)$.

The histogram and the normal probability plot for the differences are given in Figure 2.7. It indicates that their distribution is more skewed than the normal distribution. But the assumption of normality does not look too unreasonable. A $p$-value of 0.72 from the Shapiro-Wilk test of normality corroborates this claim. The 95% confidence intervals for $\mu$ and $\sigma$ are, respectively, $[-9.66, 0]$ and $(0, 2.73]$. In this case we would have evidence for satisfactory agreement between the two instruments if $\delta_\mu > 9.66$ and $\delta_\sigma > 2.73$. Such a threshold for $\delta_\sigma$ seems realistic but for $\delta_\mu$ it does not. But, recall that the two means seemed to differ roughly by 9.5 units. So, if we subtract 9.5 from all the measurements by the Nadler method, the 95% confidence interval for $\mu$ becomes $[-0.16, 0.64]$. Thus, after this transformation, we would infer the two instruments to have satisfactory agreement.

It may be noted that for this data the estimate of concordance correlation, given by (2.9), is 0.82 and its lower confidence bound is 0.81. These values are high enough to conclude “good” agreement as a cutoff of 0.75 is sometimes suggested (see Atkinson
Figure 2.6: Scatter plot for the plasma volume data (Example 2.2).
Figure 2.7: Histogram and the normal probability plot for the differences, $D = X_2 - X_1$, in Example 2.2.
and Nevill, 1997). But, this conclusion may be misleading if the bias between the two instruments is taken into account.

It is clear that the confidence intervals like (2.16) and (2.17) that are built into the decision making process indicate the nature and extent of disagreement. This information is helpful, as was the case in Example 2.2 above, in suggesting what corrective measures to take if a lack of satisfactory agreement is inferred. It is an important advantage over the existing approaches that summarize agreement in a single index.

**EXAMPLES 2.1 and 2.2 (continued):** Figure 2.8 has the plots of means versus differences for the two data discussed above. The positive association in the case of fractional area change data suggests that perhaps the variance of the computer algorithm ($\eta_2$) is greater than the variance of the gold standard ($\eta_1$). Further, the negative association in the case of plasma volume data indicates that perhaps variance of Nadler method ($\eta_1$) is greater than the variance of Hurley method ($\eta_2$). But, recall that we adjudged the variance of $D$ to be practically close to zero. It may be noted that when the correlation is very high, as is the case in this example, a difference in the two variances may fail to be evident in the variance of $D$.

If a threshold for the ratio of the two variances can be specified, we can ask whether they are practically equivalent. We also ask whether the correlation is high enough for agreement. These questions can be handled through the tools developed in the next section where we work with the bivariate data.
Figure 2.8: Plots of means versus differences for the data in Example 2.1 (top) and Example 2.2 (bottom).
2.4 Test based on the bivariate data

We now assume that \((X_{1j}, X_{2j})\), \(1 \leq j \leq n\), is a random sample from a bivariate normal distribution with parameters \((\xi_1, \xi_2, \eta_1, \eta_2, \nu)\). For perfect agreement we must have \(\xi_1 = \xi_2, \eta_1 = \eta_2\) and \(\nu = 1\). Thus we can assess satisfactory agreement by simultaneously testing if \(|\xi_2 - \xi_1| < \delta_\mu, \delta_1 < \eta_2/\eta_1 < \delta_2\) and \(\nu > \delta_\nu\), where the thresholds are specified by the experimenter. Here \(\delta_\nu \in (0,1)\) and \(0 < \delta_1 < 1 < \delta_2\). The choices for \(\delta_1, \delta_2\) depend on whether (i) one of the instruments, say the first, is a reference instrument, or (ii) there is no reference instrument. In the former case \(\delta_1 = 1 - \delta(\eta_1, \eta_2)\) and \(\delta_2 = 1 + \delta(\eta_1, \eta_2)\), where \(\delta(\eta_1, \eta_2)\) is specified such that \(|(\eta_2/\eta_1) - 1| < \delta(\eta_1, \eta_2)\). In the latter case \(\delta_1 = (1 + \delta(\eta_1, \eta_2))^{-1}\) and \(\delta_2 = 1 + \delta(\eta_1, \eta_2)\), where \(|\ln(\eta_2/\eta_1)| < \ln(1 + \delta(\eta_1, \eta_2))\). The hypotheses of satisfactory agreement then becomes:

\[
H_2: \{\nu \leq \delta_\nu\} \cup \{|\xi_2 - \xi_1| \geq \delta_\mu\} \cup \{\eta_2/\eta_1 \leq \delta_1 \text{ or } \eta_2/\eta_1 \geq \delta_2\}, \text{ versus } \\
K_2: \{\nu > \delta_\nu\} \cap \{|\xi_2 - \xi_1| < \delta_\mu\} \cap \{\delta_1 < \eta_2/\eta_1 < \delta_2\}. \quad (2.18)
\]

We now construct the intersection-union test for these hypotheses. The component hypotheses are:

\[
H_\mu: |\xi_2 - \xi_1| \geq \delta_\mu, \text{ versus } K_\mu: |\xi_2 - \xi_1| < \delta_\mu, \quad (2.19)
\]

\[
H_\nu: \nu \leq \delta_\nu, \text{ versus } K_\nu: \nu > \delta_\nu, \text{ and } \quad (2.20)
\]

\[
H_{\eta_1, \eta_2}: \eta_2/\eta_1 \leq \delta_1 \text{ or } \eta_2/\eta_1 \geq \delta_2, \text{ versus } K_{\eta_1, \eta_2}: \delta_1 < \eta_2/\eta_1 < \delta_2. \quad (2.21)
\]

The test of (2.19) has been discussed in Section 2.3 with \(\mu = \xi_2 - \xi_1\). For testing (2.20), the level \(\alpha\) critical region has the form

\[
C_\nu = \{X:\hat{\nu} > c(\alpha, \delta_\nu, n)\}, \quad (2.22)
\]
where \( c(\alpha, \delta_\nu, n) \) is the critical value (see also Lehmann, 1986, p. 340). Although the distribution of \( \hat{\nu} \) is hard to deal with analytically, one can obtain \( c(\alpha, \delta_\nu, n) \) and the confidence interval for \( \nu \) using numerical methods. FORTRAN programs for these computations are available at www.stat.ohio-state.edu/~hnn/pankaj_programs.html or at www.stat.ohio-state.edu/~pankaj. Alternatively one could use Fisher’s Z-transformation for \( \hat{\nu} \) to approximate these quantities. This transformation works reasonably well for \( \delta_\nu \geq 0.9 \) even with \( n = 5 \). Our extensive numerical computations showed that for \( \alpha = 0.05 \), the maximum relative error of the approximation of \( c(\alpha, \delta_\nu, n) \) never exceeded 0.25%, where the maximum was taken over \( \delta_\nu = 0.90(0.01)0.99 \) and \( n \geq 5 \). But the level of the test using this approximation was slightly more than the nominal level.

For testing the hypotheses (2.21) we consider the following transformations suggested by Lehmann (1986, p. 268). This approach is an extension of approach of Pitman (1939).

\[
U_{1j} = \delta_1 X_{1j} + X_{2j}, \quad U_{2j} = \delta_1 X_{1j} - X_{2j},
\]

\[
V_{1j} = \delta_2 X_{1j} + X_{2j}, \quad V_{2j} = \delta_2 X_{1j} - X_{2j}.
\]

Let \( \gamma_1 \) (\( \gamma_2 \)) be the correlation between \( U_{1j} \)'s (\( V_{1j} \)'s) and \( U_{2j} \)'s (\( V_{2j} \)'s). Then, for \( i = 1, 2 \),

\[
\gamma_i = (\delta_i^2 - \eta_1^{-2}\eta_2^{-2}) \left( \eta_1^{-2}\eta_2^2 + \delta_i^2 + 2\delta_i\nu\eta_1^{-1}\eta_2 \right)^{-1/2} \left( \eta_1^{-2}\eta_2^2 + \delta_i^2 - 2\delta_i\nu\eta_1^{-1}\eta_2 \right)^{-1/2}.
\]

Thus the hypotheses (2.21) on \( \eta_2/\eta_1 \) can be reformulated as:

\[
H_{\eta_1,\eta_2} : \{ \gamma_1 \geq 0 \} \cup \{ \gamma_2 \leq 0 \}, \quad \text{versus } K_{\eta_1,\eta_2} : \{ \gamma_1 < 0 \} \cap \{ \gamma_2 > 0 \}.
\]

Let \( \hat{\gamma}_i \) be the sample counterpart of \( \gamma_i \) given by

\[
\hat{\gamma}_i = (\hat{\delta}_i^2 - \hat{\eta}_1^{-2}\hat{\eta}_2^{-2}) \left( \hat{\eta}_1^{-2}\hat{\eta}_2^2 + \hat{\delta}_i^2 + 2\hat{\delta}_i\hat{\nu}\hat{\eta}_1^{-1}\hat{\eta}_2 \right)^{-1/2} \left( \hat{\eta}_1^{-2}\hat{\eta}_2^2 + \hat{\delta}_i^2 - 2\hat{\delta}_i\hat{\nu}\hat{\eta}_1^{-1}\hat{\eta}_2 \right)^{-1/2}.
\]
Then \( C_{\eta_1, \eta_2} = C_{1; \eta_1, \eta_2} \cap C_{2; \eta_1, \eta_2} \), where
\[
C_{1; \eta_1, \eta_2} = \{ \mathbf{X} : (n - 2)^{1/2} \hat{\gamma}_1 < -t_{n-2}(\alpha) \left( 1 - \hat{\gamma}_1^2 \right)^{1/2} \}, \quad (2.26)
\]
\[
C_{2; \eta_1, \eta_2} = \{ \mathbf{X} : (n - 2)^{1/2} \hat{\gamma}_2 > t_{n-2}(\alpha) \left( 1 - \hat{\gamma}_2^2 \right)^{1/2} \}, \quad (2.27)
\]
is a size \( \alpha \) critical region for testing (2.24). The following theorem gives the confidence interval for \( \eta_2/\eta_1 \).

**Proposition 2.2.** The \( 100(1 - \alpha)\% \) confidence interval for \( \eta_2/\eta_1 \) corresponding to the hypotheses in (2.21) is: \( [\min(\Delta_-, 1), \max(\Delta_+, 1)] \), where
\[
\Delta_- = \frac{\hat{\gamma}_2}{\hat{\eta}_1} \frac{\sqrt{1 - t_1^2 \hat{\nu}^2} - t_1 \sqrt{1 - \hat{\nu}^2}}{\sqrt{1 - t_1^2}}, \quad \Delta_+ = \frac{\hat{\gamma}_2}{\hat{\eta}_1} \frac{\sqrt{1 - t_1^2 \hat{\nu}^2} + t_1 \sqrt{1 - \hat{\nu}^2}}{\sqrt{1 - t_1^2}},
\]
and \( t_1 = (n - 2 + t_{n-2}(\alpha))^2 t_{n-2}(\alpha) \).

**Proof:** We obtain the lower and upper confidence bounds for \( \eta_2/\eta_1 \) by inverting acceptance regions for the component hypotheses in (2.24) and combine them using Theorem 2.5 in Hsu et al. (1994) to get the result. Since \( \hat{\gamma}_i / \left( 1 - \hat{\gamma}_i^2 \right)^{1/2} \) is a monotonic function of \( \hat{\gamma}_i \), the critical region \( C_{\eta_1, \eta_2} \) is equivalent to the region \( \{ \mathbf{X} : \hat{\gamma}_1 < -t_1, \hat{\gamma}_2 > t_1 \} \), where the \( \hat{\gamma}_i \)'s are given in (2.25).

We can write \( \{ \hat{\gamma}_1 \geq -t_1 \} \) as
\[
\{ \hat{\gamma}_1 \geq -t_1 \} = \{ |\hat{\gamma}_1| \leq t_1 \} \cup \{ \hat{\gamma}_1 \geq t_1 \}. \quad (2.28)
\]

Further (2.25) implies that \( |\hat{\gamma}_1| \leq t_1 \) is equivalent to:
\[
(\hat{\delta}_1^2 - \hat{\eta}_1^{-2} \hat{\gamma}_2^2) - t_1^2 \left( (\hat{\eta}_1^{-2} \hat{\gamma}_2^2 + \delta_1^2) - 4 \hat{\nu}^2 \delta_1^2 \hat{\eta}_1^{-2} \hat{\gamma}_2^2 \right) \leq 0. \quad (2.29)
\]
It follows from (2.29) that \( |\hat{\gamma}_1| \leq t_1 \iff \Delta_- \leq \delta_1^2 \leq \Delta_+ \). Furthermore, \( \hat{\gamma}_1 \geq t_1 \iff \Delta_+ \leq \delta_1^2 \). Hence, on using (2.28) we conclude that: \( \{ \hat{\gamma}_1 \geq -t_1 \} \iff \{ \Delta_+ \leq \delta_1^2 \} \), meaning that \( 100(1 - \alpha)\% \) lower confidence bound on \( \eta_2/\eta_1 \) is \( \Delta_- \).
Similarly, starting with \( \{ \hat{\gamma}_2 \leq t_1 \} \) we conclude that the 100(1 - \alpha)\% upper confidence bound on \( \eta_2/\eta_1 \) is \( \Delta_+ \). Now the desired result follows by using \( c = 1 \) in Theorem 2.5 in Hsu et al. (1994).

Finally consider the critical region,

\[
C_2 = C_\nu \cap C_\mu \cap C_{1; \eta_1, \eta_2} \cap C_{2; \eta_1, \eta_2},
\]

(2.30)

where \( C_\nu \), \( C_\mu \), \( C_{1; \eta_1, \eta_2} \) and \( C_{2; \eta_1, \eta_2} \) are defined in (2.22), (2.13), (2.26) and (2.27) respectively.

**Proposition 2.3.** The region \( C_2 \) has size \( \alpha \) for testing \( H_2 \) against \( K_2 \) in (2.18).

**Proof:** From the intersection-union principle it follows that the test has level \( \alpha \).

Define a sequence of parameter values \( \varphi_l = (\xi_{ll}, \xi_{2l}, \eta_{ll}, \eta_{2l}, \nu_l) \), such that \( \xi_{ll} - \xi_{2l} = \delta_\mu \), \( \eta_{ll} = \eta_{2l}, \forall l \), and \( \eta_{ll} \to 0, \nu_l \to 1 \) as \( l \to \infty \). Now as \( l \to \infty \) we have,

\[
Pr_{\varphi_l}(X \in C_\mu) = Pr_{\varphi_l}\left( (\hat{\mu} - n^{-1/2} t_{n-1}(\alpha) S, \hat{\mu} + n^{-1/2} t_{n-1}(\alpha) S) \subset (-\delta_\mu, \delta_\mu) \right) \to \alpha.
\]

Further, from (2.23), \( \gamma_\mu \to -1 \) and \( \gamma_{2l} \to 1 \), because \( \eta_{ll} = \eta_{2l}, \forall l \) and \( \nu_l \to 1 \). So on applying Lemma A.1 to the sample correlations \( \hat{\gamma}_1, \hat{\gamma}_2 \) and \( \hat{\nu} \) we get:

\[
Pr_{\varphi_l}(X \in C_{1; \eta_1, \eta_2}) = Pr_{\varphi_l}(\hat{\gamma}_1 < -t_1) \to 1,
\]

\[
Pr_{\varphi_l}(X \in C_{2; \eta_1, \eta_2}) = Pr_{\varphi_l}(\hat{\gamma}_2 > t_1) \to 1,
\]

and \( Pr_{\varphi_l}(X \in C_\nu) = Pr_{\varphi_l}(\hat{\nu} > c(\alpha, \delta_\nu, n)) \to 1, \)

since \( t_1 = (n - 2 + t_{n-2}^2(\alpha))^{-1/2} t_{n-2}(\alpha) \) and \( c(\alpha, \delta_\nu, n) \) are both less than one. The result now follows from Theorem 2 in Berger and Hsu (1996).
An analytical study of the power of the test of (2.18) using $C_2$ is intractable owing to the high number of dimensions involved. But we can use Monte-Carlo simulations for the computation of power and sample sizes.

**EXAMPLE 2.1 (continued):** The 95% confidence interval for $\xi_2 - \xi_1$, same as the one for $\mu$ using (2.16), was found in Section 2.3 to be $[-7.62, 0]$. For $\eta_2/\eta_1$, the 95% confidence interval using Theorem 2.2, is $[0.996, 2.547]$, and for $\nu$, using numerical methods, it is $[-0.32, 1)$. This way we can test (2.18) once $\delta_\mu$, $\delta_\nu$ and $\delta(\eta_1, \eta_2)$ are specified. There is evidence that the variance for the computer algorithm $(X_2)$ is higher than the committee of experts $(X_1)$, and the two are not likely to be practically equivalent. Further, the lower confidence bound of $-0.32$ for $\nu$ is too low to be sufficient for reasonable agreement. So, as in Section 2.3, the data does not have evidence for satisfactory agreement between the two methods.

**EXAMPLE 2.2 (continued):** The 95% confidence interval for $\xi_2 - \xi_1$ was found in Section 2.3 to be $[-9.66, 0]$. For $\eta_2/\eta_1$ it comes out be $[0.89, 1]$ and for $\nu$ it is $[0.99, 1)$. Thus there is some evidence that variance for the Nadler method $(X_1)$ is higher than that for the Hurley method $(X_2)$, but this difference is not likely to be practically significant. So overall, as noted in Section 2.3, after subtracting 9.5 from all the measurements from the Nadler method, we can infer the two methods to have a satisfactory agreement. A formal test of the hypotheses (2.18) can also be conducted using these confidence intervals by specifying $\delta_\mu$, $\delta_\nu$ and $\delta(\eta_1, \eta_2)$.
Remark 2.2. In case one specifies thresholds on the ratio of the means \( \xi_2/\xi_1 \) where both \( \xi_1 \) and \( \xi_2 \) are positive, one can modify the hypotheses (2.18) and (2.19) accordingly and use Sasabuchi’s test (see Berger and Hsu, 1996, p. 292) for the resulting hypotheses on \( \xi_2/\xi_1 \).

2.5 Discussion

We summarize the salient features of the various approaches in Table 2.1. The limits of agreement (LOA) and the concordance correlation coefficient (CCC) approaches were described in Section 2.2. The approaches introduced in Sections 2.3 and 2.4 provide two ways of expressing satisfactory agreement. They break up the overall level of agreement into several components. So, they are also more informative in the sense of suggesting where the disagreement is coming from and indicating its extent. Further, the hypotheses (2.10) and (2.18) resemble the hypotheses commonly used for average and population bioequivalence, respectively (see Berger and Hsu, 1996 for an introduction). However, both of them here have an additional component: (2.10) has the hypotheses on \( \sigma \) and (2.18) has the hypotheses on \( \rho \).

The comprehensive approach of Section 2.4 requires one more threshold than the approach of Section 2.3. The former is affected by the between-subject variation, but is also the more informative. Prior specification of thresholds is not really a pre-condition for using these approaches. The confidence intervals can be computed without the thresholds and they indicate what values of thresholds will be needed to infer satisfactory agreement. If the researcher believes that those values are too high or low enough from practical equivalence considerations, then the appropriate inference can be readily made.
<table>
<thead>
<tr>
<th>Features</th>
<th>Approach in Section</th>
<th>2.3</th>
<th>2.4</th>
<th>CCC</th>
<th>LOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure</td>
<td>$\mu, \sigma$, $\mu, \eta_2/\eta_1, \nu$</td>
<td>$\mu, \sigma$</td>
<td>$\mu, \eta_2/\eta_1, \nu$</td>
<td>1 index</td>
<td>2 limits</td>
</tr>
<tr>
<td>data</td>
<td>$D$</td>
<td>$(X_1, X_2)$</td>
<td>$(X_1, X_2)$</td>
<td>$D$</td>
<td>$N$</td>
</tr>
<tr>
<td>distribution</td>
<td>$N$</td>
<td>BVN</td>
<td>BVN</td>
<td>$N$</td>
<td></td>
</tr>
<tr>
<td>thresholds</td>
<td>$\delta_\mu, \delta_\sigma$</td>
<td>$\delta_\mu, \delta(\eta_1, \eta_2), \delta_\nu$</td>
<td>3 allowable losses</td>
<td>1 for $D$</td>
<td></td>
</tr>
<tr>
<td>inference</td>
<td>exact</td>
<td>exact*</td>
<td>approximate</td>
<td>informal</td>
<td></td>
</tr>
<tr>
<td>disagreement extent info.</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes**</td>
<td></td>
</tr>
<tr>
<td>b/w subject var. effect</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: *The test is exact but sample size computation uses simulation. **This information can informally be gathered from the plots.

Table 2.1: Salient features of various approaches.

We have written FORTRAN programs for determining sample size to plan a method comparison study using any of the two approaches. We now briefly illustrate this computation for the approach based on $D$ when it is normal. For specified $\delta_\mu/\sigma$, $\delta_\sigma/\sigma$, $\mu/\sigma$, $\alpha$ and the desired power $1 - \beta$, we can numerically solve the equation $p_1(\alpha, \delta_\mu, \delta_\sigma, n, \mu, \sigma) = 1 - \beta$ for $n$. The expression for $p_1$ is given by Proposition 2.1. In Table 2.2, we present the sample sizes required by a 5% level test to achieve 80% power at $(\mu, \sigma) = (0, \delta_\sigma/2)$, for various values of $\delta_\mu$ and $\delta_\sigma$. The choices of $(\delta_\mu, \delta_\sigma, \mu, \sigma)$ are arbitrary and intended solely for illustration. The FORTRAN programs for computing power and sample size are available at the website www.stat.ohio-state.edu/~hnn/pankaj_programs.html or at www.stat.ohio-state.edu/~pankaj.
<table>
<thead>
<tr>
<th>$\delta_\sigma$</th>
<th>0.50</th>
<th>0.55</th>
<th>0.60</th>
<th>0.65</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>21</td>
<td>18</td>
<td>16</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>2.0</td>
<td>36</td>
<td>30</td>
<td>26</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>55</td>
<td>46</td>
<td>39</td>
<td>34</td>
<td>29</td>
</tr>
<tr>
<td>3.0</td>
<td>79</td>
<td>66</td>
<td>55</td>
<td>47</td>
<td>41</td>
</tr>
<tr>
<td>3.5</td>
<td>107</td>
<td>89</td>
<td>75</td>
<td>64</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 2.2: Sample sizes required to achieve 80% power at $(\mu, \sigma) = (0, \delta_\sigma/2)$ with 5% level test of hypotheses (2.10) for various $(\delta_\mu, \delta_\sigma)$. 
CHAPTER 3

ASSESSING AGREEMENT
USING COVERAGE PROBABILITY

3.1 Introduction

Suppose we have two instruments that are such that a large proportion of the differences of the paired measurements is sufficiently close to zero. Then we can consider them to have satisfactory agreement. For example, as Lin et al. (2002) point out, two instruments for measuring diastolic blood pressure are usually said to have satisfactory agreement if at least 90% of their differences are within ± 10 mmHg. A similar criterion is sometimes used for assessing individual bioequivalence of two drug formulations (see e.g., Anderson and Hauck, 1990).

