Available online at www.sciencedirect.com
ScienceDirect
Theriogenology 73 (2010) 1167–1179
Theriogenology
www.theriojournal.com

Invited Review

Method agreement analysis: A review of correct methodology

P.F. Watson [a,*], A. Petrie [b]

[a] *The Royal Veterinary College, London, United Kingdom*
[b] *The UCL Eastman Dental Institute, London, United Kingdom*

Abstract

The correct approach to analyzing method agreement is discussed. Whether we are considering agreement between two measurements on the same samples (repeatability) or two individuals using identical methodology on identical samples (reproducibility) or comparing two methods, appropriate procedures are described, and worked examples are shown. The correct approaches for both categorical and numerical variables are explained. More complex analyses involving a comparison of more than two pairs of data are mentioned and guidance for these analyses given. Simple formulae for calculating the approximate sample size needed for agreement analysis are also given. Examples of good practice from the reproduction literature are cited, and common errors of methodology are indicated.

Keywords: Agreement analysis; Reliability; Repeatability; Reproducibility; Sample size calculation

Contents

* Corresponding author. Tel.: +44 0 1920 466941.
E-mail address: pwatson@rvc.ac.uk (P.F. Watson).

doi:10.1016/j.theriogenology.2010.01.003

1. Introduction

A common question encountered in reproductive biology is whether or not the measurement of a variable by two different methods, or by two different operators using the same method, or by one operator repeating the measurement at two different times, produces essentially the same result. We are concerned both with accuracy (the way in which an observed value of a quantity agrees with the true value) and precision (a measure of the extent to which repeated observations conform). Examples might be the measurement of blood hormone concentrations or the use of two different techniques for determining pregnancy status. It is important to use appropriate statistical methods to address a question such as this.

For many years, it was common to use one of several incorrect methods to answer this question with the consequence of unsatisfactory or sometimes misleading conclusions. In this article, we will illustrate and highlight the correct approaches to address the problem of assessing the consistency of the measuring process using some examples drawn from the literature. An overview of the procedures discussed is given according to type of variable in Table 1.

1.1. Measurement variability and measurement error

When we measure a biological variable in a number of individuals or repeatedly within an individual (either within a short time or over a longer period), the data always exhibit, to a greater or lesser extent, a scatter of values. Inter-individual variation (between individuals) as well as intra-individual variation (within individual) is thus likely to be evident. Much of this variability is due to variation in associated factors (e.g., genetic, social, or environmental factors). For example, if these individuals differ in terms of their reproductive status, age, weight or gender, blood hormone measurements may be expected to vary. Similarly, if we take repeated measurements from an individual at different times of the day, they may well vary. This variability is termed *measurement variability.* In contrast, *measurement error* is defined as that which arises because the observed (or ''measured'') values and true values of a variable differ (note that although we refer to the ''true'' measurement here, it is rarely possible to obtain this value). Two kinds of measurement error can occur:

- *Random:* The observed values may be sometimes higher or lower than the true values, but on average they tend to balance out. For example, the measurement may be read on a scale to the nearest division. Although random error is governed by chance, the degree of error can be influenced by external factors (e.g., a balance may exhibit greater random variability when sited in a drafty location).
- *Systematic:* The observed values have a tendency to be consistently high (or low) because of some extraneous factor, known or unknown, affecting the measurements in the same way (e.g., because of an instrument that has not been calibrated correctly or an observer consistently overestimating the values). This kind of error, which concerns the overall accuracy of the observations, results in biased