We use the basic notation introduced in Chapter 2. Let \( l \) and \( u \) be specified threshold limits such that a difference \( D \) lying in the interval \([l, u]\) is considered practically equivalent to zero. Here \( l < u, u > 0 \). Without any loss of generality we assume that \( l = -u \), as it amounts to transforming \( D \) to \( u(2D - u - l)/(u - l) \). Let \( \pi' \) be the coverage probability of this threshold interval, i.e., \( \pi' = Pr(l \leq D \leq u) = Pr(|D| \leq u) \). In this chapter we take this probability as a measure of agreement.
between two instruments and compare it with a specified large proportion, \( \pi_0' \) to assess the extent of agreement.

Let \( \pi_l = Pr(D < l) \), \( \pi_u = Pr(D > u) \) and \( \pi = \pi_l + \pi_u \). Further, let \( \pi'_l = 1 - \pi_l \), \( \pi'_u = 1 - \pi_u \) and \( \pi_0 = 1 - \pi'_0 \). Clearly, comparing \( \pi' \) with \( \pi'_0 \) is equivalent to comparing \( \pi \) with \( \pi_0 \). We now consider two criteria to assess satisfactory agreement. The first requires that \( \pi < \pi_0 \), without regard to how the total proportion outside \([l, u]\) is distributed between the two tails. This way the hypotheses of satisfactory agreement introduced in (2.1) become

\[
H_1: \pi \geq \pi_0, \text{ versus } K_1: \pi < \pi_0. \quad (3.1)
\]

The second criterion is more specific and requires that both \( \pi_l \) and \( \pi_u \) be less than \( \pi_0/2 \), thus ensuring \( \pi < \pi_0 \). So the hypotheses of satisfactory agreement become

\[
H_2: \pi_l \geq \pi_0/2 \text{ or } \pi_u \geq \pi_0/2, \text{ versus } K_2: \pi_l < \pi_0/2 \text{ and } \pi_u < \pi_0/2. \quad (3.2)
\]

This criterion is more stringent than the first one but it protects against having more than \( \pi_0/2 \) proportion in any one tail. The hypotheses (3.2) are suitable for the application of the intersection-union principle. They are tested by first testing both the component hypotheses

\[
H_{2l}: \pi_l \geq \pi_0/2, \text{ versus } K_{2l}: \pi_l < \pi_0/2, \quad (3.3)
\]

and \( H_{2u}: \pi_u \geq \pi_0/2, \text{ versus } K_{2u}: \pi_u < \pi_0/2, \quad (3.4)\)

and then combining the results.

This chapter is organized as follows. Section 3.2 contains a brief review of literature. We describe the distribution-free tests of hypotheses (3.1) and (3.2) in Section 3.3. A test of (3.2) based on the normality of \( D \) is presented in Section 3.4.
Section 3.5 investigates the connection between tolerance intervals and the tests of Section 3.3 and 3.4. The connection between the hypotheses (3.1) and the ones used in the acceptance sampling literature is pointed out in Section 3.6. We conclude the chapter in Section 3.7 with a discussion.

3.2 Literature review

In the literature on assessing agreement only the test of (3.1) under the normality of $D$ has been discussed. The two relevant references are Lin (2000) and Lin et al. (2002). Lin (2000) first computes $\kappa$ such that $Pr(|D| < \kappa) = \pi'_0$, where $\pi'_0$ is pre-specified, and compares it with the threshold $u$. He refers to this as the TDI (total deviation index) approach. On the other hand, Lin et al. (2002) first compute $\pi' = Pr(|D| < u)$ for a pre-specified $u$, and compare it with the threshold $\pi'_0$. They refer to it as the CP (coverage probability) approach.

3.2.1 TDI approach

Since $D$ follows $N(\mu, \sigma)$ distribution, solving $Pr(|D| < \kappa) = \pi'_0$ for $\kappa$ gives

$$\kappa = \sigma \left( \chi^2_1(1 - \pi'_0, \mu^2/\sigma^2) \right)^{1/2},$$

(3.5)

where $\chi^2_1(\alpha, \delta)$ is the upper $\alpha$-th quantile of a non-central $\chi^2_1$-distribution with non-centrality parameter $\delta$. In terms of $\kappa$ the hypotheses (3.1) become

$$H_1 : \; \kappa \geq u \; \text{ (i.e., at most } \pi'_0 \text{ area of } N(\mu, \sigma) \text{ lies in } [-u, u]), \; \text{ versus,}$$

$$K_1 : \; \kappa < u \; \text{ (i.e., more than } \pi'_0 \text{ area of } N(\mu, \sigma) \text{ lies in } [-u, u]).$$

(3.6)
Lin (2000) argues that the inference based on the estimate of \( \kappa \) in (3.5) is intractable. So approximates it by

\[
\kappa^* = \left( (\mu^2 + \sigma^2) \chi_1^2 (1 - \pi_0') \right)^{1/2} = (\mu^2 + \sigma^2)^{1/2} z(\pi_0'/2),
\]

where \( \chi_1^2(\alpha) \) and \( z(\alpha) \) are the upper \( \alpha \)-th quantiles of central \( \chi_1^2 \) and \( N(0, 1) \) distributions, respectively. For assessing agreement, he modifies the hypotheses (3.6) to

\[
H_1^*: \kappa^* \geq u, \text{ versus } K_1^*: \kappa^* < u.
\]

Lin estimates \( (\mu^2 + \sigma^2) \) by \( \sum_{i=1}^n D_i^2 / (n-1) \) and performs statistical inference through the asymptotic normality of \( \ln \left( \sum_{i=1}^n D_i^2 / (n-1) \right) \).

Clearly, the approximation \( \kappa^* \) for \( \kappa \) will be good only when \( \mu^2/\sigma^2 \) is small. Lin gives a range of values of \( \pi_0' \) and \( \mu^2/\sigma^2 \) where this approximation can be considered reasonable. However, \( \mu^2/\sigma^2 \) is usually not known in advance. This limits the practical utility of this procedure. Further, the test of (3.8) in Lin (2000) has asymptotic level \( \alpha \) and is consistent. This leads to the undesirable property that in the limiting case where \( n \to \infty \), with probability 1, there will be some regions where the agreement is satisfactory (\( \pi < \pi_0 \)), but the test will infer on the contrary (\( \pi \geq \pi_0 \)) and vice versa. Such regions depend on whether the approximation \( \kappa^* \) is conservative (i.e., \( \kappa^* > \kappa \)) or liberal (i.e., \( \kappa^* < \kappa \)).

### 3.2.2 CP approach

Lin et al. (2002) consider the hypotheses (3.1) in the following equivalent form:

\[
H_1: \pi' \leq \pi_0', \text{ versus } K_1: \pi' > \pi_0'.
\]
Since $D$ follows $N(\mu, \sigma)$ distribution, \[ \pi' = \Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right) \]. It is estimated as \[ \hat{\pi}' = \Phi\left(\frac{u-\hat{\mu}}{\hat{\sigma}^*}\right) - \Phi\left(\frac{l-\mu}{\hat{\sigma}^*}\right) \], where \( \hat{\sigma}^2 = (n-1)\hat{\sigma}^2/(n-3) \). They perform inference by invoking the large sample normality of \( \ln \left(\frac{\hat{\pi}'/(1 - \hat{\pi}')}{\pi'/(1 - \pi')}\right) \) with mean \( \ln \left(\frac{\pi'/(1 - \pi')}{\pi'/(1 - \pi')}\right) \) and variance

\[
\frac{1}{(n-3)(\pi')^2(1-\pi')^2} \left\{ \left[ \phi\left(\frac{u-\mu}{\sigma}\right) - \phi\left(\frac{l-\mu}{\sigma}\right) \right]^2 + \frac{1}{2} \left[ \phi\left(\frac{u-\mu}{\sigma}\right) - \frac{1}{2} \phi\left(\frac{u-\mu}{\sigma}\right) \phi\left(\frac{l-\mu}{\sigma}\right) \left[ \phi\left(\frac{u-\mu}{\sigma}\right) - \phi\left(\frac{l-\mu}{\sigma}\right) \right]^2 \right] \right\}.
\]

Our preliminary empirical investigation indicates that the test is very conservative.

### 3.3 Distribution-free approach

In this section we discuss distribution-free tests of the hypotheses of satisfactory agreement. Let us start with the hypotheses (3.1). We use the sign test for them. This test rejects $H_1$ at level $\alpha$ if $B \leq b_\alpha(n, \pi_0)$, where $B$ is the number of observed $D$’s outside $[l, u]$, and the critical point $b_\alpha(n, \pi_0)$ is the $\alpha$-th quantile of binomial $(n, \pi_0)$ distribution. Thus $b_\alpha(n, \pi_0)$ is the largest integer such that $Pr(\text{Binomial}(n, \pi_0) \leq b_\alpha(n, \pi_0)) \leq \alpha$. It may be noted that the randomized sign test is uniformly most powerful among the class of distribution-free level-$\alpha$ tests of (3.2) (see Lehmann, 1986, pp. 106-107). For determining sample size to ensure a power of at least $1 - \beta$ at $\pi = \pi_1 < \pi_0$, one can numerically compute $n$ such that $Pr(B \leq b_\alpha(n, \pi_0) | \pi = \pi_0) \leq \alpha$ and $Pr(B \leq b_\alpha(n, \pi_0) | \pi = \pi_1) \geq 1 - \beta$. See Rosner (2000, ch. 7) for normal approximation of the sign test and the associated sample size formula.

We consider the hypotheses (3.2) next. The level-$\alpha$ sign test of (3.3) rejects $H_2$ if $B_l \leq b_\alpha(n, \pi_0/2)$, where $B_l$ is number of $D$’s that are less than $l$. Likewise, the level-$\alpha$ sign test of (3.4) rejects $H_2u$ if $B_u \leq b_\alpha(n, \pi_0/2)$, where $B_u$ is the number of
$D$'s that are more than $u$. So, the level-$\alpha$ intersection-union test of (3.2) rejects $H_2$ if $B_l \leq b_\alpha(n, \pi_0/2)$ and $B_u \leq b_\alpha(n, \pi_0/2)$.

Let $D_{(i)}$, $i = 1, 2, \ldots, n$, denote the order statistics of the $D_i$'s. Note that $B_l \leq b_\alpha(n, \pi_0/2)$ is equivalent to $D_{(b_\alpha(n, \pi_0/2)+1)} \geq l$. Similarly, $B_u \leq b_\alpha(n, \pi_0/2)$ is equivalent to $D_{(n-b_\alpha(n, \pi_0/2))} \leq u$. Hence, in terms of the order statistics, the intersection-union test of (3.2) rejects $H_2$ if $[D_{(b_\alpha(n, \pi_0/2)+1)}, D_{(n-b_\alpha(n, \pi_0/2))}] \subseteq [l, u]$. The sample size computation for this test has not been studied yet.

The confidence intervals for $\pi$, $\pi_l$ and $\pi_u$ related with the respective hypotheses (3.1), (3.3) and (3.4) have been discussed in Rosner (2000, sec 6.8 and 6.10). See also the recommendations of Brown, Cai and DasGupta (2001).

**EXAMPLE 3.1.** These data, taken from Bland and Altman (1986), were introduced in Chapter 1. Here Oxygen saturation measurements are taken on 72 subjects using the pulsed oximeter saturation ($X_1$) method and the oxygen saturation meter ($X_2$). Figure 3.1 has their scatter plot. The measurements are highly correlated. Some of the $X_2$ measurements are a bit higher than $X_1$ measurements and others are a bit lower. But their means appear to be the same. Further, the mean versus difference plot in Figure 3.1 does not have any obvious pattern, indicating that the variances of the two methods are also similar. Overall, the level of agreement between the two instruments looks good. The parameter estimates are: $(\hat{\xi}_1, \hat{\xi}_2, \hat{\eta}_1, \hat{\eta}_2, \hat{\nu}) = (89.08, 89.50, 8.76, 8.70, 0.99)$ and $(\hat{\mu}, \hat{\sigma}) = (-0.42, 1.21)$.

The histogram of the differences, $D = X_2 - X_1$, is given in Figure 3.2. Clearly, normality is not a reasonable assumption for them. In addition, the Shapiro-Wilk test of normality yields a $p$-value of 0.005. Hence the distribution-free approach described in this section would be appropriate.
Figure 3.1: Scatter plot (top) and the plot of means versus differences (bottom) for the Oxygen saturation data (Example 3.1).

43
Figure 3.2: Histogram for the differences, $D = X_2 - X_1$, in Example 3.1.
Suppose the threshold interval \([l, u] = [-2, 2]\) and the cutoff probability \(\pi_0 = 0.20\) are specified. Then, \(B_l = \text{number of } D's < -2 = 1\), \(B_u = \text{number of } D's > 2 = 6\), and \(B = \text{number of } D's \text{ outside } [l, u] = B_l + B_u = 7\). Note that only 1.4% of the \(D's\) lie in the left tail whereas 8.3% of them lie in the right tail. Let us first consider a 5% level test of (3.1). Since \(B \leq b_{0.05}(72, 0.20) = 8\), the sign test rejects \(H_1\). Thus with these thresholds the data support the claim of satisfactory agreement. Next, we consider a 5% test of (3.2). The critical value in this case is \(b_{0.05}(72, 0.10) = 2\). Since \(B_l \leq 2\) and \(B_u > 2\), the individual 5% level sign test rejects \(H_{2l}\) but fails to reject \(H_{2u}\). Hence, the 5% level intersection-union test of (3.2) fails to reject \(H_2\). So the data do not offer sufficient evidence of satisfactory agreement through the hypotheses (3.2). These data emphasize the difference between the kind of protection the hypotheses (3.1) and (3.2) provide.

\(\square\)

**Remark 3.1.** The hypotheses (3.1) and (3.2) can also be used for assessing agreement when \(D\) represents a ratio instead of a difference. This ratio form is usually suggested for assessing individual bioequivalence (see e.g., Anderson and Hauck, 1990).

### 3.4 Normality based approach

In this section we assume that \(D\) follows \(N(\mu, \sigma)\) distribution and describe the intersection-union test of (3.2). This test was first considered by Liu and Chow (1997) in the context of assessing individual bioequivalence. It may be noted that the region under \(K_2\) is an open triangle with \((\mu, \sigma)\) co-ordinates \((l, 0)\), \((u, 0)\) and \((0, u/z(\pi_0/2))\).

From Lehmann (1986, p. 296), \(\{D : \hat{\sigma}^{-1}(l - \hat{\mu}) \leq -a_n\}\) is a level-\(\alpha\) critical region for the hypotheses (3.3) where

\[
a_n = a(n, \alpha, z(\pi_0/2)) = \sqrt{n} t_{n-1}^*(\alpha, n^{1/2} z(\pi_0/2)),
\]

(3.9)
and \( t_k^*(\alpha, \delta) \) is the upper \( \alpha \)-th quantile of a non-central \( t_k \)-distribution with non-centrality parameter \( \delta \). Likewise, \( \{ D : \hat{\sigma}^{-1}(u - \hat{\mu}) \geq a_n \} \) is a level-\( \alpha \) critical region for the hypotheses (3.4). It may be noted that these two tests are uniformly most powerful in the class of invariant tests for their respective hypotheses (see Lehmann, 1986, p. 296). From the intersection-union principle the critical region

\[
C = \{ D : [\hat{\mu} - a_n \hat{\sigma}, \hat{\mu} + a_n \hat{\sigma}] \subseteq [l, u] \},
\]

has level \( \alpha \) for testing (3.2). From Theorem 2 in Berger and Hsu (1996), it follows that this test actually has size \( \alpha \).

**Proposition 3.1.** The critical region \( C \) for testing (3.2) has the power function:

\[
p(u, \pi_0, \alpha, n, \mu, \sigma) = \int_0^{(n \sigma)^{-1/2}(n-1)^{1/2}} \left( \Phi \left( n^{1/2} \varphi_u - a_n (n - 1)^{-1/2} (n w)^{1/2} \right) 
- \Phi \left( - n^{1/2} \varphi_l + a_n (n - 1)^{-1/2} (n w)^{1/2} \right) \right) 
\times h_{n-1}(w) dw,
\]

where \( \varphi_u = (u - \mu)/\sigma, \varphi_l = (l - \mu)/\sigma \) and \( h_{n-1}(\cdot) \) is the density function of \( \chi^2_{n-1} \)-distribution.

**Proof:** It follows on the lines of Proposition 2.1. \( \Box \)

The above expression for the power function appears similar to the one given by Proposition 2.1 for the test of (2.10). However, the upper limits of integration and the coefficients of \( w^{1/2} \) are different in the two expressions.

The power function depends on \((\mu, \sigma)\) only through \((\mu/\sigma, u/\sigma)\), and is symmetric in \( \mu \) about zero. Figure 3.3 presents the contours of \( p(1, 0.20, 0.05, n, \mu, \sigma) = 0.05, 0.90 \) in the \((\mu, \sigma)\)-plane for \( n = 15, 30 \). Notice that there is no loss of generality in taking
Figure 3.3: Contours of the power function $p(-1, 0.20, 0.05, n, \mu, \sigma)$ for $n = 15, 30$.

$u = 1$. Further, due to the symmetry of $p$ with respect to $\mu$, we only focus on $\mu \geq 0$. We observe that the type-I error of 0.05 is achieved on the boundary dividing the null and alternative regions of (3.2). This verifies that the test has size $\alpha$. The surface plot of $p(1, 0.20, 0.05, 15, \mu, \sigma)$ is given in Figure 3.4. The power function increases as $\sigma$ and $|\mu|$ decrease. It is close to one when $\sigma$ is close to zero. The expression for $p$ given by Theorem 3.1 is also useful for sample size computation.

EXAMPLE 2.2 (continued): The normality of $D$ was discussed on page 24. Suppose $\pi_0 = 0.20$ is specified as the cutoff for $\pi$. Let us now consider a 5% level test of (3.1). The value of $a_{99}$, defined in (3.9), is 15.20. Hence the test would reject $H_2$ if $[-12.93, -5.59] \subseteq [l, u]$. Thus to infer satisfactory agreement $u$ must be more
than 12.93. Such a threshold for $D$ looks unrealistic. But recall that the two methods seemed to differ by about 9.5 units on average. So, if we subtract 9.5 from every measurement by the Nadler method, the 5% level test of (3.1) would reject if $[-3.43, 3.91] \subseteq [l, u]$. Thus, if $u = 4$ were specified, we would infer satisfactory agreement.

3.5 Tests based on tolerance intervals

The tolerance intervals play a natural role in assessing agreement through coverage probabilities. See Guttman (1988) for an introduction to this topic. Several authors have suggested using them for assessing individual bioequivalence (see, e.g., Brown, Iyer and Wang, 1997 and Chinchilli, 1996). However, a formal relationship between
tolerance intervals and the tests of hypotheses (3.1)-(3.4) does not appear to be investigated, at least in the assessing agreement literature. The purpose of this section is to explore this relationship. In particular, we describe how tolerance intervals can be used for testing hypotheses on one-tailed or two-tailed coverage probabilities, and show that:

(a) For testing (3.1), the test based on two-tailed tolerance interval is not a good choice.

(b) For testing (3.3) and (3.4), the tests based on appropriate one-tailed tolerance intervals and the best tests are equivalent.

A consequence of (b) is that the intersection-union test of (3.2) by combining the tests of (3.3) and (3.4) discussed in the previous two sections, and by combining their tests based on one-tailed tolerance intervals are equivalent. Recall that this intersection-union test has the form: reject $H_2$ if $[V_1, V_2] \subseteq [l, u]$, where the statistics $V_1$ and $V_2$ are associated with the tests of (3.3) and (3.4), respectively. We also show that:

(c) The interval $[V_1, V_2]$ can be interpreted as a large sample two-tailed tolerance interval.

In the present context, it is natural to work with $\pi'_{l}, \pi'_{u}, \pi'$ and $\pi'_{0}$ instead of $\pi_l, \pi_u, \pi$ and $\pi_0$. So, we will consider the following equivalent forms for the respective hypotheses (3.1), (3.3) and (3.4):

\begin{align*}
    H_1 : \pi' &\leq \pi'_{0}, \quad \text{versus} \quad K_1 : \pi' > \pi'_{0} \quad (3.11) \\
    H_{2l} : \pi'_{l} &\leq (1 + \pi'_{0})/2, \quad \text{versus} \quad K_{2l} : \pi'_{l} > (1 + \pi'_{0})/2, \quad (3.12) \\
    \text{and} \quad H_{2u} : \pi'_{u} &\leq (1 + \pi'_{0})/2, \quad \text{versus} \quad K_{2u} : \pi'_{u} > (1 + \pi'_{0})/2. \quad (3.13)
\end{align*}

49
First we describe the test of (3.11) using a two-tailed tolerance interval. Let \([T_1, T_2]\) be a two-tailed \(\pi'_0\)-content tolerance interval for \(D\) with confidence level \(1 - \alpha\), \(0.5 < \alpha < 1\). This interval is such that

\[
Pr\left( F(T_2) - F(T_1) \geq \pi'_0 \right) \geq 1 - \alpha, \tag{3.14}
\]

where the probability is computed under the assumption that \(D\) has cdf \(F\). Consider the critical region

\[
C' = \{ D : [T_1, T_2] \subseteq [l, u] \}. \tag{3.15}
\]

**Proposition 3.2.** The critical region \(C'\) has level \(\alpha\) for testing (3.11).

**Proof:** By definition we have,

\[
Pr\left( D \in C' \mid F \in H_1 \right) = Pr\left( [T_1, T_2] \subseteq [l, u] \mid F \in H_1 \right)
\]

\[
\leq Pr\left( F(T_2) - F(T_1) \leq \pi' \mid F \in H_1 \right)
\]

\[
\leq Pr\left( F(T_2) - F(T_1) \leq \pi'_0 \right)
\]

\[
\leq \alpha,
\]

from (3.14). \(\square\)

In this way we can use tolerance intervals to test (3.11), but notice that \(C'\) may give a conservative test.

Next, we consider the test of (3.12) using a right-tailed tolerance interval. The test of (3.13) using a left-tailed test is analogous. Let \([T_i, \infty)\) be a right-tailed tolerance interval for \(D\) with content \((1 + \pi'_0)/2\) and confidence level \(1 - \alpha\). On the lines of Proposition 3.2 it is easy to see that the critical region

\[
C'_i = \{ D : T_i \geq l \}. \tag{3.16}
\]

has level \(\alpha\) for testing (3.12).
3.5.1 Distribution-free tests

We now establish the superiority of the sign test for (3.1) or equivalently (3.11) over its test using a two-tailed distribution-free tolerance interval. This addresses the issue (a). The sign test was discussed in Section 3.3.

An interval $[D_{(i)}, D_{(j)}]$ is a $\pi'_0$-content distribution-free tolerance interval with confidence level $1 - \alpha$, if $j - i$ is the smallest integer that satisfies

$$Pr \left( \text{Beta}(j - i, n - j + i + 1) \geq \pi'_0 \right) \geq 1 - \alpha. \quad (3.17)$$

For specified $\alpha$ and $\pi'_0$ one may have to take sufficiently large $n$ for $j - i$ satisfying (3.17) to exist. The smallest such $n$ satisfies $n(1 - \pi'_0)(\pi'_0)^{n-1} + (\pi'_0)^n \leq \alpha$ (see Arnold, Balakrishnan and Nagaraja, 1992, pp. 185-186), and in this case $i = 1, j = n$. For a given $n$ satisfying this condition, there may exist several $i$ and $j$ such that (3.17) holds. Hence the choice of a two-tailed tolerance interval is not unique. To accommodate this non-uniqueness we enlarge the critical region $C'$ to

$$C^* = \{ D : [D_{(i)}, D_{(j)}] \subseteq [l, u] \text{ for some } 1 \leq i < j \leq n \text{ and } j - i \text{ satisfies (3.17)} \}. \quad (3.18)$$

A simple calculation shows that the resulting test of (3.11) also has level $\alpha$.

**Proposition 3.3.** The critical region $C^*$ gives a distribution-free test of (3.11) at level $\alpha$.

**Proof:** The test is distribution-free since the tolerance interval is. We have

$$Pr(D \in C^* \mid F \in H_1) = Pr(\text{At least } j - i + 1 \text{ of the } D's \text{ are in } [l, u] \mid F \in H_1)$$

$$= Pr(\text{Binomial}(n, \pi') \geq j - i + 1), \text{ with } \pi' \leq \pi'_0$$

$$\leq Pr(\text{Binomial}(n, \pi'_0) \geq j - i + 1)$$

51
\[ \leq Pr(\text{Binomial}(n, \pi'_0) \geq j - i) \]
\[ = Pr(\text{Beta}(j - i, n - j + i + 1) \leq \pi'_0) \]
\[ \leq \alpha, \]

from (3.17). Hence the result holds. \qed

A comparison of \( C^* \) and the critical region of the non-randomized sign test for (3.11) gives \( j - i = n - b_a(n, \pi_0) \). Since \([D(i), D(j)] \subseteq [l, u]\) for some \( i, j \) iff at least \( j - i + 1 \) (\( = n - b_a(n, \pi_0) + 1 \)) of the \( D \)'s are in \([l, u]\), it follows that the sign test rejects \( H \) of (3.11) more often than its test using \( C^* \), at the same level. Consequently, the former test is uniformly more powerful than the latter test.

The issue (b) is taken up next. The right-tailed tolerance interval for \( D \) with content \((1 + \pi'_0)/2\) and confidence level \(1 - \alpha\) has the form \([D(n-k+1), \infty)\), where \( k \) is the smallest integer that satisfies

\[ Pr(\text{Beta}(k, n - k + 1) \geq (1 + \pi'_0)/2) \geq 1 - \alpha. \quad (3.19) \]

The test of (3.12) defined by the critical region (3.16) rejects \( H_2 \) if \( D(n-k+1) \geq l \). Note that the condition (3.19) is same as \( Pr(\text{Binomial}(n, \pi_0/2) \leq n - k) \leq \alpha \), since \( \pi_0 = 1 - \pi'_0 \). Thus, it follows that \( n - k = b_a(n, \pi_0/2) \) and hence this test is equivalent to the sign test described in Section 3.3. This result contrasts with the two-tailed case discussed above.

Finally, we consider the issue (c). The following result shows that the interval \([D(b_a(n,\pi_0/2)+1), D(n-b_a(n,\pi_0/2))]\), that comes up in the intersection-union test of (3.2), can be interpreted as a large sample two-tailed tolerance interval with content \( \pi'_0 \) and confidence level \(1 - \alpha\).
**Proposition 3.4.** \( \lim_{n \to \infty} Pr\left(F(D(n-b_{n}(n,\pi_0/2))) - F(D(b_{n}(n,\pi_0/2)+1)) \geq \pi'_0 \right) \geq 1 - \alpha, \)

where \( F \) is the cdf of \( D. \)

**Proof:** The interval \([D(b_{n}(n,\pi_0/2)+1), D(n-b_{n}(n,\pi_0/2))]\) consists of \( n - 2b_{n}(n,\pi_0/2) - 1 \) blocks of order statistics. Hence the coverage probability of this interval follows \( \text{Beta}(n - 2b_{n}(n,\pi_0/2) - 1, 2b_{n}(n,\pi_0/2) + 2) \) distribution (see Guttman, 1988). From the relationship between beta and binomial distributions it suffices to establish that

\[
\lim_{n \to \infty} Pr\left(\text{Binomial}(n, \pi'_0) \geq n - 2b_{n}(n,\pi_0/2) - 1\right) \leq \alpha.
\]

Further, the result would hold from the central limit theorem if we show that \( e_{n} = (n(1 - \pi'_0) - 2b_{n}(n,\pi_0/2) - 1)/(n\pi'_0(1 - \pi'_0))^{1/2} \) goes to a limit that is at least \( z(\alpha) \).

By definition, \( b_{n}(n,\pi_0/2) \) is such that \( Pr\left(\text{Binomial}(n, \pi_0/2) \leq b_{n}(n,\pi_0/2)\right) \leq \alpha, \) for every \( n \). Now an application of the central limit theorem and some algebraic manipulations show that \( (n(1 - \pi'_0) - 2b_{n}(n,\pi_0/2))/(n(1 - \pi'_0)(1 + \pi'_0))^{1/2} \) converges to \( z(\alpha) \). Consequently, \( e_{n} \to ((1 + \pi'_0)/\pi'_0)^{1/2} z(\alpha) \) as \( n \to \infty \). This limit is more than \( z(\alpha) \) because \( (1 + \pi'_0)/\pi'_0 > 1 \), and \( z(\alpha) > 0 \) as \( 0.5 < \alpha < 1. \) \( \square \)

### 3.5.2 Normality based tests

Here we discuss the issues (a)-(c) assuming that \( D \) follows \( N(\mu, \sigma) \) distribution. We first consider the issue (a) by showing that the test of (3.1) or equivalently (3.11) using the usual normal theory tolerance interval is not consistent.

A tolerance interval of the form \([T_1, T_2] = [\hat{\mu} - b_n \hat{\sigma}, \hat{\mu} + b_n \hat{\sigma}]\) is used for \( N(\mu, \sigma) \) distribution, where \( b_n \) is chosen to satisfy (3.14). Thus for testing (3.11), the critical region \( C' \) in (3.15) becomes \( C' = \{D : [\hat{\mu} - b_n \hat{\sigma}, \hat{\mu} + b_n \hat{\sigma}] \subseteq [l, u]\} \). Only a large sample approximation to \( b_n = b(n, \alpha, \pi'_0) \) is available in the form

\[
b_n = b_1(n, \pi'_0) b_2(n, \alpha),
\] (3.20)
where \( b_1(n, \pi'_0) \) satisfies

\[
\Phi\left(n^{-1/2} + b_1(n, \pi'_0)\right) - \Phi\left(n^{-1/2} - b_1(n, \pi'_0)\right) = \pi'_0. \tag{3.21}
\]

and \( b_2(n, \alpha) \) is such that

\[
b_2(n, \alpha) = \left(\frac{(n - 1)}{\chi^2_{n-1}(1 - \alpha)}\right)^{1/2}. \tag{3.22}
\]

From Guttman (1988) it follows that for this \( b_n \),

\[
\lim_{n \to \infty} Pr\left(\Phi\left(\frac{\hat{\mu} + b_n \hat{\sigma} - \mu}{\sigma}\right) - \Phi\left(\frac{\hat{\mu} - b_n \hat{\sigma} - \mu}{\sigma}\right) \geq \pi'_0\right) = 1 - \alpha. \tag{3.23}
\]

Hence the confidence level of the tolerance interval is approximately \( 1 - \alpha \) for large \( n \). Note that the interval is designed to cover only the central 100\( \pi'_0 \)% area of \( N(\mu, \sigma) \) distribution. Let

\[
\varphi_l = (l - \mu)/\sigma, \quad \varphi_u = (u - \mu)/\sigma, \quad q_0 = z\left((1 - \pi'_0)/2\right),
\]

\[
\psi_0 = z(\alpha) q_0/(2 + q_0^2)^{1/2} \quad \text{and} \quad \rho_0 = (2 - q_0^2)/(2 + q_0^2). \tag{3.24}
\]

**Proposition 3.5.** \( \lim_{n \to \infty} Pr(D \in C') = \)

\[
\begin{cases}
0, & \text{if } \varphi_l > -q_0 \text{ or } \varphi_u < q_0,
\Phi(-\psi_0), & \text{if } \varphi_l = -q_0, \varphi_u > q_0, \text{ or } \varphi_l < -q_0, \varphi_u = q_0,
\Phi(-\psi_0) - \Phi_2(\psi_0, -\psi_0; \rho_0), & \text{if } \varphi_l = -q_0, \varphi_u = q_0,
1, & \text{if } \varphi_l < -q_0, \varphi_u > q_0.
\end{cases}
\]

Notice that the region \( \{\varphi_l < -q_0, \varphi_u = q_0\} \cup \{\varphi_l = -q_0, \varphi_u > q_0\} \) falls under \( K_1 \).

Thus the test of (3.11) using \( C' \) is not consistent. In fact there are several parameter configurations in \( K_1 \) where the asymptotic power is zero.

**Proof of Proposition 3.5:** By definition,

\[
Pr(D \in C') = Pr([\hat{\mu} - b_n \hat{\sigma}, \hat{\mu} + b_n \hat{\sigma}] \subseteq [l, u])
\]

\[
= Pr\left(n^{1/2}(\varphi_l + b_n) \leq Z - b_n n^{1/2}((n - 1)^{-1/2} W^{1/2} - 1), \right.
\]

\[
\left. n^{1/2}(\varphi_u - b_n) \geq Z + b_n n^{1/2}((n - 1)^{-1/2} W^{1/2} - 1)\right), \tag{3.25}
\]

54
where \( Z = n^{1/2}(\hat{\mu} - \mu)/\sigma \) and \( W = (n - 1)\sigma^2/\sigma^2 \).

First we show that

\[
\lim_{n \to \infty} n^{1/2} (b_n - q_0) = 2^{-1/2} q_0 z(\alpha). \tag{3.26}
\]

From (3.20), \( b_n = b_1(n, \pi_0') b_2(n, \alpha) \), and from (3.21), \( b_1(n, \pi_0') \to q_0 \) as \( n \to \infty \). Using Taylor’s Theorem we get,

\[
\Phi\left(n^{-1/2} + b_1(n, \pi_0')\right) = \Phi(q_0) + \left(n^{-1/2} + b_1(n, \pi_0') - q_0\right) \phi(y_1), \tag{3.27}
\]

\[
\Phi\left(n^{-1/2} - b_1(n, \pi_0')\right) = \Phi(-q_0) + \left(n^{-1/2} - b_1(n, \pi_0') + q_0\right) \phi(y_2), \tag{3.28}
\]

where \( \phi(\cdot) \) is the standard normal density, \( y_1 \) lies within \( n^{-1/2} + b_1(n, \pi_0') \) and \( q_0 \), and \( y_2 \) is within \( n^{-1/2} - b_1(n, \pi_0') \) and \( -q_0 \). Now, on subtracting (3.28) from (3.27), using (3.21) and re-arranging terms we obtain

\[
n^{1/2}(b_1(n, \pi_0') - q_0) = \left(\phi(y_2) + \phi(y_1)\right)^{-1} \left(\phi(y_2) - \phi(y_1)\right). \tag{3.29}
\]

Since \( b_1(n, \pi_0') \to q_0 \), \( \lim_{n \to \infty} n^{1/2}(b_1(n, \pi_0') - q_0) = 0 \) by the symmetry of \( \phi \).

For \( b_2(n, \alpha) \) defined by (3.22), we have

\[
Pr\left(n^{1/2}\left((n - 1)^{1/2} W^{-1/2} - 1\right) \geq n^{1/2}(b_2(n, \alpha) - 1)\right) = \alpha. \tag{3.30}
\]

From the delta method, it follows that \( n^{1/2}\left((n - 1)^{1/2} W^{-1/2} - 1\right) \xrightarrow{d} N(0, 2^{-1/2}) \).

Hence from (3.30), \( n^{1/2}(b_2(n, \alpha) - 1) \) must converge to \( 2^{-1/2} z(\alpha) \).

Finally from (3.20), we can write

\[
n^{1/2}(b_n - q_0) = b_1(n, \pi_0') n^{1/2}(b_2(n, \alpha) - 1) + n^{1/2}(b_1(n, \pi_0') - q_0).
\]

But the first term on the right hand side converges to \( 2^{-1/2} q_0 z(\alpha) \). Thus (3.26) is established.
We now return to (3.25). From (3.26), \( b_n \to q_0 \) as \( n \to \infty \). Hence upon using the delta method, and the independence of \( \hat{\mu} \) and \( \hat{\sigma} \) we can conclude that 
\[
Z \pm n^{1/2} \left( (n - 1)^{-1/2} W^{1/2} - 1 \right) \overset{d}{\to} N\left(0, (1 + q_0^2/2)^{1/2}\right).
\]
Furthermore, the random variables 
\[
Z + n^{1/2} \left( (n - 1)^{-1/2} W^{1/2} - 1 \right) \quad \text{and} \quad Z - n^{1/2} \left( (n - 1)^{-1/2} W^{1/2} - 1 \right)
\]
jointly converge in distribution to a bivariate normal random variable with correlation \( \rho_0 \), where \( \rho_0 \) is defined in (3.24). Now the result follows upon applying (3.26) and using the properties of a cdf.

Next, we consider the issue (b). The right-tailed tolerance interval for \( D \) with content \((1 + \pi_0^2)/2\) and confidence level \( 1 - \alpha \) has form \([\hat{\mu} - a_n \hat{\sigma}, \infty)\), where \( a_n \) is defined by (3.9). Thus, the test of (3.12) defined by the critical region (3.16) and the UMP invariant test discussed in Section 3.4 are equivalent.

Finally, we consider the issue (c). The following result shows that the interval 
\[[\hat{\mu} - a_n \hat{\sigma}, \hat{\mu} + a_n \hat{\sigma}]\]
can be interpreted as a large sample tolerance interval.

**Proposition 3.6.**

\[
\lim_{n \to \infty} Pr \left( \Phi \left( \frac{\hat{\mu} + a_n \hat{\sigma} - \mu}{\sigma} \right) - \Phi \left( \frac{\hat{\mu} - a_n \hat{\sigma} - \mu}{\sigma} \right) \geq \pi_0 \right) = 1 - \alpha.
\]

**Proof:** By the definition of \( a_n \), we have for every fixed \( n \), 
\[
Pr \left( \hat{\sigma}^{-1} (1 - \hat{\mu}) < -a_n \big| \pi_1 = \pi_0/2 \right) = \alpha. \]
As in Proposition 3.5, we write this probability as

\[
Pr \left( Z - a_n n^{1/2} \left( (n - 1)^{-1/2} W^{1/2} - 1 \right) > n^{1/2} (a_n - q_0) \right) = \alpha,
\]
where \( Z = n^{1/2}(\hat{\mu} - \mu)/\sigma \), \( W = (n - 1)\hat{\sigma}^2/\sigma^2 \) and \( q_0 \) is defined in (3.24). There we also saw that \( Z - a_n n^{1/2} \left( (n-1)^{-1/2} W^{1/2} - 1 \right) \overset{d}{\to} N\left(0, (1 + q_0^2/2)^{1/2}\right). \) Hence the above equation implies that \( n^{1/2} (a_n - q_0) \) must converge to \((1 + q_0^2/2)^{1/2} z(\alpha)\). Consequently, from (3.26) it follows that \( \lim_{n \to \infty} (a_n - b_n) = 0. \) Moreover, since the normal density
function is bounded, we can see that \( \Phi((\hat{\mu}+a_n\hat{\sigma}-\mu)/\sigma) - \Phi((\hat{\mu}-a_n\hat{\sigma}-\mu)/\sigma) - \Phi((\hat{\mu}+b_n\hat{\sigma}-\mu)/\sigma) + \Phi((\hat{\mu}-b_n\hat{\sigma}-\mu)/\sigma) \xrightarrow{n \to \infty} 0 \). The result is now established by using (3.23).

3.6 Connection with acceptance sampling

In this section we point out the connection between the hypotheses (3.1) and the ones that are used in the acceptance sampling literature. Acceptance sampling (or sampling inspection) refers to the examination of the quality of items in a lot by inspecting a random sample from it and deciding usually whether to accept or reject the lot. This approach of providing quality assurance has a long history. For a description of this area see Bowker and Goode (1952), Lieberman and Resnikoff (1955), Montgomery (2001, ch. 14-15), Hamilton and Lesperance (1995), and Lei and Vardeman (1998).

In acceptance sampling, the parameter of interest is the lot quality measured by the proportion of non-conforming items in the lot, say \( \pi \). The higher this proportion is the worse the lot quality is. There is either a lower specification limit, an upper specification limit or both that define a range of acceptable quality. Thus \( \pi \) is the proportion outside the specification limits.

The decision to accept or reject the lot is based on a test of the hypotheses

\[ H'_1 : \pi \leq \pi_0 \quad \text{versus} \quad K'_1 : \pi > \pi_0, \]  

(3.31)

where \( \pi_0 \) is a specified acceptable quality level. Acceptance of \( H'_1 \) amounts to accepting the lot and its rejection amounts to rejecting the lot.

In the terminology of acceptance sampling, we can regard a difference \( D \) to be of "good" quality if it falls in the threshold interval \([l, u]\), otherwise it is of "bad" quality. Thus \( \pi = 1 - Pr(l \leq D \leq u) \), our measure of agreement between the two instruments,
actually measures the proportion of “bad” quality differences. The resemblance of the
hypotheses (3.1) of satisfactory agreement and the above hypotheses (3.31) should
now be noted. The null of (3.1) is the alternative of (3.31) and vice versa. Due to
this connection, the techniques of one area can be adapted in the other area, without
much effort in most of the cases.

The issue of performance of tests of (3.31) has been discussed at length in the ac-
ceptance sampling literature. A test is evaluated in terms of the operating characteris-
tic (OC) curve. This curve plots the probability of accepting the lot, \( Pr(\text{Accept } H_0) \),
against the lot quality \( \pi \). It turns out that when the quality characteristic follows
\( N(\mu, \sigma) \) distribution and the specification limit is two-tailed (the case of our interest),
the probability of acceptance associated with the available tests depend not only on
the lot quality \( \pi \) but also on \( (\mu, \sigma) \). For a fixed \( \pi \), there are uncountably many com-
binations of the proportions outside \( l \) and \( u \) such that the total proportion outside \([l, u]\)
is \( \pi \). In other words, there are uncountably many combinations of \( (\mu, \sigma) \) that produce
the same \( \pi \). So, the OC curve is actually a band. The attention then focuses on the
thickness and the steepness of the OC band (see e.g., Lei and Vardeman, 1998). Its
thickness is defined as the difference of the maximum and the minimum probabilities
of acceptance for a given \( \pi \). A thin band is desirable so that all the lots of the same
quality \( \pi \) are dealt with in the same way. This usually ensures that a lot of “better”
quality will most likely have a higher probability of acceptance than a lot of “worse”
quality. The steepness of OC band is also an attractive property since the steeper the
band is, the greater is the discriminatory power of the test to classify “good” lots as
“good” and “bad” lots as “bad”. These issues are also important in evaluating the
tests for assessing agreement through coverage probability.
3.7 Discussion

The coverage probability approach provides an intuitively appealing way for assessing agreement. To claim satisfactory agreement a large proportion of differences must lie close to zero. These bounds on differences and the target proportion are readily interpretable. Hence this formulation makes it easier for the investigator to specify various thresholds.

In this chapter we summarized the distribution-free and normality based tests of hypotheses of satisfactory agreement (3.1) and (3.2). We explained the relationship between these tests and the tolerance intervals. We also touched upon the connection between hypotheses (3.1) and the ones that are of interest in the acceptance sampling literature. This connection brings to our disposal a whole body of tools and some issues that were hitherto unknown in the literature on assessing agreement. However, more work is needed to give a complete solution for the normality based test of (3.1).

A FORTRAN program for computing sample size to plan a method comparison study that would use the normality based test of (3.2) is available at www.stat.ohio-state.edu/~hnn/pankaj_programs.html or at www.stat.ohio-state.edu/~pankaj. We briefly illustrate this computation here. For specified \((\alpha, \pi_0, u/\sigma, \mu/\sigma)\) and the desired power \(1 - \beta\) one can solve the equation \(p(u, \pi_0, \alpha, n, \mu, \sigma) = 1 - \beta\) for \(n\) numerically. The expression for the power function \(p(\cdot)\) was derived in Proposition 3.5. A natural approach for determining sample size for specified \((\pi_0, \alpha)\) is to ensure \(1 - \beta\) power for the test when, for some specified \(\pi_1 < \pi_0\), there is exactly \(100(1 - \pi_1)\%\) central area of \(N(\mu, \sigma)\) distribution in \([l, u] = [-1, 1]\). This area requirement is equivalent to \(1 - \Phi((1-\mu)/\sigma) = \Phi((-1-\mu)/\sigma) = \pi_1/2\), which in turn gives \((\mu, \sigma) = (0, 1/z(\pi_1/2))\).
<table>
<thead>
<tr>
<th>$(\pi_0, \pi_1)$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.15, 0.05)</td>
<td>77</td>
</tr>
<tr>
<td>(0.20, 0.05)</td>
<td>44</td>
</tr>
<tr>
<td>(0.20, 0.10)</td>
<td>134</td>
</tr>
<tr>
<td>(0.25, 0.05)</td>
<td>30</td>
</tr>
<tr>
<td>(0.25, 0.10)</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 3.1: Sample sizes required to achieve 80% power with normality based 5% level test of hypotheses (3.2).

To give an idea of how the sample sizes look like with these choices, Table 3.1 shows their values for various combinations of $(\pi_0, \pi_1)$ associated with $(\alpha, \beta) = (0.05, 0.20)$.

The approaches of this and the previous chapter together provide several ways to assess agreement. In a given situation, the appropriate approach to use would depend on the goals of the experiment and the investigator’s ability to specify the thresholds. If the goal is just to check for satisfactory agreement then one of the coverage probability approaches can be used. However, if one is also interested in knowing the nature and the extent of disagreement, the approaches of Section 2.3 and Section 2.4 are recommended.
CHAPTER 4

SELECTING THE INSTRUMENT THAT IS CLOSEST TO A GOLD STANDARD

4.1 Introduction and notation

In the last two chapters, we discussed the problem of assessing satisfactory agreement between two instruments. Now we consider the problem of comparing two instruments with a gold standard to find the one that agrees the most with it. We will refer to this instrument as the best one.

In the literature on measuring agreement, St. Laurent (1998), and Hutson, Wilson and Geiser (1998) are the two relevant references for this problem. St. Laurent considers the estimation of spin-spin relaxation times in magnetic resonance with spline multi-echo, Lorentz single-echo and spline single-echo methods. He uses the multi-echo method as the gold standard and compares it with the two single-echo methods. Hutson et al. deal with fractional area change for detecting the limiting boundaries of endocardium surface. They treat a committee of experienced echocardiographers as gold standard, and are interested in finding whether a computer algorithm or a human echocardiographer agrees more with it.
Let us now introduce some additional notation. The triplet \((G, X_1, X_2)\) denotes the measurements on a typical subject by the gold standard, the first and the second instrument, respectively, and \(D_i\) is the difference \(G - X_i\) \((i = 1, 2)\). We take \(\theta_i = E(D_i^2)\) as the measure of agreement between the \(i\)-th instrument and the gold standard. Let \([1], [2]\) be the unknown indices such that \(\theta_{[1]} \leq \theta_{[2]}\). Thus the instrument associated with \(\theta_{[1]}\) is the best one.

We assume that \((D_1, D_2)\) follows bivariate normal distribution with mean \(\mu = (\mu_1, \mu_2)\), standard deviation \(\sigma = (\sigma_1, \sigma_2)\) and correlation \(\rho\), where \((\mu, \sigma, \rho) \in \Omega = \{(\mu, \sigma, \rho) : -\infty < \mu_i < \infty, 0 < \sigma_i < \infty, |\rho| < 1; i = 1, 2\}\). Thus \((D_1^2, D_2^2)\) follows a continuous bivariate distribution with mean \((\theta_1, \theta_2)\), standard deviation \((\psi_1, \psi_2)\) and correlation \(\gamma\), where \(\theta_i = \mu_i^2 + \sigma_i^2, \psi_i^2 = 2\sigma_i^4 + 4\mu_i^2\sigma_i^2\) and \(\gamma = (\psi_1\psi_2)^{-1}[2\rho\sigma_1\sigma_2(\rho\sigma_1\sigma_2 + 2\mu_1\mu_2)]\). Next, let \(\psi_a^2 = Var(D_1^2 - D_2^2) = \psi_1^2 + \psi_2^2 - 2\gamma\psi_1\psi_2, \lambda_i = \ln(\theta_i), \tau_i^2 = \psi_i^2/\theta_i^2\) and \(\tau^2_a = \tau_1^2 + \tau_2^2 - 2\gamma\tau_1\tau_2\). The unknown indices \([1]\) and \([2]\) refer to the true best and the true worst between the two instruments. Thus \(E(D_{[i]}) = \mu_{[i]}\) and \(E(D_{[i]}^2) = \theta_{[i]}\). Similarly \(\sigma_{[i]}\) and \(\psi_{[i]}\) are the standard deviations of \(D_{[i]}\) and \(D_{[i]}^2\), respectively.

Suppose \((D_{1j}, D_{2j}), j = 1, 2, \ldots,\) is a sequence of i.i.d. observations on \((D_1, D_2)\). This way \((D_{1j}^2, D_{2j}^2)\)'s are also i.i.d. observations on \((D_1^2, D_2^2)\). Based on the first \(m\) observations, let \(\hat{\mu}_i(m), \hat{\sigma}_i^2(m)\) and \(\hat{\rho}(m)\) be the usual unbiased estimators of \(\mu_i, \sigma_i^2\) and \(\rho\). The estimators of functions of these five parameters are constructed by plugging-in their sample counterparts, and are denoted by the usual hat notation. When it is clear from the context, we will suppress the sample size as an argument.

Finally, let \(z(\alpha)\) and \(t_k(\alpha)\) denote the upper \(\alpha\)-th quantiles of \(N(0, 1)\) distribution and \(t_k\)-distribution, respectively.
All the simulation results reported in this chapter focus on the subset

\[ \{(\mu, \sigma) : \mu_i^2 + \sigma_i^2 = \theta_i, \rho = 0.5, \sigma_i > 0; \ i = 1, 2\} \]

of the parameter space \( \Omega \) of the bivariate normal distribution of \( (D_1, D_2) \). The value of \( \rho \) is fixed at 0.5 as the results do not vary markedly across \( \rho \) unless it is close to \( \pm 1 \).

Note that when \( (\theta_1, \theta_2) \) is fixed, any function \( g \) of \( (\mu, \sigma, \rho) \), such that \( \mu_i^2 + \sigma_i^2 = \theta_i \ (i = 1, 2) \) satisfies \( g(\mu_1, \mu_2, \sigma, \rho) = g(\mu_1, -\mu_2, \sigma, \rho) = g(-\mu_1, \mu_2, \sigma, \rho) = g(-\mu_1, -\mu_2, \sigma, \rho) \).

Due to this symmetry we only need to investigate half of the above subset, namely, \( \{(\mu, \sigma) : \mu_i^2 + \sigma_i^2 = \theta_i, \rho = 0.5, \sigma_i > 0, \theta_1^{1/2} < \mu_1 \leq 0; \ i = 1, 2\} \). Let

\[
\begin{align*}
\Omega_1(\theta_1, \theta_2) &= \{ (\mu_1, \mu_2) : \mu_1 \in [-\theta_1^{1/2} + 0.001, 0], \mu_2 \in [-\theta_2^{1/2} + 0.001, \theta_2^{1/2} - 0.001] \}, \\
\Omega_2(\theta_1, \theta_2) &= \{ (\mu_1, \mu_2) : \mu_1 = -\theta_1^{1/2} + 0.001, 0, \text{ or } \mu_2 = \pm \left( \theta_2^{1/2} - 0.001 \right) \}, \\
\Omega_3(\theta_1, \theta_2) &= \Omega_1(\theta_1, \theta_2) - \Omega_2(\theta_1, \theta_2).
\end{align*}
\]

(4.1)

The region \( \Omega_3(\theta_1, \theta_2) \) excludes the situation where \( \sigma_1 \) or \( \sigma_2 \) are close to zero as they are not usually encountered in real applications. We simulate on a \( 11 \times 21 \) grid of equally spaced values of \( (\mu_1, \mu_2) \in \Omega_1(\theta_1, \theta_2) \). Thus the number of combinations of \( (\mu_1, \mu_2) \) in \( \Omega_1(\theta_1, \theta_2) \) are \( 11 \times 21 = 231 \), and in \( \Omega_3(\theta_1, \theta_2) \) are \( 10 \times 19 = 190 \). Frequently, we will use the five-number summary, i.e., (min, 1st quartile, median, 3rd quartile, max), to summarize the simulation results.

Rest of this chapter is organized as follows. In Section 4.2 we outline the two current approaches and draw attention to their drawbacks. We present two single-stage procedures in Section 4.3 and illustrate their use with the fractional area change data of Hutson et al. (1998). Section 4.4 describes a two-stage procedure. The chapter concludes in Section 4.5 with a discussion.
4.2 Current approaches

St. Laurent (1998) assumes a random effects model $X_i = G + \epsilon_i$, where $\epsilon_i, i = 1, 2,$ are correlated random variables with zero means and distributed independently of $G$. This model assumes that the two instruments and the gold standard have the same means. Hence the equal agreement is equivalent to the equality of variances of $\epsilon_1$ and $\epsilon_2$. For the inference he uses a nonparametric bootstrap confidence interval for the difference of the intraclass correlations between $(X_1, G)$ and between $(X_2, G)$. This approach is ad hoc and the equality of means assumption cannot always be justified from practical considerations.

Hutson et al. (1998) consider the large sample $100(1 - \alpha)\%$ confidence interval for $\theta_1/(\theta_1 + \theta_2)$ constructed from a sample of size $n$, and given as

$$\frac{\hat{\theta}_1}{\theta_1 + \theta_2} \pm \frac{z(\alpha/2)}{\sqrt{n}} \frac{1}{(\theta_1 + \theta_2)^2} \left( \hat{\psi}_1^2 \hat{\theta}_2^2 - 2 \hat{\gamma} \hat{\psi}_1 \hat{\psi}_2 \hat{\theta}_1 \hat{\theta}_2 + \hat{\psi}_2^2 \hat{\theta}_1^2 \right)^{1/2}.$$

This approach infers first (second) instrument as the true best if the upper (lower) bound of the confidence interval in (4.2) is less (greater) than 0.5, and is indecisive if the interval contains 0.5.

We now present simulation results showing that, for moderate values of $n$, the actual coverage probability of this confidence interval is much less than its nominal level. Without loss of generality we assume $\theta_1 = 1.0$ as it amounts to re-scaling all the observations on $D_1$ by $\theta_1^{-1/2}$. For $\theta_2$, we first take $\theta_2 = 1.2$, and later explore the effect of larger $\theta_2$ values.

We compute the proportion of times the 95% confidence interval in (4.2) covers the true value of $\theta_1/(\theta_1 + \theta_2)$ in 50,000 repetitions over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$, defined by (4.1). Figure 4.1 has surface plots of these empirical coverage probabilities.
for $n = 20, 30$. In the former case the five-number summary for the proportions is 
$(0.904, 0.912, 0.915, 0.925, 0.935)$, and in the latter case it is $(0.916, 0.923, 0.926, 
0.931, 0.936)$. It may be noted that roughly 65% of the values are less than 0.92 when $n = 20$, and roughly 70% values are less than 0.93 when $n = 30$. Further, these 
summaries for $n = 20$ over the grids $\Omega_1(1.0, 1.5)$ and $\Omega_1(1.0, 2.0)$ are $(0.905, 0.912, 
0.916, 0.925, 0.935)$ and $(0.906, 0.913, 0.917, 0.925, 0.935)$, respectively. Thus the 
overall picture does not change as $\theta_2$ moves further away from $\theta_1$. Hence it is clear 
that the confidence interval (4.2) results in substantial under-coverage. Notice also 
that the under-coverage is more severe in the region $\Omega_3(\theta_1, \theta_2)$, which excludes the 
region where $\sigma_1$ or $\sigma_2$ are close to zero.

Using sample moments based estimators of $(\theta_1, \theta_2, \psi_1, \psi_2, \gamma)$ instead of estimating 
them by substituting the estimates of $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ leads to even more under-
coverage than shown in Figure 4.1. In this case the five number summaries of the 
coverage probabilities in the region $\Omega_1(1.0, 1.2)$ are $(0.878, 0.882, 0.887, 0.891, 0.929)$ 
for $n = 20$ and $(0.901, 0.905, 0.907, 0.911, 0.936)$ for $n = 30$.

4.3 Single-stage procedures

Since we are concerned with finding the instrument associated with $\theta_{[1]}$, a natural 
approach is to use the multiple comparisons with the best (MCB) procedure to con-
struct large sample (normality based) confidence interval for either $\theta_1 - \theta_2$ or $\lambda_1 - \lambda_2$, 
$\lambda_i = \ln(\theta_i)$.

Our simulation studies reveal that the normal approximation does not work well 
for the studentized estimator of $\theta_1 - \theta_2$ with moderate values of $n$, whereas it does
Figure 4.1: Empirical coverage probability of 95% confidence interval for $\theta_1/(\theta_1 + \theta_2)$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$, for $n = 20, 30$. 
for $\lambda_1 - \lambda_2$. This can be seen from Figure 4.2 where the average Shapiro-Wilk $p$-values for testing normality of estimators based on 15 observations are plotted over the grid $\Omega_1(1.0, 1.2)$. These averages are obtained in the following manner: for a fixed value of $(\mu_1, \mu_2)$ in $\Omega_1(1.0, 1.2)$, (i) generate 15 i.i.d. observations on $(D_1, D_2)$ and compute the studentized estimate of the parametric function, (ii) repeat step (i) 2000 times, then a fit normal distribution to these 2000 observations and compute $p$-value for the Shapiro-Wilk test, and finally (iii) repeat steps (i)-(ii) 1000 times to get the distribution of $p$-values and compute its average.
Roughly 90% of the averages shown in the Figure 4.2 for $\theta_1 - \theta_2$ are less than 0.05 and about the same percentage of them are more than 0.10 for $\lambda_1 - \lambda_2$. Further, the five-number summary of averages in the latter case is (0.006, 0.380, 0.694, 0.744, 0.781). Most of the points where the averages in this case are close to zero fall in the region $\Omega_2(1.0, 1.2)$, which is not of concern in applications. The five-number summary over the region $\Omega_3(1.0, 1.2)$ becomes (0.065, 0.541, 0.716, 0.746, 0.781). Thus apparently the normal approximation works well in the case of $\lambda_1 - \lambda_2$ for most of the parameter combinations even with $n = 15$, and it does not in the case of $\theta_1 - \theta_2$. This inference is in agreement with Lambert and Hall (1982) who mention that a $p$-value is stochastically at least as large as a uniform (0, 1) random variable under the associated null hypothesis and it is stochastically smaller than this random variable under the associated alternative hypothesis.

Furthermore, since the $\theta_i$’s are mean squared deviations, it may be easier to specify a threshold for practical equivalence of the two instruments in terms of $\lambda_1 - \lambda_2 = \ln(\theta_1/\theta_2)$ than in terms of $\theta_1 - \theta_2$. In this sense, $\lambda_1 - \lambda_2$ is a more interpretable parameter than $\theta_1 - \theta_2$.

**Remark 4.1.** The normal approximation for the studentized estimator of $\theta_1/(\theta_1 + \theta_2)$ is worse than that for $\lambda_1 - \lambda_2$ for $n = 15$. The plot of average Shapiro-Wilk $p$-values for testing normality in the former case is given in Figure 4.3 for the grid $\Omega_4(1.0, 2.0)$, and their five-number summary is (0, 0.033, 0.095, 0.189, 0.741).

These considerations lead us to settle on $\lambda_1 - \lambda_2$ as the metric for comparing the two instruments. With the help of standard asymptotic arguments, it is easy to see
Figure 4.3: Average Shapiro-Wilk p-values for testing normality of studentized estimator of $\theta_1/(\theta_1 + \theta_2)$ for $n = 15$ and $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$. 
that the coverage probability of the interval
\[
\hat{\lambda}_1 - \hat{\lambda}_2 \pm n^{-1/2} t_{n-1}(\alpha/2) \hat{\tau}_d.
\] (4.3)
converges to $1 - \alpha$ as $n \to \infty$. Hence this interval is a large sample 100(1 - $\alpha$)% confidence interval for $\lambda_1 - \lambda_2$. Besides, it can also be interpreted as a large sample unconstrained MCB confidence interval for $\lambda_1 - \lambda_2$ derived using Tukey’s method (see Hsu, 1996, p. 103). Using this interval, one infers the first instrument to be the best one if its upper (lower) bound is negative (positive), and is indecisive otherwise. We now turn to the following interval for $\lambda_1 - \lambda_2$, which is constrained to contain zero:
\[
\left[ \min \left\{ 0, \hat{\lambda}_1 - \hat{\lambda}_2 - n^{-1/2} t_{n-1}(\alpha) \hat{\tau}_d \right\}, \max \left\{ 0, \hat{\lambda}_1 - \hat{\lambda}_2 + n^{-1/2} t_{n-1}(\alpha) \hat{\tau}_d \right\} \right].
\] (4.4)

**Proposition 4.1.** In the limiting case as $n \to \infty$, the coverage probability of the interval in (4.4) is at least $1 - \alpha$, for every fixed $(\mu, \sigma, \rho) \in \Omega$.

**Proof:** For every fixed $n$, the interval in (4.4) covers zero with probability one. So, the result holds when $\lambda_1 = \lambda_2$. Now, we show that the limiting coverage probability of this interval is $1 - \alpha$ when $(\mu, \sigma, \rho) \in \Omega - \{ \lambda_1 = \lambda_2 \}$. Following Hsu (1996, pp. 84-85), the event in (4.4) is implied by the event \( \left\{ \hat{\lambda}_{[2]} - \hat{\lambda}_{[1]} - \lambda_{[2]} + \lambda_{[1]} > -n^{-1/2} t_{n-1}(\alpha) \hat{\tau}_d \right\} \), where [1] and [2] are the unknown indices of the true best and the true worst instruments, respectively. To see this, let $Y = t_{n-1}(\alpha) \hat{\tau}_d / n^{1/2}$ and note that
\[
\begin{align*}
\left\{ \hat{\lambda}_{[2]} - \hat{\lambda}_{[1]} - \lambda_{[2]} + \lambda_{[1]} > -Y \right\} \\
= \left\{ \lambda_{[2]} - \lambda_{[1]} < \hat{\lambda}_{[2]} - \hat{\lambda}_{[1]} + Y \right\} \\
\subseteq \left\{ 0 < \lambda_1 - \lambda_2 < \hat{\lambda}_1 - \hat{\lambda}_2 + Y \right\} \cap \left\{ 0 < \lambda_2 - \lambda_1 < \hat{\lambda}_2 - \hat{\lambda}_1 + Y \right\} \\
= \left\{ \lambda_1 - \lambda_2 < \max \{0, \hat{\lambda}_1 - \hat{\lambda}_2 + Y\} \right\} \cap \left\{ \lambda_2 - \lambda_1 < \max \{0, \hat{\lambda}_2 - \hat{\lambda}_1 + Y\} \right\}.
\end{align*}
\]
The last event is equivalent to the event in (4.4) with probability one. Hence it is enough to show that

$$\lim_{n \to \infty} Pr\left( \hat{\lambda}_{[2]} - \hat{\lambda}_{[1]} - \lambda_{[2]} + \lambda_{[1]} > -n^{-1/2} t_{n-1}(\alpha) \hat{\tau}_d \right) = 1 - \alpha.$$

Using standard asymptotic arguments we can show that $n^{1/2}(\hat{\lambda}_{[2]} - \hat{\lambda}_{[1]} - \lambda_{[2]} + \lambda_{[1]})/\hat{\tau}_d \overset{d}{\to} N(0,1)$ as $n \to \infty$. Hence the limit on the RHS exists and the equality holds because $t_{n-1}(\alpha) \to z(\alpha)$ as $n \to \infty$. 

Thus the interval in (4.4) is a large sample constrained MCB confidence interval for $\lambda_1 - \lambda_2$. If we rule out the possibility that $\lambda_1 = \lambda_2$, we can infer the first (second) instrument to be the best one if the upper (lower) bound of this interval equals zero; otherwise we are indecisive. Note however that when this procedure identifies one instrument to be the best, say the first one, it does not provide a negative upper bound on $\theta_1 - \theta_2$. Hence we cannot infer how much better the first instrument is compared to the second one. On the other hand, the unconstrained interval (4.3) allows such an inference. But, sacrificing this ability and ruling out the possibility that $\lambda_1 = \lambda_2$ allows the constrained interval to infer an instrument to be the best one more frequently than the unconstrained interval with the same asymptotic level. This sharper inference is desirable as our main goal is to find the best of the two instruments. In addition, assuming $\lambda_1 \neq \lambda_2$ is reasonable from practical viewpoint as it amounts to assuming that the two instruments do not agree equally well with the gold standard.

Figure 4.4 has the empirical coverage probabilities for the unconstrained and the constrained 95% confidence intervals for $\lambda_1 - \lambda_2$ with $n = 15$ over the grid $\Omega_1(1.0, 1.2)$. The five-number summaries for the two sets of probabilities are respectively (0.946,
0.949, 0.951, 0.955, 0.962) and (0.940, 0.946, 0.948, 0.952, 0.968). Moreover, in the latter case, only 10% values are less than 0.945, half of them lying in the region $\Omega_2(1.0,1.2)$. This overall situation does not change much when simulated on the grid $\Omega_1(1.0,1.5)$ or $\Omega_2(1.0, 2.0)$. Hence it may be concluded that both these intervals have reasonably good small sample properties for $n \geq 15$.

**EXAMPLE 4.1.** We now apply the two single-stage procedures developed in Section 4.3 to the fractional area change data from Hutson et al. (1998). This data was introduced in Chapter 1. We focus on the first expert echocardiographer ($X_1$), the Autonomous Boundary Detection Algorithm ($X_2$) and the committee of expert echocardiographers ($G$), and ask whether the computer algorithm or the human expert is closer to the gold standard.

Figure 4.5 has the scatter plot and Figure 4.6 has the marginal normal probability plots for the differences $D_1 = G - X_1$ and $D_2 = G - X_2$. (The scatter plot is also given in Hutson et al.'s paper.) Apparently, the marginal normality assumption for $D_1$ and $D_2$ look fine. Besides, the Shapiro-Wilk test of normality yields $p$-values of 0.46 for $D_1$ and 0.70 for $D_2$. For the null hypotheses that $(D_1, D_2)$ follows a bivariate normal distribution, Mardia's test gives $p$-values of 0.44 for skewness and 0.70 for kurtosis (see Mardia and Foster, 1983). Hence we may assume $(D_1, D_2)$ to have a bivariate normal distribution. This conclusion is further supported by the Q-Q plots of radii and angles. These plots are suggested by Hutchinson and Lai (1990, p. 60) for graphically assessing bivariate normality. It may be noted that this bivariate normal assumption is not discussed in Hutson et al. (1998).
Figure 4.4: Empirical coverage probabilities of 95% unconstrained MCB confidence interval for $\lambda_1 - \lambda_2$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0,1.2)$ (top) and its 95% constrained version (bottom), for $n = 15$. 

73
Figure 4.5: Scatter plot for the differences in the fractional area change data.
The estimates of the parameters of \((D_1, D_2)\) are \((\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho}) = (1.67, 4.68, 8.41, 6.47, 0.34)\). Hence, \((\hat{\theta}_1, \hat{\theta}_2, \hat{\psi}_1, \hat{\psi}_2, \hat{\gamma}) = (73.56, 63.78, 103.96, 84.72, 0.15)\) and \((\hat{\lambda}_1, \hat{\lambda}_2, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_d) = (4.30, 4.16, 1.41, 1.33, 1.79)\). The 95% unconstrained and constrained confidence intervals for \(\lambda_1 - \lambda_2\) are \([-0.85, 1.14]\) and \([-0.67, 0.96]\), respectively. Thus both these procedures remain indecisive in choosing between the human expert and the computer algorithm. Hutson et al. (1998) also have the same conclusion but we now have more faith in it as the confidence intervals here have good small sample properties, in particular with \(n = 15\). Due to the one-to-one correspondence, the confidence intervals for \(\theta_1 / (\theta_1 + \theta_2)\) can be derived from those for \(\lambda_1 - \lambda_2\). On using the 95% unconstrained version it is obtained as \([0.30, 0.76]\), and on using the 95% constrained version it is obtained as \([0.34, 0.72]\). Curiously, the confidence interval
reported in Hutson et al. is the same as the latter when rounded to two decimal places. But this is due to round off errors and is not true in general.

\[ \square \]

### 4.4 A two-stage procedure

As was the case in the above example, a single-stage procedure may fail to distinguish between the two instruments. However, a two-stage procedure does not have this disadvantage. For this procedure, the experimenter pre-specifies \( \delta (> 0) \) and \( \alpha \) \((0, 0.5)\) such that (i) whenever \(|\lambda_1 - \lambda_2| < \delta\) the two instruments are considered 

*practically equivalent*, and (ii) the probability requirement

\[
P_r(\text{correct selection } | |\lambda_1 - \lambda_2| \geq \delta) \geq 1 - \alpha,
\]

is satisfied. Here “correct selection” refers to the inference of selecting the true best instrument. The threshold \( \delta \) is called the threshold for indifference in the terminology of the indifference-zone selection procedures. In a two-stage procedure, the first-stage sample is used to estimate the covariance matrix, and this estimate is used to compute a second-stage sample size that ensures the desired accuracy. Here we present a two-stage procedure, say \( \mathcal{D}(\delta, \alpha, m) \), that satisfies the above probability requirement approximately.

**Procedure** \( \mathcal{D}(\delta, \alpha, m) \):

**Stage 1:** Select a random sample of size \( m \), compute \( \hat{\tau}_d^2(m) \) and define

\[
N_m = \max \left\{ \left[ \tau_{m-1}^2(\alpha) \hat{\tau}_d^2(m) \delta^{-2} \right], m \right\},
\]

where \( \hat{\tau}_d^2(m) \) is the estimate of \( \tau_d^2 = \tau_1^2 + \tau_2^2 - 2\gamma \tau_1 \tau_2 \) with \( \tau_i^2 = \psi_i^2/\theta_i^2 \), \( \theta_i = \mu_i^2 + \sigma_i^2 \), \( \psi_i^2 = 2\sigma_i^4 + 4\mu_i^2\sigma_i^2 \) and \( \gamma = (\psi_1\psi_2)^{-1} \left( 2\rho \sigma_1 \sigma_2 (\rho \sigma_1 \sigma_2 + 2\mu_1\mu_2) \right) \). Here the parameters \((\mu_i, \sigma_i, \rho)\) are estimated by their respective sample counterparts.
Stage 2: Take \( N_m - m \) additional i.i.d. observations if necessary and compute the estimates \( \hat{\lambda}_i(N_m), i = 1, 2 \). Then select the instrument that produces the smaller estimate as the best of the two.

Consider the interval

\[
\min \left\{ 0, \hat{\lambda}_1(N_m) - \hat{\lambda}_2(N_m) - \delta \right\} \leq \lambda_1 - \lambda_2 \leq \max \left\{ 0, \hat{\lambda}_1(N_m) - \hat{\lambda}_2(N_m) + \delta \right\}.
\] (4.7)

Following Hsu (1996, pp. 101-102) we now show that for a fixed \((\delta, \alpha, m)\), the event given by (4.7) implies the event that the procedure \( \mathcal{D}(\delta, \alpha, m) \) selects the true best instrument whenever \(|\lambda_1 - \lambda_2| \geq \delta\).

Let the subscripts \((1)\) and \((2)\) denote the indices of the best and the worst instruments in the sample, respectively. Thus \(\lambda_{(1)}\) and \(\lambda_{(2)}\) are the true means of the best and the worst instruments in the sample. Note that

\[
\left\{ \min \left\{ 0, \hat{\lambda}_1(N_m) - \hat{\lambda}_2(N_m) - \delta \right\} \leq \lambda_1 - \lambda_2 \leq \max \left\{ 0, \hat{\lambda}_1(N_m) - \hat{\lambda}_2(N_m) + \delta \right\} \right\}
= \left\{ \lambda_1 - \lambda_2 \leq \max \left\{ 0, \hat{\lambda}_1(N_m) - \hat{\lambda}_2(N_m) + \delta \right\} \right\}
= \left\{ \lambda_2 - \lambda_1 \leq \max \left\{ 0, \hat{\lambda}_2(N_m) - \hat{\lambda}_1(N_m) + \delta \right\} \right\}
\subseteq \left\{ \lambda_{(1)} - \lambda_{(2)} \leq \max \left\{ 0, \hat{\lambda}_{(1)}(N_m) - \hat{\lambda}_{(2)}(N_m) + \delta \right\} \right\}
\subseteq \left\{ \lambda_{(1)} - \lambda_{(2)} < \delta \right\},
\]

with probability one since \(\hat{\lambda}_{(1)}(N_m) - \hat{\lambda}_{(2)}(N_m) < 0\) with probability one. When \(\lambda_{(2)} - \lambda_{(1)} \geq \delta\), the last event implies the event \(\{\lambda_{(1)} = \lambda_{(1)}\}\), which is equivalent to the event \(\{\hat{\lambda}_{(1)}(N_m) \leq \hat{\lambda}_{(2)}(N_m)\}\). Consequently, to establish the validity of this procedure it suffices to show that for large \(m\), the coverage probability of the interval (4.7) is \(1 - \alpha\) approximately.
Proposition 4.2. Consider the procedure \( \mathcal{D}(\delta_m, \alpha, m) \) where \( \delta_m = \Delta / m^{1/2} (\Delta > 0) \). As \( m \to \infty \), the limiting coverage probability of the interval (4.7) with \( \delta_m \) in place of \( \delta \) is at least \( 1 - \alpha \), for every fixed \((\mu, \sigma, \rho) \in \Omega\).

Proof: For every fixed \( m \), the interval (4.7) covers zero with probability one. So, when \( \lambda_1 = \lambda_2 \), the result holds. Let us now focus on \((\mu, \sigma, \rho) \in \Omega - \{\lambda_1 = \lambda_2\}\). As in the case of Proposition 4.1, it is enough to establish that

\[ \lim_{m \to \infty} Pr \left( \hat{\lambda}_{[2]}(N_m) - \hat{\lambda}_{[1]}(N_m) - \lambda_{[2]} + \lambda_{[1]} > -\delta_m \right) \geq 1 - \alpha. \]

Using the definition of \( N_m \) in (4.6), it is easy to see that \( N_m \delta_m^2 \overset{p}{\to} \max \{ z^2(\alpha) \tau_d^2, \Delta^2 \} \), since \( \tilde{\tau}_d^2(m) \overset{p}{\to} \tau_d^2 \). Note that \( \hat{\tau}_d^2(m) / \tilde{\tau}_d^2(N_m) \overset{p}{\to} 1 \). Hence from part (b) of Lemma A.3 we conclude that \( N_m^{1/2} \left( \hat{\lambda}_{[2]}(N_m) - \hat{\lambda}_{[1]}(N_m) - \lambda_{[2]} + \lambda_{[1]} \right) / \hat{\tau}_d(m) \overset{d}{\to} N(0, 1) \). Now the result readily follows by noticing that the limit of \( N_m^{1/2} \delta_m / \hat{\tau}_d(m) \) is at least \( z(\alpha) \). \qed

Hence when \( m \) is large, the interval (4.7) is an approximate \( 100(1 - \alpha)\% \) confidence interval for \( \lambda_1 - \lambda_2 \). This interval may be conservative. Further, with probability \( 1 - \alpha \) approximately, it guarantees to infer that the mean of the best instrument in the sample is within \( \delta \) of the mean of the true best. It may be noted that the two-stage MCB procedure of Matejcik and Nelson (1995) is not applicable here as their assumptions are not satisfied.

One important issue now is to find a reasonable value of \( m \), the first-stage sample size. We would also like to know how conservative the procedure is and how large \( N_m \) can be expected. We use simulation for this investigation. In Section 4.3 we discovered that \( n = 15 \) was large enough for using the asymptotic methods in the case of \( \lambda_1 - \lambda_2 \). So a natural starting point is \( m = 15 \).

78
The first plot in Figure 4.7 has empirical coverage probabilities of the 95% confidence interval (4.7) obtained through 50,000 replications of the procedure with \( m = 15 \). The second plot contains the associated average two-stage sample sizes, hereafter termed as the average sample number (ASN). These probabilities and ASN’s are plotted for \( \alpha = 0.05, 0.20 \) over the grid \( \Omega_1(1.0, 1.2) \), defined in (4.1). It is observed that in the case of \( \alpha = 0.20 \), the smallest probability shown on the top plot is 0.791, roughly 7% of the values are less than 0.795 and only 15% are more than 0.805. Similarly, in the case of \( \alpha = 0.05 \), no value is less than 0.949 and only 20% are more than 0.955. In both cases, the probabilities are much above the target only when \( \sigma_1 \) and \( \sigma_2 \) are both close to zero. Hence overall it appears that the procedure \( \mathcal{D}(\ln(1.2), \alpha, 15) \) satisfies the probability requirement (4.5) well and is not too conservative. Further, the five-number summary of ASN over the grid \( \Omega_3(1.0, 1.2) \) is (18.5, 56.3, 64.6, 70.4, 83.0) when \( \alpha = 0.20 \), and it is (68.2, 230.0, 264.8, 288.2, 340.1) when \( \alpha = 0.05 \). It may be noted that the behavior of probabilities or ASN’s over the grid \( \Omega_1(1.0, 1.5) \) or \( \Omega_1(1.0, 2.0) \) remain similar to those over \( \Omega_1(1.0, 1.2) \) described above.

**EXAMPLE 4.1 (continued):** We take the 15 observations on \((D_1, D_2)\) in the fractional area change data as the first-stage sample. The parameter estimates based on this sample were given on page 75. We compute the second-stage sample size \( N_{15} \) using (4.6) such that the procedure \( \mathcal{D}(\delta, \alpha, 15) \) satisfies the probability requirement (4.5) approximately. These values are presented in Table 4.1 for \( \delta = \ln(1.1), \ln(1.2), \ln(1.3) \) and \( \alpha = 0.05, 0.10, 0.20 \). Usually \( \delta = \ln(1.2) \) is taken as the threshold. □

**Remark 4.2.** It may not be reasonable to assume that \( \tau_d^2 \) is known as it cannot be determined without the knowledge of the parameters that also determine \( \lambda_1 \) and
Figure 4.7: Empirical coverage probability of the confidence interval for $\lambda_1 - \lambda_2$ given by the procedure $\mathcal{D}(\ln(1.2), \alpha, 15)$ (top) and the associated ASN (bottom), over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$. 

80
<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$1 - \alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(1.1)</td>
<td>0.80</td>
</tr>
<tr>
<td>ln(1.2)</td>
<td>0.90</td>
</tr>
<tr>
<td>ln(1.3)</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 4.1: Two-stage sample sizes required by the procedure $\mathcal{D}(\delta, \alpha, 15)$ to satisfy the probability requirement (4.5) approximately.

$\lambda_2$. This is why we used the first-stage sample to estimate $\tau_d^2$. However, when it can be assumed to be known, a single-stage procedure with the sample size $z_\alpha^2 \tau_d^2 / \delta^2$ and the decision rule same as in the two-stage procedure $\mathcal{D}(\delta, \alpha, m)$ would satisfy the probability requirement (4.5) approximately.

### 4.5 Discussion

In this chapter we presented two large sample single-stage procedures and one two-stage procedure for finding the best of two instruments when compared with a gold standard. These procedures use $\lambda_1 - \lambda_2$ as the metric for comparison, where $\lambda_i = \ln(\text{mean squared difference of } X_i \text{ and } G)$. We discovered that a sample size of 15 is adequate for the single-stage procedures and also as the initial sample size for the two-stage procedure. It may be noted that the ASN varies very little when the first stage sample size increases from 15 to 30, unless $\sigma_1$ or $\sigma_2$ are close to zero. Hence from practical considerations a larger initial sample would be preferred as it would provide more reliable estimates of $\tau_d$.

Among the proposed single-stage procedures we prefer the constrained confidence interval over its unconstrained version as the sharper inference with the former is
desirable. The two-stage procedure always infers one instrument to be the best and gives a confidence interval for $\lambda_1 - \lambda_2$. This procedure is useful when a threshold for practical equivalence can be specified. If the primary concern is the accuracy of the procedure then following the tradition of requiring 80% power for a test, one can set $1 - \alpha = 0.80$ as the lower bound on the probability of correct selection.

The procedures introduced in this chapter remain valid if the bivariate normality of $(D_1, D_2)$ is replaced by the weaker assumptions that $Var(D_1^4), Var(D_2^4)$ and $Var(D_2^2D_2^2)$ be non-zero and finite, and $|Corr(D_1^2, D_2^2)| < 1$. In this setup, however, $\hat{\theta}_i$ remains the same as in Section 4.1, but $\hat{\psi}_i^2$ and $\hat{\gamma}$ are modified to be, respectively, the variance of $D_i^2$ and the correlation of $(D_1^2, D_2^2)$ in the sample.
CHAPTER 5

ASSESSING AGREEMENT WITH A GOLD STANDARD
AND SELECTING THE INSTRUMENT
THAT IS CLOSEST TO IT

5.1 Introduction

In the last chapter we discussed the problem of finding the best of two instruments when they are compared with a gold standard. This best instrument, however, is of limited practical utility unless its satisfactory agreement with the gold standard is also established. In this chapter we develop a methodology to first check if there is any instrument that agrees sufficiently well with the gold standard before proceeding to find the best one.

We use the basic notation introduced in Section 4.1. Let $\theta_0$ be a pre-specified small positive real number such that when $0 < \theta_i < \theta_0$, $i = 1, 2$, the $i$-th instrument is considered to have satisfactory agreement with the gold standard. Let $\lambda_0 = \ln(\theta_0)$. Thus $\theta_i < \theta_0$ is equivalent to $\lambda_i < \lambda_0$. Our strategy is to first test the hypothesis

$$H : \lambda_{[1]} \geq \lambda_0, \text{ versus } K : \lambda_{[1]} < \lambda_0.$$  \hspace{1cm} (5.1)

If $H$ is rejected there is evidence that the best instrument agrees sufficiently well with the gold standard, and we select the instrument that produces smaller $\theta$ estimate as

83
the best of the two. However, if $H$ is accepted, we stop inferring that no instrument is good enough. It is desirable to compute the sample size that guarantees

$$Pr(\text{reject } H, \text{ correct selection}) \geq 1 - \beta, \text{ when } \lambda_0 - \lambda_{[1]} \geq \delta, \lambda_{[2]} - \lambda_{[1]} \geq \delta,$$

(5.2)

where the thresholds $\delta$ and $\beta$ are pre-specified. The above condition assures that when the true best instrument has sufficiently close agreement with the gold standard, i.e., $\lambda_0 - \lambda_{[1]} \geq \delta$, and the two competing instruments are not practically equivalent, i.e., $\lambda_{[2]} - \lambda_{[1]} \geq \delta$, then the simultaneous probability of correctly rejecting $H$ and selecting the true best instrument is sufficiently high. Here $\delta$ serves as a threshold for both $\lambda_0 - \lambda_{[1]}$ and $\lambda_{[2]} - \lambda_{[1]}$.

We assume that $\lambda_1 \neq \lambda_2$. The rationality of this assumption was noted in Section 4.3. Let $\Omega_H$ denote the parameter region under the null hypothesis, i.e., $\{ (\mu, \sigma, \rho) : \lambda_{[1]} \geq \lambda_0 \} - \{ (\mu, \sigma, \rho) : \lambda_1 = \lambda_2 \}$, and $\Omega_K$ denote the region under the alternative hypothesis, i.e., $\{ (\mu, \sigma, \rho) : \lambda_{[1]} < \lambda_0 \} - \{ (\mu, \sigma, \rho) : \lambda_1 = \lambda_2 \}$. We denote the best and the worst instruments in the sample by indices (1) and (2), respectively. Hence, for a sample of size $m$, the indices (1) and (2) are such that $\hat{\theta}(1)(m) \leq \hat{\theta}(2)(m)$.

This chapter is organized as follows. Section 5.2 introduces a two-stage decision rule for this comparison problem. We study its small sample properties in Section 5.3. Section 5.4 concludes with a discussion of the results.

5.2 A two-stage procedure

We first suggest a two-stage procedure that satisfies the requirement

$$Pr(\text{reject } H, \text{ correct selection}) \geq 1 - \beta, \text{ when } (\lambda_0 - \lambda_{[1]}, \lambda_{[2]} - \lambda_{[1]}) = (\delta, \delta),$$

84
where the probability is computed using the limit distribution of the statistic involved.

Later we verify using simulation that subject to \( \lambda_0 - \lambda_{[1]} \geq \delta, \lambda_{[2]} - \lambda_{[1]} \geq \delta \) the probability on the LHS of (5.2) is minimized when \( (\lambda_0 - \lambda_{[1]}, \lambda_{[2]} - \lambda_{[1]}) = (\delta, \delta) \) or equivalently \( (\lambda_{[1]}, \lambda_{[2]}) = (\lambda_0 - \delta, \delta) \).

Based on pre-specified \( \alpha, \beta, \lambda_0, \delta \) and \( m \), where \( 0 < \alpha < 1 - \beta < 1, \delta > 0 \), and \( m \geq 2 \); consider the following two-stage procedure.

**PROCEDURE** \( \mathcal{D}(\alpha, \beta, \lambda_0, \delta, m) \):

Stage 1: Take a random sample of size \( m \) and compute the estimates \( \hat{\tau}^2_{[1]}(m) = \hat{\tau}^2_{[1]}(m) + \hat{\tau}^2_{[2]}(m) - 2\hat{\gamma}(m) \hat{\tau}_1(m) \hat{\tau}_2(m), \hat{\tau}^2_{(1)}(m) = \left( \hat{\tau}^2_{[1]}(m) - \hat{\tau}^2_{[2]}(m) \right) I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) + \hat{\tau}^2_{2}(m) \)

and \( \hat{\nu}(m) = (\hat{\tau}_{(1)}(m) - \hat{\gamma}(m) \hat{\tau}_{(2)}(m)) / \hat{\tau}_d(m) \), where the various estimates are defined in Section 4.1. Solve the equation

\[
\Phi_2 \left( \frac{K^{1/2}_m \delta}{\hat{\tau}_{(1)}(m)} - \frac{K^{1/2}_m \delta}{\hat{\tau}_d(m)}; \hat{\nu}(m) \right) = 1 - \beta, \tag{5.3}
\]

for \( K_m \) with \( b_m(\alpha) = (z(\alpha) + t_{m-1}(\alpha))/2 \) and define

\[
N_m = \max \left\{ \left\lfloor K_m \right\rfloor, m \right\} \tag{5.4}
\]

as the second stage sample size.

Stage 2: Take \( N_m - m \) additional i.i.d. observations. Based on this two-stage sample, compute \( \hat{\lambda}_{(1)}(N_m) \) and \( \hat{\tau}_{(1)}(N_m) \). Reject \( H \) if \( \hat{\lambda}_{(1)}(N_m) + N_m^{-1/2} b_m(\alpha) \hat{\tau}_{(1)}(N_m) \leq \lambda_0 \) and accept \( H \) otherwise. Further, if \( H \) is rejected, infer the instrument that produces \( \hat{\lambda}_{(1)}(N_m) \) as the best instrument.

In Section 5.3 we show that \( m = 15 \) is a reasonable choice of first-stage sample size to start the experiment, and using \( b_m(\alpha) \) in (5.3) instead of \( z(\alpha) \) or \( t_{m-1}(\alpha) \) leads to better small sample properties.
Proposition 5.1. (i) Let \((\mu, \sigma, \rho) \in \Omega - \{(\mu, \sigma, \rho): \lambda_1 = \lambda_2\}\) be fixed. Then for the two-stage procedure \(D(\alpha, \beta, \lambda_0, \delta, m)\) described above by (5.3) and (5.4), we have

\[
\lim_{m \to \infty} Pr(\text{reject } H \text{ of (5.1)}) = \begin{cases} 
0, & \text{if } \lambda_{[1]} > \lambda_0; \\
\alpha, & \text{if } \lambda_{[1]} = \lambda_0; \\
1, & \text{if } \lambda_{[1]} < \lambda_0. 
\end{cases}
\]

(ii) Set \(\lambda_{[2]} = \lambda_0\), find \((\mu_0, \sigma_0)\), \(\sigma_0 > 0\), such that \(\ln(\mu_0^2 + \sigma_0^2) = \lambda_0\), and fix \((\mu_{[2]}, \sigma_{[2]}) = (\mu_0, \sigma_0)\). Consider a sequence \((\mu_{[1,m]}, \sigma_{[1,m]}^2)\) of values of \((\mu_{[1]}, \sigma_{[1]}^2) \in \Omega_K\) such that \(\mu_{[1,m]}^2 - \mu_0^2 \asymp m^{-1/2}, \sigma_{[1,m]}^2 - \sigma_0^2 \asymp m^{-1/2}\), and \(\lambda_0 - \lambda_{[1,m]} = \delta_m \sim m^{-1/2} \Delta\), where \(\lambda_{[1,m]} = \ln(\mu_{[1,m]}^2 + \sigma_{[1,m]}^2)\) and \(\Delta > 0\). Then for the procedure \(D(\alpha, \beta, \lambda_0, \delta_m, m)\) we have

\[
\lim_{m \to \infty} Pr(\text{reject } H \text{ of (5.1), correct selection } | \lambda_{[1,m]} = \lambda_0 - \delta_m) \in [1 - \beta, 1),
\]

when \(\lambda_{[2]} = \lambda_0\).

Part (i) of the above result shows that the test of hypotheses (5.1) has asymptotic level \(\alpha\) and part (ii) establishes the validity of the sample size formula given in (5.4). In part (ii) the mean and the standard deviation of the worse instrument \(D_{[2]}\), \((\mu_{[2]}, \sigma_{[2]}^2)\), are held fixed at \((\mu_0, \sigma_0)\) ensuring \(\lambda_{[2]} = \lambda_0\). Then sequences \((\mu_{[1,m]}, \sigma_{[1,m]}^2) \in \Omega_K\) of mean and standard deviation of the best instrument \(D_{[1]}\) are chosen such that they converge to \((\mu_0, \sigma_0)\) as \(m \to \infty\). This way \(\lambda_{[1,m]} \to \lambda_0\) and \(Pr(\text{reject } H, \text{ correct selection})\) goes to a limit that is no less than \(1 - \beta\). So, in a sense as \(m\) keeps increasing, \(\lambda_{[1,m]}\) comes closer to the cutoff \(\lambda_0\) making it more difficult for the test to reject \(H\), and simultaneously the two instruments come closer making the correct selection more difficult as well. Such a limiting argument is usually the basis of sample size computation for large sample tests (see Lehman, 1999, sec 3.3).
For a fixed \( m \), the five parameters of \((D^{12}, D^{22})\) are \((\mu^{12}_m, \mu_0, \sigma^{12}_m, \sigma_0, \rho)\) and the parameters of \((D^{12}, D^{22})\) are \((\theta^{12}_m, \theta_0, \psi^{12}_m, \psi_0, \gamma_m)\). In addition, let \( \psi^2_{d_m}, \tau^2_{[1]_m}, \tau^2_{0}\) and \( \tau^2_{d_m} \) be the sequences associated with \( \psi^2_{d}, \tau^2_{[1]}, \tau^2_{[2]} \) and \( \tau^2_{d} \), respectively. These quantities are obtained by substituting \((\mu^{12}_m, \mu_0, \sigma^{12}_m, \sigma_0, \rho)\) for \((\mu^{12}, \mu_0, \sigma^{12}, \sigma_0, \rho)\).

**Proof of Proposition 5.1:** (i) Consider a fixed \((\mu, \sigma, \rho) \in \Omega - \{(\mu, \sigma, \rho) : \lambda_1 = \lambda_2\}\). For given values of \( \alpha, \beta, \delta, \lambda_0 \) and \( m \), we have by construction

\[
Pr(\text{reject } H) = Pr(\hat{\lambda}_{(1)}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{(1)}(N_m) \leq \lambda_0).
\] (5.5)

First we establish that \( N_m/m \xrightarrow{P} 1 \) as \( m \to \infty \) so that Lemma A.3 is applicable. Upon using the weak law of large numbers we can conclude that \( \hat{\tau}^2_{d}(m) \xrightarrow{P} \tau^2_{d} \) and \( \hat{\gamma}(m) \xrightarrow{P} \gamma \). Further, an application of Lemma A.4 (a) shows \( \hat{\tau}^2_{(1)}(m) \xrightarrow{P} \tau^2_{[1]} \). Hence

\[
\hat{\nu}(m) \xrightarrow{P} \nu = (\tau_{[1]} - \gamma \tau_{[2]})/\tau_d, \quad \text{where} \quad \hat{\nu}(m) = \left( \hat{\tau}_{(1)}(m) - \hat{\gamma}(m) \hat{\tau}_{(2)}(m) \right)/\hat{\tau}_d(m).
\] So, on taking limit \( m \to \infty \) on both sides of (5.3) we conclude that \( K_m \) converges in probability to a finite positive real number. Hence, from the definition of \( N_m \) in (5.4) it follows that \( N_m/m \xrightarrow{P} 1 \).

As a consequence of part (a) of Lemma A.3, \( \hat{\theta}_{(1)}(N_m) = \min \{ \hat{\theta}_1(N_m), \hat{\theta}_2(N_m) \} \xrightarrow{P} \min \{ \theta_1, \theta_2 \} = \theta_{[1]} \), which implies that \( \hat{\lambda}_{(1)}(N_m) \xrightarrow{P} \lambda_{[1]} \). Moreover, since \( \lambda_1 \neq \lambda_2 \), we have \( \hat{\tau}_{(1)}(N_m) \xrightarrow{P} \tau_{[1]} \) from Lemma A.4 (b). Further, \( b_m(\alpha) \to z(\alpha) \) as \( m \to \infty \). It then follows that \( \hat{\lambda}_{(1)}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{(1)}(N_m) \xrightarrow{P} \lambda_{[1]} \). Hence on taking limit on both sides of (5.5) we get,

\[
\lim_{m \to \infty} Pr(\text{Reject } H) = \lim_{m \to \infty} Pr(\hat{\lambda}_{(1)}(N_m) \leq \lambda_0) = \begin{cases} 
0, & \text{if } \lambda_{[1]} > \lambda_0; \\
1, & \text{if } \lambda_{[1]} < \lambda_0.
\end{cases}
\]

87
Now it just remains to show that the limit of \( Pr(\text{Reject } H) \) is \( \alpha \) when \( \lambda_{[1]} = \lambda_0 \).

Rewrite (5.5) as

\[
Pr(\text{Reject } H) = Pr(\text{Reject } H, \hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m)) \\
+ Pr(\text{Reject } H, \hat{\lambda}_{[1]}(N_m) > \hat{\lambda}_{[2]}(N_m)).
\]

(5.6)

Note that \( \lim_{m \to \infty} Pr(\hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m)) = 1 \) since \( \lambda_{[1]} < \lambda_{[2]} \). Consequently, the second term on the RHS of (5.6) goes to zero and

\[
\lim_{m \to \infty} Pr(\text{Reject } H) = \lim_{m \to \infty} Pr(\text{Reject } H, \hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m)) \\
= \lim_{m \to \infty} Pr(\hat{\lambda}_{[1]}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{[1]}(N_m) \leq \lambda_0, \hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m)) \\
= \lim_{m \to \infty} Pr(\hat{\lambda}_{[1]}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{[1]}(N_m) \leq \lambda_0) \\
= \lim_{m \to \infty} Pr\left( N^{-1/2}_m (\hat{\lambda}_{[1]}(N_m) - \lambda_{[1]}) / \hat{\tau}_{[1]}(N_m) \leq -b_m(\alpha) \right),
\]

(5.7)

when \( \lambda_{[1]} = \lambda_0 \). Now \( N^{-1/2}_m (\hat{\lambda}_{[1]}(N_m) - \lambda_{[1]}) / \hat{\tau}_{[1]}(N_m) \) \( \xrightarrow{p} \) \( N(0,1) \) from the part (b) of Lemma A.3. Hence the limit in (5.7) equals 1 - \( \alpha \) because \( b_m(\alpha) \to z(\alpha) \).

(ii) Fix \( m \geq 2 \) so that \( (\mu_{[1]_m}, \mu_0, \sigma_{[1]_m}, \sigma_0, \rho) \in \Omega_K \) is fixed. By construction \( Pr(\text{reject } H, \text{correct selection}) \)

\[
= Pr\left( \hat{\lambda}_{[1]}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{[1]}(N_m) \leq \lambda_0, \hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m) \right) \\
= Pr\left( \hat{\lambda}_{[1]}(N_m) + N^{-1/2}_m b_m(\alpha) \hat{\tau}_{[1]}(N_m) \leq \lambda_0, \hat{\lambda}_{[1]}(N_m) \leq \hat{\lambda}_{[2]}(N_m) \right) \\
= Pr\left( \frac{N^{-1/2}_m}{\hat{\tau}_{[1]}(N_m)} (\hat{\lambda}_{[1]}(N_m) - \lambda_{[1]_m}) \leq \frac{N^{-1/2}_m \delta_m}{\hat{\tau}_{[1]}(N_m)} - b_m(\alpha), \right) \\
\frac{N^{-1/2}_m}{\hat{\tau}_{d}(N_m)} (\hat{\lambda}_{[1]}(N_m) - \hat{\lambda}_{[2]}(N_m) - \lambda_{[1]_m} + \lambda_0) \leq \frac{N^{-1/2}_m \delta_m}{\hat{\tau}_{d}(N_m)} \right),
\]

(5.8)

since \( \lambda_0 - \lambda_{[1]_m} = \delta_m = \Delta/m^{1/2} \). We now proceed in two steps. Step 1 establishes \( N_m/m \xrightarrow{p} 1 \) and in step 2 we use part (b) of Lemma A.6 to obtain the joint limiting
distribution of the random variables on the LHS of the inequalities in (5.8).

Step 1: From Lemma A.6 (a), \( \hat{\tau}(m) \xrightarrow{p} \tau_0^2 \), as \( m \to \infty \). Further, \( \hat{\tau}_d^2(m) \xrightarrow{p} \tau_{d_0}^2 \), \( \hat{\gamma}(m) \xrightarrow{p} \gamma_0 \) and hence \( \hat{\nu}(m) \xrightarrow{p} \nu_0 = \tau_0(1 - \gamma_0)/\tau_{d_0} \). Let \( K'_m \) be the solution of (5.3) with \( \delta_m \) in place of \( \delta \), that is

\[
\Phi_2 \left( \frac{K'_m^{1/2} \delta_m}{\hat{\tau}(m)} - b_m(\alpha), \frac{K'_m^{1/2} \delta_m}{\hat{\tau}_d(m)}; \hat{\nu}(m) \right) = 1 - \beta.
\]

Since \( 0 < \alpha < 1 - \beta < 1 \), it follows that \( K'_m/m \xrightarrow{p} k' \), where \( k' \) is such that

\[
\Phi_2 \left( \frac{k^{1/2} \Delta}{\tau_0} - z(\alpha), \frac{k^{1/2} \Delta}{\tau_{d_0}}; \nu_0 \right) = 1 - \beta. \tag{5.9}
\]

Hence from (5.4), \( N_m/m \xrightarrow{p} \max\{k',1\} \), a finite positive real number.

Step 2: We apply part (b) of Lemma A.5 (a) and use delta method to conclude that the joint limiting distribution of the random variables

\[
\frac{N_m^{1/2}}{\hat{\tau}[m](N_m)} \left( \hat{\lambda}[m](N_m) - \lambda[m] \right), \frac{N_m^{1/2}}{\hat{\tau}_d(N_m)} \left( \hat{\lambda}[m](N_m) - \hat{\lambda}[m](N_m) - \lambda[m] + \lambda_0 \right)
\]

is \( BVN(0,0,1,1,\nu_0) \). Moreover, as \( m \to \infty \),

\[
\frac{N_m^{1/2} \delta_m}{\hat{\tau}[m](N_m)} - b_m(\alpha) \xrightarrow{p} \max \left\{ \frac{k \Delta}{\tau_0}, \frac{\Delta}{\tau_0} \right\} - b_m(\alpha),
\]

and \( \frac{N_m^{1/2} \delta_m}{\hat{\tau}_d(N_m)} \xrightarrow{p} \max \left\{ \frac{k \Delta}{\tau_{d_0}}, \frac{\Delta}{\tau_{d_0}} \right\} \).

Hence as \( m \to \infty \), the probability in (5.8) converges to a limit that is no less than the LHS of (5.9). Thus the result is established.

5.3 Small sample properties

In this section we examine the small sample performance of the proposed procedure \( \mathcal{D} \) using simulation. In particular, we choose \( m = 15 \) and investigate the following:
1. $Pr(\text{reject } H)$ when $\lambda_{[1]} = \lambda_0$ and the magnitude of $E(N_{15})$.

2. $Pr(\text{reject } H, \text{ correct selection})$ when $\lambda_0 - \lambda_{[1]} = \delta, \lambda_{[2]} - \lambda_{[1]} = \delta$ and the magnitude of $E(N_{15})$.

3. Whether $Pr(\text{reject } H, \text{ correct selection})$ is minimized at $(\lambda_0 - \lambda_{[1]}, \lambda_{[2]} - \lambda_{[1]}) = (\delta, \delta)$ subject to $\lambda_0 - \lambda_{[1]} \geq \delta, \lambda_{[2]} - \lambda_{[1]} \geq \delta$.

4. Adequacy of $m = 15$ as the first-stage sample size.

In applications usually one chooses $5\%$ as the level of a test, $80\%$ as the cutoff power for sample size computation (see e.g., Rössner, 2000, sec. 7.6), and $20\%$ relative difference as the threshold for practical equivalence (e.g., bioequivalence studies). So, for the simulation study we choose $\alpha = 0.05, \beta = 0.2, \delta = \ln(1.2) = \delta_0$ (say). In addition, we take $\theta_0 = 1$ or equivalently $\lambda_0 = 0$ as the threshold for satisfactory agreement between an instrument and the gold standard. There is no loss of generality in assuming $\theta_0 = 1$ since it amounts to re-scaling all the observations on $(D_1, D_2)$ by $\theta_0^{-1/2}$. Thus $D(0.05, 0.2, 0.0, \delta_0, 15)$, say $D_0$, is the specific procedure we are studying. Further, we assume that the first instrument is the best one. Hence $\theta_1 < \theta_2$ and the indices $(\{1\}, \{2\}) = (1, 2)$. We use the grid of $(\mu_1, \mu_2)$ values, $\Omega_i(\theta_1, \theta_2), i = 1, 2, 3$, introduced in (4.1).

We now study $Pr(\text{reject } H)$ when $\lambda_1 = \lambda_0 = 0$ or $\theta_1 = \theta_2 = 1$. For $\theta_2$ we first take $\theta_2 = 1.2$ and later discuss the effect of moving $\theta_2$ away from $\theta_1$.

The top of Figure 5.1 contains the plot of empirical type-I error probability of the test computed using 50,000 replications of the procedure $D_0$ for $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$. The five-number summary for these 231 values is (0.034, 0.045, 0.046, 0.047, 0.061). Only 3\% of them are more than 0.05 and they fall in the region $\{\mu_1 = -0.999, \mu_2 =$
\[-1.094 \leq \mu_2 \leq 0.4, 1.094\]}, which is a subset of \(\Omega_2(1.0, 1.2)\). Further, 21\% of the values are less than 0.045 and most of them also belong to \(\Omega_2(1.0, 1.2)\). Thus the parameter points where the test is liberal or overly conservative are those where \(\sigma_1\) or \(\sigma_2\) are close to zero. Fortunately, these regions are not of much practical interest as they refer to situations when one or both the instruments agree almost perfectly with the gold standard except for a constant bias. The five-number summary of the 190 values in \(\Omega_3(1.0, 1.2)\) is (0.042, 0.046, 0.046, 0.047, 0.049).

The bottom of Figure 5.1 has the surface plot of the corresponding average two-stage sample sizes, which will now be referred as average sample number (ASN). The five-number summary of these 231 ASN’s is (15, 264.1, 333, 372.5, 389.9). As expected, the minimum occurs when \(\{\mu_1 = -0.999, \mu_2 = \pm 1.094\}\), i.e., when both \(\sigma_1\) and \(\sigma_2\) are the smallest; and the maximum occurs when \(\{\mu_1 = 0 = \mu_2\}\), i.e., when both \(\sigma_1\) and \(\sigma_2\) are the largest. For a fixed \(\mu_2\), the ASN increases as \(\mu_1\) increases. On the other hand, for a fixed \(\mu_1\), it first increases with \(\mu_2\) and then starts decreasing. But the point where it starts decreasing varies with \(\mu_1\). The five-number summary for the 190 ASN’s in \(\Omega_3(1.0, 1.2)\) is (136.7, 299.3, 348.2, 376.9, 389.9).

Since the region where \(\sigma_1\) or \(\sigma_2\) are close to zero is not of practical interest only the results belonging to \(\Omega_3(\theta_1, \theta_2)\) will now be presented.

It may be noted that if we had used \(z(0.05)\) in place of \(b_{14}(0.05) = (z(0.05) + t_{14}(0.05))/2\) in (5.3) the test would have been liberal and if we had used \(t_{14}(0.05)\) it would have been more conservative. The difference between the empirical type-I error probability of the procedure with \(z(0.05)\) in (5.3) and the nominal level 0.05 is plotted in Figure 5.2 for \((\mu_1, \mu_2) \in \Omega_3(1.0, 1.2)\). The five-number summary for these differences is (-0.002, 0.001, 0.002, 0.003, 0.005) with roughly 89\% of the differences
Figure 5.1: Empirical type-I error probability for testing hypotheses (5.1) using the procedure $D_0$ over the grid $(\mu_1, \mu_2) \in \Omega_1(1.0, 1.2)$ (top) and the associated ASN (bottom).
being positive. This summary for the associated ASN’s is (130.5, 286.1, 333.5, 360.6, 373.0). The reduction in ASN is expected since \( z(0.05) < b(0.05) \).

As \( \theta_2 \) is increased from \( \theta_1 = 1.0 \), the empirical type-I error probabilities come a little closer to 0.05. The five-number summaries for these and the associated ASN’s over the grid are (0.043, 0.045, 0.046, 0.047, 0.05) and (135.9, 306.5, 350.7, 377.6, 389.8), respectively. In addition when \( \theta_2 = 2.0 \), these summaries are (0.043, 0.046, 0.046, 0.047, 0.049) and (135.5, 300.1, 361.9, 381.9, 389.9). Thus it appears that as \( \theta_2 \) moves away from \( \theta_1 \) the procedure becomes less conservative without considerable effect on the ASN’s.
Next, we study \( Pr(\text{reject } H, \text{ correct selection}) \) when \( \lambda_0 - \lambda_1 = -\lambda_1 = \delta_0 \) and \( \lambda_2 - \lambda_1 = \ln(1.2) \) or equivalently \((\theta_1, \theta_2) = (1/1.2, 1.0)\). The empirical probabilities that \( \mathcal{D}_0 \) rejects \( H \) and simultaneously selects the first instrument are presented in Figure 5.3 (top) for \((\mu_1, \mu_2) \in \Omega_3(1/1.2, 1.0)\). These empirical values are computed using 50,000 repetitions of the procedure, and their five-number summary is \((0.659, 0.770, 0.796, 0.801, 0.879)\). For every fixed \( \mu_2 \), the values are increasing functions of \( \mu_1 \). On the other hand, for every fixed \( \mu_1 \), the values increase with \( \mu_2 \) and then start decreasing. Figure 5.3 (bottom) has the surface plot for the associated ASN’s. The five-number summary of these averages is \((136.5, 294.4, 348.3, 376.9, 389.9)\) and their overall pattern remains similar to Figure 5.1.

Our next task is to verify if \( Pr(\text{reject } H, \text{ correct selection}) \) for the procedure \( \mathcal{D}_0 \) is minimized at \((-\lambda_1, \lambda_2 - \lambda_1) = (\delta_0, \delta_0) \) provided \(-\lambda_1 \geq \delta_0, \lambda_2 - \lambda_1 \geq \delta_0 \). The above process of computing empirical probabilities when \((\lambda_1, \lambda_2) = (-\delta_0, 0) = \zeta_0 \) (say) is repeated for four values of \((\lambda_1, \lambda_2)\): \( \zeta_1 = (-\delta_0, 0.5\delta_0) \), \( \zeta_2 = (-1.5\delta_0, -0.5\delta_0) \), \( \zeta_3 = (-1.5\delta_0, 0) \) and \( \zeta_4 = (-\delta_0, 0.5\delta_0) \). The points \( \zeta_0 \), \( \zeta_1 \) and \( \zeta_2 \) lie on the boundary of the region \( \{-\lambda_1 \geq \delta_0, \lambda_2 - \lambda_1 \geq \delta_0\} \), whereas \( \zeta_3 \) and \( \zeta_4 \) lie in its interior with \( \zeta_3 \) being closer to the boundary than \( \zeta_4 \). The locations of these points are illustrated in Figure 5.4.

The five-number summaries of the 190 empirical probabilities and ASN’s when \((\mu_1, \mu_2) \in \Omega_3(\theta_1, \theta_2)\) for the various values of \((\lambda_1, \lambda_2)\) are presented in Table 5.1. Since the quantiles of the empirical probability related to \( \zeta_0 \) are the least, the above assertion is verified. It may be noted that when \( \lambda_1 \) is fixed and \( \lambda_2 \) is increased, all the quantiles of empirical probability also increase. The same is true when \( \lambda_2 \) is fixed and \( \lambda_1 \) is decreased, however the increase in this case is more substantial. Although
Figure 5.3: Empirical probability that the procedure $\mathcal{D}_0$ rejects $H$, given in (5.1), and simultaneously selects the first instrument (top) and the associated ASN (bottom), over the grid $(\mu_1, \mu_2) \in \Omega_3(1/1.2, 1.0)$. 
Figure 5.4: Location of the points $\zeta_0,\ldots,\zeta_4$ in the region $\{\lambda_0 - \lambda_1 \geq \delta_0, \lambda_2 - \lambda_1 \geq \delta_0\}$ with $\lambda_0 = 0$ and $\delta_0 = \ln(1.2)$. 
there is no considerable difference in the overall distributions of ASN’s, the quantiles of ASN for $\zeta_0$ and $\zeta_2$, and those for $\zeta_1$ and $\zeta_3$ are roughly the same. This implies that when $\lambda_2 - \lambda_1$ is held fixed, any increase in $\lambda_0 - \lambda_1$ has little bearing on the distribution of ASN’s.

Overall, the choice of $m = 15$ appears to be reasonable unless $\sigma_1$ or $\sigma_2$ is close to zero. Further, as one would expect, when $m$ is increased from 15 to 30, the empirical type-I error probability of the procedure $\mathcal{D}(0.05, 0.2, 0.0, \delta_0, 30)$ comes closer to 0.05. With $m = 30$, the five-number summary for the 190 empirical probabilities over the grid $\Omega_3(1.0, 1.2)$ is (0.045, 0.049, 0.049, 0.05, 0.053). Here 13% of the values are more than 0.05 with only 1% more than 0.051. The summary for associated ASN’s is (133.0, 303.0, 346.3, 374.2, 383.9). Further, this summary for the empirical probabilities that
\[ \begin{array}{cccc}
\delta & 1 - \beta \\
\ln(1.1) & 0.70 & 0.80 & 0.90 \\
\ln(1.2) & 986 & 1283 & 1762 \\
\ln(1.3) & 270 & 351 & 482 \\
\ln(1.4) & 131 & 170 & 233 \\
\end{array} \]

Table 5.2: Two-stage sample sizes required by the procedure \( D(0.05, \beta, \lambda_0, \delta, 15) \) to satisfy the probability requirement (5.2) approximately.

the procedure rejects \( H \) and simultaneously selects the first instrument over the grid \( \Omega_3(1/1.2, 1.0) \) is \((0.681, 0.779, 0.795, 0.813, 0.879)\), and for the associated ASN’s it is \((133.4, 303.3, 346.3, 374.1, 384)\). These probabilities are higher than those with \( m = 15 \) and usually the ASN’s are a bit lower. This is a desirable feature as the procedure performs better with a larger first-stage sample size and it comes at a reduced cost in terms of a little lower average second-stage sample size.

**EXAMPLE 4.1 (continued):** We take the 15 observations on \((D_1, D_2)\) in the fractional area change data as the first-stage sample. The parameter estimates based on this sample were given on page 75. We compute the second-stage sample size \( N_{15} \) using (5.4) such that the procedure \( D(0.05, \beta, \lambda_0, \delta, 15) \) satisfies the probability requirement (5.2) approximately. These values are presented in Table 5.2 for \( \delta = \ln(1.1), \ln(1.2), \ln(1.3) \) and \( \beta = 0.10, 0.20, 0.30 \). Now one can use the decision rule given on page 85 for a given \( \lambda_0 \) after taking the second-stage sample. □

### 5.4 Discussion

In this chapter we considered an extension of the problem of selecting the best of two instruments when compared with a gold standard. Before the selection, we
assessed using a test of hypotheses whether the agreement between the best and the
gold standard was satisfactory. Only when we inferred satisfactory agreement, we
proceeded to the selection step. We described a large sample two-stage procedure
for this problem and studied its small sample properties. We found that 15 is a
reasonable choice for the first stage sample size.

It is natural to think that if we forego the testing step and proceed directly to
the selection step, the two-stage procedure \( \mathcal{D}(\alpha, \beta, \lambda_0, \delta, m) \) presented here should
reduce to the one presented in Section 4.4. It happens if we take \( \alpha = 1 \), because with
this choice of \( \alpha \), the procedure \( \mathcal{D}(1, \beta, \lambda_0, \delta, m) \) would always reject \( H \) irrespective
of the value of \( \lambda_0 \). Consequently, the formula (5.4) for the sample size \( N_m \) becomes
\[
\max \left\{ \frac{z^2(\beta) \hat{\tau}_d^2(m)}{\delta^2}, m \right\}.
\]
Thus, if we replace \( z(\beta) \) by \( t_{m-1}(\beta) \) in this formula, it turns
into the sample size formula (4.6) for the two-stage selection procedure introduced in
Section 4.4. In that section, we had used the \( t \)-quantile as it had led to better small
sample properties.

The bivariate normal assumption for \( (D_1, D_2) \) can be weakened. We only need
that the 8-th order moments of \( (D_1, D_2) \) be finite. In addition, when they depend
on \( m \), we assume that they remain finite as \( m \to \infty \) and the limiting correlation of
\( (D_1^2, D_2^2) \) is away from \( \pm 1 \). In this setup, \( \hat{\theta}_i \) remains the same as in Section 4.1, but
\( \hat{\psi}_i^2 \) and \( \hat{\gamma} \) are modified to be the variance of \( D_i^2 \) and the correlation of \( (D_1^2, D_2^2) \) in the
sample, respectively. The proof of Proposition 5.1 can be modified accordingly.
CHAPTER 6

SUMMARY AND FUTURE WORK

In this dissertation, we introduced statistical methodologies to handle several problems arising in method comparison studies. We discussed the basic issue of assessing satisfactory agreement between two instruments in Chapters 2 and 3. In Chapter 2 we presented a review of existing approaches and noted their limitations. They provided the motivation for performing inference on individual parameters of the underlying distribution. We then combined these separate inferences with intersection-union principle to make an overall statement about the degree of evidence in the data for satisfactory agreement. In future we will explore distribution-free procedures for inference on individual parameters. They will be helpful when the normality assumption for the data cannot be justified.

In Chapter 3 we described the assessment of agreement through the coverage probability of a threshold interval. There we established connection between various tests and tolerance intervals. We also remarked on the similarity of this approach and the technique of acceptance sampling. Future work involves adapting the procedures from acceptance sampling and comparing them with the existing approaches. The hope is to find a test that has good small sample properties in the sense of having
type-I error rate close to the nominal level and that produces a thin and steep power curve (see Lei and Vardeman, 1998).

In chapters 4 and 5 we discussed the comparison of two instruments with a gold standard to find the best of the two. We presented single-stage and two-stage procedures for selecting the best instrument in Chapter 4. In Chapter 5 we introduced a two-stage procedure for determining beforehand if the best instrument is good enough for practical purposes. In both the chapters we commented on the adequacy of 15 as the sample size for a single-stage and first-stage sample size for a two-stage procedure. We plan to extend these procedures to handle the comparison of more than two instruments with a gold standard.

There are several important problems in this area that warrant attention. Of particular interest are the following:

1. Extending the procedures to handle replicate measurements on every subject by every instrument.

2. Investigating robustness of the procedures to deviations from the assumed normality.

3. Taking into account situations when parameters cannot be assumed to remain fixed over the entire range of measurement.

4. Comparison of instruments with a gold standard in terms of coverage probability of a threshold interval.

5. Incorporating covariate information in comparison.

6. Comparison with respect to multivariate measurements.
APPENDIX A

TECHNICAL LEMMAS

Here we collect the lemmas that were used to prove various results.

A lemma for Proposition 2.3:

Lemma A.1. For any $c < 1$,

$\lim_{\rho \to 1^+} P_{\rho}(\hat{\rho} > c) = 1,$

and $\lim_{\rho \to 1^-} P_{\rho}(\hat{\rho} < -c) = 1.$

Proof: (a) holds if $\lim_{\rho \to 1^-} P_{\rho}(\hat{\rho} \leq c) = 0$. The density function of $\hat{\rho}$ is,

$$f_{\hat{\rho}}(r, \rho) = \frac{n - 2}{\pi} (1 - \rho^2)^{(n-1)/2}(1 - r^2)^{(n-4)/4}$$

$$\times \int_0^1 \frac{t^{n-2}}{(1 - \rho rt)^{n-1} \sqrt{1 - t^2}} dt; \quad |r| \leq 1, |\rho| < 1,$$

and the factor $(1 - \rho^2)^{(n-1)/2}$ approaches 0 as $\rho \to 1^-$. Thus $P_{\rho}(\hat{\rho} \leq c) \to 0$ as $\rho \to -1$ if

$$\lim_{\rho \to 1^-} \int_{-1}^c (1 - r^2)^{(n-4)/4} \int_0^1 \frac{t^{n-2}}{(1 - \rho rt)^{n-1} \sqrt{1 - t^2}} dt dr < \infty.$$

Note that for $\rho > 0$, the inside integrand is an increasing function of $r$ and is non-negative. Hence for $\rho > 0$, the above integrand is bounded by

$$\int_{-1}^c (1 - r^2)^{(n-4)/4} dr \int_0^1 \frac{t^{n-2}}{(1 - \rho ct)^{n-1} \sqrt{1 - t^2}} dt.$$
Here the first integral is bounded by $1 + c$. The integrand of the second integral is an increasing function of $\rho$ for $\rho > 0$. Hence on taking limit $\rho \to 1-$ and using the monotone convergence theorem, we conclude that as $\rho \to 1-$, the second integral converges to
\[
\int_0^1 \frac{t^{n-2}}{(1 - ct)^{n-1}\sqrt{1 - t^2}} dt < \frac{1}{(1 - c)^{n-1}} \int_0^1 \frac{1}{\sqrt{1 - t^2}} dt < \infty,
\]
since $c < 1$. Thus (a) holds.

(b) Note that $Pr_{-\rho}(\hat{\rho} < -c) \equiv Pr_{-\rho}(-\hat{\rho} > c) = Pr_{\rho}(\hat{\rho} > c)$, since $f_{\rho}(r, \rho) = f_{-\rho}(r, -\rho)$. Hence the result follows from (a). □

**Lemmas for Propositions 4.2 and 5.1:**

**Lemma A.2.** For a fixed $m$, let $Z_1, Z_2, \ldots, Z_{N_m}$ be a sequence of i.i.d. random variables with mean $E_m(Z_1)$ and variance $Var_m(Z_1)$. Assume that $\lim_{m \to \infty} Var_m(Z_1) < \infty$. Define $S(m) = \sum_{i=1}^{m} (Z_i - E_m(Z_1))$. Then if $N_m/m \overset{p}{\to} c$ as $m \to \infty$, we have $S(N_m)/N_m \overset{p}{\to} 0$, where $c$ is a finite positive real number.

**Proof:** Let $c_m = \lceil cm \rceil$. Then $N_m/c_m \overset{p}{\to} 1$ as $m \to \infty$ because $[c_m]/(cm) \to 1$. Hence it suffices to establish that $S(N_m)/c_m \overset{p}{\to} 0$. We can write
\[
\frac{S(N_m)}{c_m} = \frac{S(c_m)}{c_m} + \frac{S(N_m) - S(c_m)}{c_m}.
\]
The first term on the RHS above converges to zero in probability from the weak law of large numbers. We now show that the second term also converges to zero in probability.

Let $\epsilon_1 > 0$ and $\epsilon_2 \in (0, 1]$. Then for a fixed $m$, we have
\[
Pr \left( |S(N_m) - S(c_m)| > \epsilon_1 c_m \right) = Pr \left( |S(N_m) - S(c_m)| > \epsilon_1 c_m, |c_m^{-1} N_m - 1| \leq \epsilon_2 \right) + Pr \left( |S(N_m) - S(c_m)| > \epsilon_1 c_m, |c_m^{-1} N_m - 1| > \epsilon_2 \right).
\]

103
The first term on the RHS is bounded above by \((2\varepsilon_2 Var_m(Z_1))/(c_m\varepsilon_1^2)\) from Kolmogorov’s inequality (see Billingsley, 1995, p. 287). Since \(\lim_{m \to \infty} Var_m(Z_1) < \infty\), this term approaches zero as \(m \to \infty\). The second term also goes to zero since \(N_m/c_m \xrightarrow{p} 1\). \(\square\)

**Lemma A.3.** Let \(N_m\) be a sequence of positive random integers such that \(N_m/m \xrightarrow{p} c\) as \(m \to \infty\), where \(c\) is a finite positive real number. Then for any fixed \((\mu, \sigma, \rho)\) in the parameter space \(\Omega\),

(a) \(\hat{\theta}_i(N_m), \hat{\psi}_i^2(N_m)\) and \(\hat{\gamma}(N_m)\) converge in probability to \(\theta_i, \psi_i^2\) and \(\gamma\), respectively, for \(i = 1, 2\).

(b) \(\left(\frac{N_m^{1/2}(\hat{\lambda}_1(N_m) - \lambda_1)}{\hat{\tau}_1(N_m)}, \frac{N_m^{1/2}(\hat{\lambda}_2(N_m) - \lambda_2)}{\hat{\tau}_2(N_m)}\right) \xrightarrow{L} BVN(0, 0, 1, 1, \gamma)\).

**Proof:** (a) Since \(\hat{\theta}_i(N_m), \hat{\psi}_i^2(N_m)\) and \(\hat{\gamma}(N_m)\) are continuous functions of \(\hat{\mu}_i(N_m), \hat{\sigma}_i^2(N_m)\) and \(\hat{\rho}(N_m)\), it is enough to show that the latter sequences converge in probability to their respective parameters.

First consider \(\hat{\mu}_i(N_m) - \mu_i\). We can write it as \(N_m^{-1} \sum_{j=1}^{N_m} (D_{ij} - E_m(D_i))\), where \(E_m(D_i) = \mu_i\). Also, \(Var(D_i) = \sigma_i^2 = Var_m(D_i)\). Now it follows from Lemma A.2 that \(\hat{\mu}_i(N_m) - \mu_i \xrightarrow{p} 0\).

Since \((D_1, D_2)\) has a bivariate normal distribution, \(Var(D_1^2)\) and \(Var(D_1D_2)\) are also finite. Hence upon proceeding in a similar fashion we can establish the convergence for the other two sequences.

(b) Note that \(\hat{\theta}_i(N_m) = \left(\frac{\sum_{j=1}^{N_m} D_{ij}^2 - \hat{\mu}_i^2(N_m)}{N_m - 1}\right)\). Since \(\hat{\mu}_i(N_m) \xrightarrow{p} \mu_i\), it follows from the Slutsky’s theorem that the limiting distributions of \(\left(\frac{N_m^{1/2}(\hat{\theta}_1(N_m) - \theta_1)}{\sqrt{\sum_{j=1}^{N_m} D_{ij}^2 - \hat{\theta}_1^2(N_m)}}, \frac{N_m^{1/2}(\hat{\theta}_2(N_m) - \theta_2)}{\sqrt{\sum_{j=1}^{N_m} D_{ij}^2 - \hat{\theta}_2^2(N_m)}}\right)\) and \(\left(\frac{N_m^{-1/2} \sum_{j=1}^{N_m} D_{ij}^2}{N_m^{-1/2} \sum_{j=1}^{N_m} (D_{ij}^2 - \theta_1)}, \frac{N_m^{-1/2} \sum_{j=1}^{N_m} D_{ij}^2}{N_m^{-1/2} \sum_{j=1}^{N_m} (D_{ij}^2 - \theta_2)}\right)\) are the

104
same. The bivariate version of the random index central limit theorem (see Billingsley, 1995, p. 369 and Lehmann, 1999, p. 295) establishes that the limiting distribution of the latter is bivariate normal with parameters \((0,0,\psi_1,\psi_2,\gamma)\). Further, an application of the delta method (see Lehmann, 1999, p. 284) shows that \((N_m^{1/2}(\hat{\lambda}_1(N_m) - \lambda_1), N_m^{1/2}(\hat{\lambda}_2(N_m) - \lambda_2)) \overset{d}{\to} \text{BVN}(0,0,\tau_1,\tau_2,\gamma)\). The result now follows from Slutsky’s theorem since \(\hat{\tau}_i^2(N_m) \overset{p}{\to} \tau_i^2\), for \(i = 1, 2\), using part (a). \(\Box\)

**Lemma A.4.** For any fixed \((\mu, \sigma, \rho) \in \Omega - \{(\mu, \sigma, \rho) : \theta_1 = \theta_2\}\), as \(m \to \infty\),

(a) \(\hat{\tau}^2_{(1)}(m) \overset{p}{\to} \tau^2_{[1]}\), and

(b) \(\hat{\tau}^2_{(1)}(N_m) \overset{p}{\to} \tau^2_{[1]}\), where \(N_m\) is a sequence of positive random integers such that \(N_m / m \overset{p}{\to} c; c\) is a finite positive real number.

**Proof:** (a) Since \(\hat{\tau}^2_{(1)}(m) = \hat{\psi}^2_{(1)}(m) / \hat{\theta}^2_{(1)}(m)\), it suffices to establish that \(\hat{\theta}_{(1)}(m) \overset{p}{\to} \theta_{[1]}\) and \(\hat{\psi}^2_{(1)}(m) \overset{p}{\to} \psi^2_{[1]}\) as \(m \to \infty\). First note that \(\hat{\mu}_i(m) \overset{p}{\to} \mu_i\), \(\hat{\sigma}_i^2(m) \overset{p}{\to} \sigma_i^2\) and \(\hat{\rho}(m) \overset{p}{\to} \rho\) from the weak law of large numbers. Consequently, \(\hat{\theta}_i(m) \overset{p}{\to} \theta_i\), \(\hat{\psi}^2_i(m) \overset{p}{\to} \psi^2_i\) and \(\hat{\theta}_{(1)}(m) = \min \{\hat{\theta}_1(m), \hat{\theta}_2(m)\} \overset{p}{\to} \theta_{[1]} = \min\{\theta_1, \theta_2\}\).

Next, we write

\[
\hat{\psi}^2_{(1)}(m) = (\hat{\psi}^2_1(m) - \hat{\psi}^2_2(m))I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) + \hat{\psi}^2_2(m),
\]

where \(I(\cdot)\) is the indicator function. So, to establish \(\hat{\psi}^2_{(1)}(m) \overset{p}{\to} \psi^2_{[1]}\) we just need to show that as \(m \to \infty\),

\[
I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) \overset{p}{\to} \begin{cases} 1, & \text{if } \theta_1 < \theta_2, \\ 0, & \text{if } \theta_1 > \theta_2. \end{cases} \tag{A.1}
\]
We have \( \hat{\theta}_1(m) - \hat{\theta}_2(m) \to \theta_1 - \theta_2 \), since \( \hat{\theta}_i(m) \to \theta_i \). Hence,

\[
\lim_{m \to \infty} Pr\left( \hat{\theta}_1(m) \leq \hat{\theta}_2(m) \right) = \begin{cases} 
1, & \text{if } \theta_1 < \theta_2, \\
0, & \text{if } \theta_1 > \theta_2. 
\end{cases}
\]  
(A.2)

Since \( I(\cdot) \) is 0 or 1, for \( 0 < \epsilon < 1 \),

\[
Pr\left( \left| I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) - 1 \right| > \epsilon \right) = Pr\left( I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) = 0 \right),
\]
and

\[
Pr\left( \left| I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) \right| > \epsilon \right) = Pr\left( I(\hat{\theta}_1(m) \leq \hat{\theta}_2(m)) = 1 \right). \]
(A.3)

Now (A.1) follows upon combining (A.2) and (A.3).

(b) This can be shown on the lines of part (a) above by using Lemma A.3 (a). \( \square \)

**Lemma A.5.** Let \( N_m \) be a sequence of positive random integers such that \( N_m/m \to c \) as \( m \to \infty \), where \( c \) is a finite positive real number. Consider the sequence \( (\mu_{[1]_m}, \mu_0, \sigma_{[1]_m}, \sigma_0, \rho) \) of values of \( (\mu_{[1]}, \mu_{[2]}, \sigma_{[1]}, \sigma_{[2]}, \rho) \in \Omega \). Then as \( m \to \infty \)

(a) all the sequences \( \theta_{[1]}(N_m) - \theta_{[1]_m}, \hat{\theta}_{[2]}(N_m) - \theta_0, \hat{\psi}_{[1]}(N_m) - \psi_{[1]_m}, \hat{\psi}_{[2]}(N_m) - \psi_0 \), and \( \gamma(N_m) - \gamma_m \) converge to zero in probability,

(b) the limiting distribution of \( \left( N_m^{1/2} (\hat{\lambda}_{[1]}(N_m) - \lambda_{[1]_m}) / \hat{\tau}_{[1]}(N_m), N_m^{1/2} (\hat{\lambda}_{[2]}(N_m) - \lambda_0) / \hat{\tau}_{[2]}(N_m) \right) \) is BVN(0, 0, 1, 1, \( \gamma_0 \)).

**Proof:** (a) The convergence of \( \hat{\theta}_{[2]}(N_m) \) and \( \hat{\psi}_{[2]}^2(N_m) \) holds on the lines of part (a) of Lemma A.3 since the corresponding parameters do not depend on \( m \). For the remaining three sequences it is enough to establish that the sequences \( \hat{\mu}_{[1]}(N_m) - \mu_{[1]_m}, \hat{\sigma}_{[1]}^2(N_m) - \sigma_{[1]_m}^2, \) and \( \hat{\rho}(N_m) - \rho \) converge to zero in probability. They follow from applications of Lemma A.2 since \( \sigma_{[2]_m} = Var_m(D_{[1]}) = E(D_{[1]} - \mu_{[1]_m})^2 \to \sigma_0^2 < \infty \) and the limits of \( Var_m(D_{[1]}^2) \) and \( Var_m(D_1D_2) \) are also finite.
(b) From part (a), we conclude that \( \psi_{[i]m}(N_m) \) converge to \( \psi_0^2 \) in probability for both \( i = 1, 2 \). Consequently, \( \tilde{\tau}_{[i]}^2(N_m) \) converge to \( \tau_0^2 \) in probability. Hence from the Slutsky’s theorem it is enough to establish that the limiting distribution of

\[
\left( N_m^{1/2}(\tilde{\lambda}_{[1]}(N_m) - \lambda_{[1]m})/\tau_{[1]m}^2, N_m^{1/2}(\tilde{\lambda}_{[2]}(N_m) - \lambda_0)/\tau_0 \right) \text{ is } \text{BVN}(0, 0, 1, 1, \gamma_0).
\]

This can by established by showing that

\[
\left( \frac{1}{N_m^{1/2} \psi_{[1]m}} \sum_{j=1}^{N_m} (D_{[1]j}^2 - \theta_{[1]m}^2), \frac{1}{N_m^{1/2} \psi_0} \sum_{j=1}^{N_m} (D_{[2]j}^2 - \theta_0) \right) \overset{\mathcal{L}}{\rightarrow} \text{BVN}(0, 0, 1, 1, \gamma_0),
\]

as in part (b) of Lemma A.3. Let \( (Z_1, Z_2) \sim \text{BVN}(0, 0, 1, 1, \gamma_0) \). By Theorem 5.1.8 in Lehmann (1999), the above is equivalent to showing

\[
\frac{b_1}{N_m^{1/2} \psi_{[1]m}} \sum_{j=1}^{N_m} (D_{[1]j}^2 - \theta_{[1]m}^2) + \frac{b_2}{N_m^{1/2} \psi_0} \sum_{j=1}^{N_m} (D_{[2]j}^2 - \theta_0) \overset{\mathcal{L}}{\rightarrow} b_1 Z_1 + b_2 Z_2,
\]

(A.4)

for all real \( b_1, b_2 \) such that \( (b_1, b_2) \neq (0, 0) \). Let \( c_m = \lceil c m \rceil \) so that \( N_m/c_m \rightarrow 1 \).

For a fixed \( m \), let \( U_j = (b_1/\psi_{[1]m}) D_{[1]j}^2 + (b_2/\psi_0) D_{[2]j}^2 \). Then, the random variables \( U_j, j = 1, 2, \ldots, N_m \), are i.i.d. with mean \( E_m(U_1) = (b_1/\psi_{[1]m}) \theta_{[1]m} + (b_2/\psi_0) \theta_0 \) and variance \( \text{Var}_m(U_1) = b_1^2 + b_2^2 + 2b_1b_2 \gamma_m \). Moreover, \( \lim_{m \rightarrow \infty} E_m(U_1) = (b_1/\psi_0) \theta_0 + (b_2/\psi_0) \theta_0 \) and \( \lim_{m \rightarrow \infty} \text{Var}_m(U_1) = b_1^2 + b_2^2 + 2b_1b_2 \gamma_0 = \psi_{u_0}^2 \) (say). Notice that \( \gamma_0 = \rho(\rho \sigma_0^2 + 2 \mu_0^2)/(\sigma_0^2 + 2 \mu_0^2) \), which is away from \( \pm 1 \) since \( \sigma_0 > 0 \) and \( |\rho| < 1 \). As a consequence, \( \psi_{u_0}^2 > 0 \). Thus for (A.4) we just need to establish

\[
\frac{1}{\psi_{u_0} c_m^{1/2}} \sum_{j=1}^{N_m} (U_j - E_m(U_1)) \overset{\mathcal{L}}{\rightarrow} N(0, 1).
\]

(A.5)

Further, if \( T_m = \sum_{j=1}^{m}(U_j - E_m(U_1)) \) we can write

\[
\frac{1}{\psi_{u_0} c_m^{1/2}} \sum_{j=1}^{N_m} (U_j - E_m(U_1)) = \frac{T_{c_m}}{\psi_{u_0} c_m^{1/2}} + \frac{T_{N_m} - T_{c_m}}{\psi_{u_0} c_m^{1/2}}.
\]

(A.6)

From the Slutsky’s theorem it suffices to show that the first term on the RHS above converges to \( N(0, 1) \) in distribution and the second term converges to 0 in probability.
Due to the bivariate normality of \((D_{[1]}, D_{[2]})\), \(\psi_{\omega_0}^{-3} E_m(|U_1 - E_m(U_1)|)^3\) remains bounded as \(m \to \infty\). So an application of Corollary 2.4.1 in Lehmann (1999) shows \(T_{cm}/\left(\psi_{\omega_0} c_m^{1/2}\right) \xrightarrow{d} N(0, 1)\). Finally, since \(N_m/c_m \xrightarrow{p} 1\), we can conclude using Kolmogorov’s inequality that \(T_{N_m} - T_{cm}/\left(\psi_{\omega_0} c_m^{1/2}\right) \xrightarrow{p} 0\). This completes the proof. 

**Lemma A.6.** Consider the sequence \((\mu_{[1]}, \mu_0, \sigma_{[1]}, \sigma_0, \rho)\) of values of parameters \((\mu_{[1]}, \mu_{[2]}, \sigma_{[1]}, \sigma_{[2]}, \rho) \in \Omega\). Then as \(m \to \infty\), we have

(a) \(\hat{\tau}_1^2(m) \xrightarrow{p} \tau_0^2\), and

(b) \(\hat{\tau}_1^2(N_m) \xrightarrow{p} \tau_0^2\), where \(N_m\) is a sequence of positive random integers such that \(N_m/m \xrightarrow{p} c; c\) is a finite positive real number.

**Proof:** Here we just prove (b). The proof of (a) can be constructed similarly. A consequence of part (a) of Lemma A.5 is that \(\hat{\theta}_{[1]}(N_m) \xrightarrow{p} \theta_0\) and \(\hat{\psi}_{[1]}^2(N_m) \xrightarrow{p} \psi_0^2\) as \(m \to \infty\). Hence \(\hat{\theta}_{[1]}(N_m) \xrightarrow{p} \theta_0\). Now it is enough to show that \(\hat{\psi}_{[1]}^2(N_m) \xrightarrow{p} \psi_0^2\). By definition,

\[
\hat{\psi}_{[1]}^2(N_m) = \left(\hat{\psi}_{[1]}^2(N_m) - \hat{\psi}_{[2]}^2(N_m)\right) I\left(\hat{\theta}_{[1]}(N_m) \leq \hat{\theta}_{[2]}(N_m)\right) + \hat{\psi}_{[2]}^2(N_m).
\]

The result follows by noting that \(I\left(\hat{\theta}_{[1]}(N_m) \leq \hat{\theta}_{[2]}(N_m)\right)\) remains bounded in probability as \(m \to \infty\), and \(\hat{\psi}_{[i]}^2 \xrightarrow{p} \psi_0^2\), for \(i = 1, 2\). \(\Box\)
APPENDIX B

DATASETS

<table>
<thead>
<tr>
<th>ID</th>
<th>GSC</th>
<th>EXP1</th>
<th>ABD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.25</td>
<td>26.96</td>
<td>26.78</td>
</tr>
<tr>
<td>2</td>
<td>37.95</td>
<td>39.87</td>
<td>26.88</td>
</tr>
<tr>
<td>3</td>
<td>37.23</td>
<td>31.69</td>
<td>39.08</td>
</tr>
<tr>
<td>4</td>
<td>40.62</td>
<td>36.41</td>
<td>24.63</td>
</tr>
<tr>
<td>5</td>
<td>34.67</td>
<td>44.77</td>
<td>39.82</td>
</tr>
<tr>
<td>6</td>
<td>32.31</td>
<td>31.46</td>
<td>28.53</td>
</tr>
<tr>
<td>7</td>
<td>28.53</td>
<td>35.13</td>
<td>28.84</td>
</tr>
<tr>
<td>8</td>
<td>39.25</td>
<td>40.69</td>
<td>39.06</td>
</tr>
<tr>
<td>9</td>
<td>35.37</td>
<td>33.93</td>
<td>30.24</td>
</tr>
<tr>
<td>10</td>
<td>40.17</td>
<td>43.99</td>
<td>34.77</td>
</tr>
<tr>
<td>11</td>
<td>38.53</td>
<td>22.86</td>
<td>42.43</td>
</tr>
<tr>
<td>12</td>
<td>36.39</td>
<td>40.99</td>
<td>34.48</td>
</tr>
<tr>
<td>13</td>
<td>40.82</td>
<td>32.78</td>
<td>34.21</td>
</tr>
<tr>
<td>14</td>
<td>37.52</td>
<td>24.09</td>
<td>25.69</td>
</tr>
<tr>
<td>15</td>
<td>42.01</td>
<td>51.99</td>
<td>37.03</td>
</tr>
</tbody>
</table>

Table B.1: Fractional area change data from Hutson et al. (1998). Here GSC, EXP1 and ABD refer to the fractional area changes computed using the measurements by the committee of experts, the first expert and the computer algorithm, respectively.
<table>
<thead>
<tr>
<th>ID</th>
<th>PM</th>
<th>OM</th>
<th>ID</th>
<th>PM</th>
<th>OM</th>
<th>ID</th>
<th>PM</th>
<th>OM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>99.7</td>
<td>25</td>
<td>99</td>
<td>98.9</td>
<td>49</td>
<td>84</td>
<td>84.5</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>99.8</td>
<td>26</td>
<td>99</td>
<td>98.8</td>
<td>50</td>
<td>93</td>
<td>94.6</td>
</tr>
<tr>
<td>3</td>
<td>99</td>
<td>99.1</td>
<td>27</td>
<td>95</td>
<td>96.6</td>
<td>51</td>
<td>89</td>
<td>91.0</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>99.7</td>
<td>28</td>
<td>97</td>
<td>96.3</td>
<td>52</td>
<td>87</td>
<td>88.4</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
<td>98.5</td>
<td>29</td>
<td>97</td>
<td>96.8</td>
<td>53</td>
<td>89</td>
<td>91.2</td>
</tr>
<tr>
<td>6</td>
<td>98</td>
<td>97.2</td>
<td>30</td>
<td>99</td>
<td>98.7</td>
<td>54</td>
<td>88</td>
<td>90.1</td>
</tr>
<tr>
<td>7</td>
<td>98</td>
<td>98.0</td>
<td>31</td>
<td>93</td>
<td>93.9</td>
<td>55</td>
<td>87</td>
<td>85.6</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
<td>98.1</td>
<td>32</td>
<td>96</td>
<td>96.8</td>
<td>56</td>
<td>87</td>
<td>85.5</td>
</tr>
<tr>
<td>9</td>
<td>94</td>
<td>94.6</td>
<td>33</td>
<td>98</td>
<td>98.6</td>
<td>57</td>
<td>86</td>
<td>84.8</td>
</tr>
<tr>
<td>10</td>
<td>96</td>
<td>97.1</td>
<td>34</td>
<td>99</td>
<td>98.6</td>
<td>58</td>
<td>87</td>
<td>87.8</td>
</tr>
<tr>
<td>11</td>
<td>97</td>
<td>97.3</td>
<td>35</td>
<td>94</td>
<td>94.6</td>
<td>59</td>
<td>89</td>
<td>90.5</td>
</tr>
<tr>
<td>12</td>
<td>95</td>
<td>95.7</td>
<td>36</td>
<td>90</td>
<td>91.2</td>
<td>60</td>
<td>89</td>
<td>90.8</td>
</tr>
<tr>
<td>13</td>
<td>96</td>
<td>98.2</td>
<td>37</td>
<td>89</td>
<td>90.0</td>
<td>61</td>
<td>89</td>
<td>88.1</td>
</tr>
<tr>
<td>14</td>
<td>72</td>
<td>70.9</td>
<td>38</td>
<td>94</td>
<td>92.6</td>
<td>62</td>
<td>88</td>
<td>90.0</td>
</tr>
<tr>
<td>15</td>
<td>98</td>
<td>96.8</td>
<td>39</td>
<td>95</td>
<td>96.9</td>
<td>63</td>
<td>70</td>
<td>71.8</td>
</tr>
<tr>
<td>16</td>
<td>85</td>
<td>83.8</td>
<td>40</td>
<td>75</td>
<td>72.5</td>
<td>64</td>
<td>72</td>
<td>72.1</td>
</tr>
<tr>
<td>17</td>
<td>98</td>
<td>97.7</td>
<td>41</td>
<td>81</td>
<td>83.1</td>
<td>65</td>
<td>72</td>
<td>73.5</td>
</tr>
<tr>
<td>18</td>
<td>98</td>
<td>99.2</td>
<td>42</td>
<td>82</td>
<td>82.6</td>
<td>66</td>
<td>77</td>
<td>75.2</td>
</tr>
<tr>
<td>19</td>
<td>93</td>
<td>93.7</td>
<td>43</td>
<td>82</td>
<td>83.6</td>
<td>67</td>
<td>75</td>
<td>75.2</td>
</tr>
<tr>
<td>20</td>
<td>93</td>
<td>92.0</td>
<td>44</td>
<td>81</td>
<td>83.2</td>
<td>68</td>
<td>74</td>
<td>75.0</td>
</tr>
<tr>
<td>21</td>
<td>96</td>
<td>94.7</td>
<td>45</td>
<td>83</td>
<td>84.5</td>
<td>69</td>
<td>74</td>
<td>75.6</td>
</tr>
<tr>
<td>22</td>
<td>80</td>
<td>82.3</td>
<td>46</td>
<td>81</td>
<td>82.6</td>
<td>70</td>
<td>78</td>
<td>77.0</td>
</tr>
<tr>
<td>23</td>
<td>97</td>
<td>96.0</td>
<td>47</td>
<td>83</td>
<td>84.9</td>
<td>71</td>
<td>77</td>
<td>77.6</td>
</tr>
<tr>
<td>24</td>
<td>93</td>
<td>94.3</td>
<td>48</td>
<td>82</td>
<td>82.3</td>
<td>72</td>
<td>76</td>
<td>74.8</td>
</tr>
</tbody>
</table>

Table B.2: Oxygen saturation data from Bland and Altman (1986). Here PM and OM refer to the oxygen saturation measurements (in %) using the pulse oximeter saturation method and the oxygen saturation meter, respectively.
<table>
<thead>
<tr>
<th>ID</th>
<th>NM</th>
<th>HM</th>
<th>ID</th>
<th>NM</th>
<th>HM</th>
<th>ID</th>
<th>NM</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56.9</td>
<td>52.9</td>
<td>34</td>
<td>93.5</td>
<td>86.0</td>
<td>67</td>
<td>104.8</td>
<td>97.1</td>
</tr>
<tr>
<td>2</td>
<td>63.2</td>
<td>59.2</td>
<td>35</td>
<td>94.5</td>
<td>84.3</td>
<td>68</td>
<td>105.1</td>
<td>97.3</td>
</tr>
<tr>
<td>3</td>
<td>65.5</td>
<td>63.0</td>
<td>36</td>
<td>94.6</td>
<td>87.6</td>
<td>69</td>
<td>105.5</td>
<td>95.1</td>
</tr>
<tr>
<td>4</td>
<td>73.6</td>
<td>66.2</td>
<td>37</td>
<td>95.0</td>
<td>84.0</td>
<td>70</td>
<td>105.7</td>
<td>95.8</td>
</tr>
<tr>
<td>5</td>
<td>74.1</td>
<td>64.8</td>
<td>38</td>
<td>95.2</td>
<td>85.9</td>
<td>71</td>
<td>106.1</td>
<td>95.5</td>
</tr>
<tr>
<td>6</td>
<td>77.1</td>
<td>69.0</td>
<td>39</td>
<td>95.3</td>
<td>84.4</td>
<td>72</td>
<td>106.8</td>
<td>95.9</td>
</tr>
<tr>
<td>7</td>
<td>77.3</td>
<td>67.1</td>
<td>40</td>
<td>95.6</td>
<td>85.2</td>
<td>73</td>
<td>107.2</td>
<td>95.4</td>
</tr>
<tr>
<td>8</td>
<td>77.5</td>
<td>70.1</td>
<td>41</td>
<td>95.9</td>
<td>85.2</td>
<td>74</td>
<td>107.4</td>
<td>97.3</td>
</tr>
<tr>
<td>9</td>
<td>77.8</td>
<td>69.2</td>
<td>42</td>
<td>96.4</td>
<td>89.2</td>
<td>75</td>
<td>107.5</td>
<td>97.7</td>
</tr>
<tr>
<td>10</td>
<td>78.9</td>
<td>73.8</td>
<td>43</td>
<td>97.2</td>
<td>87.8</td>
<td>76</td>
<td>107.5</td>
<td>93.0</td>
</tr>
<tr>
<td>11</td>
<td>79.5</td>
<td>71.8</td>
<td>44</td>
<td>97.5</td>
<td>88.0</td>
<td>77</td>
<td>108.0</td>
<td>97.6</td>
</tr>
<tr>
<td>12</td>
<td>80.8</td>
<td>73.3</td>
<td>45</td>
<td>97.9</td>
<td>88.7</td>
<td>78</td>
<td>108.2</td>
<td>96.1</td>
</tr>
<tr>
<td>13</td>
<td>81.2</td>
<td>73.1</td>
<td>46</td>
<td>98.2</td>
<td>91.2</td>
<td>79</td>
<td>108.6</td>
<td>96.2</td>
</tr>
<tr>
<td>14</td>
<td>81.9</td>
<td>74.7</td>
<td>47</td>
<td>98.5</td>
<td>91.8</td>
<td>80</td>
<td>109.1</td>
<td>99.5</td>
</tr>
<tr>
<td>15</td>
<td>82.2</td>
<td>74.1</td>
<td>48</td>
<td>98.8</td>
<td>92.5</td>
<td>81</td>
<td>110.1</td>
<td>99.8</td>
</tr>
<tr>
<td>16</td>
<td>83.1</td>
<td>74.1</td>
<td>49</td>
<td>98.9</td>
<td>88.0</td>
<td>82</td>
<td>111.2</td>
<td>105.3</td>
</tr>
<tr>
<td>17</td>
<td>84.4</td>
<td>76.0</td>
<td>50</td>
<td>99.0</td>
<td>93.5</td>
<td>83</td>
<td>111.7</td>
<td>103.6</td>
</tr>
<tr>
<td>18</td>
<td>84.9</td>
<td>75.4</td>
<td>51</td>
<td>99.3</td>
<td>89.0</td>
<td>84</td>
<td>111.7</td>
<td>100.2</td>
</tr>
<tr>
<td>19</td>
<td>86.0</td>
<td>74.6</td>
<td>52</td>
<td>99.3</td>
<td>89.4</td>
<td>85</td>
<td>112.0</td>
<td>100.0</td>
</tr>
<tr>
<td>20</td>
<td>86.3</td>
<td>79.2</td>
<td>53</td>
<td>99.9</td>
<td>89.2</td>
<td>86</td>
<td>113.1</td>
<td>98.8</td>
</tr>
<tr>
<td>21</td>
<td>86.3</td>
<td>77.8</td>
<td>54</td>
<td>100.1</td>
<td>91.3</td>
<td>87</td>
<td>116.0</td>
<td>110.0</td>
</tr>
<tr>
<td>22</td>
<td>86.6</td>
<td>80.8</td>
<td>55</td>
<td>101.0</td>
<td>90.4</td>
<td>88</td>
<td>116.7</td>
<td>103.5</td>
</tr>
<tr>
<td>23</td>
<td>86.6</td>
<td>77.6</td>
<td>56</td>
<td>101.0</td>
<td>91.2</td>
<td>89</td>
<td>118.8</td>
<td>109.4</td>
</tr>
<tr>
<td>24</td>
<td>86.6</td>
<td>77.5</td>
<td>57</td>
<td>101.5</td>
<td>91.4</td>
<td>90</td>
<td>119.7</td>
<td>112.1</td>
</tr>
<tr>
<td>25</td>
<td>87.1</td>
<td>78.6</td>
<td>58</td>
<td>101.5</td>
<td>93.0</td>
<td>91</td>
<td>120.7</td>
<td>111.3</td>
</tr>
<tr>
<td>26</td>
<td>87.5</td>
<td>78.7</td>
<td>59</td>
<td>101.5</td>
<td>91.2</td>
<td>92</td>
<td>122.8</td>
<td>108.6</td>
</tr>
<tr>
<td>27</td>
<td>87.8</td>
<td>81.5</td>
<td>60</td>
<td>101.8</td>
<td>92.0</td>
<td>93</td>
<td>124.7</td>
<td>112.4</td>
</tr>
<tr>
<td>28</td>
<td>88.6</td>
<td>79.3</td>
<td>61</td>
<td>101.8</td>
<td>91.8</td>
<td>94</td>
<td>126.4</td>
<td>113.8</td>
</tr>
<tr>
<td>29</td>
<td>89.3</td>
<td>78.9</td>
<td>62</td>
<td>102.8</td>
<td>96.8</td>
<td>95</td>
<td>127.6</td>
<td>115.6</td>
</tr>
<tr>
<td>30</td>
<td>89.6</td>
<td>85.9</td>
<td>63</td>
<td>102.9</td>
<td>92.8</td>
<td>96</td>
<td>128.2</td>
<td>118.1</td>
</tr>
<tr>
<td>31</td>
<td>90.3</td>
<td>80.7</td>
<td>64</td>
<td>103.2</td>
<td>94.0</td>
<td>97</td>
<td>129.6</td>
<td>116.8</td>
</tr>
<tr>
<td>32</td>
<td>91.1</td>
<td>80.6</td>
<td>65</td>
<td>103.8</td>
<td>93.5</td>
<td>98</td>
<td>130.4</td>
<td>121.6</td>
</tr>
<tr>
<td>33</td>
<td>92.1</td>
<td>82.8</td>
<td>66</td>
<td>104.4</td>
<td>95.8</td>
<td>99</td>
<td>133.2</td>
<td>115.8</td>
</tr>
</tbody>
</table>

Table B.3: Plasma volume data from Bland and Altman (1999). Here NM and HM refer to the plasma volume measurements (as % of normal) using the sets of normal values due to Nadler and Hurley, respectively.
BIBLIOGRAPHY


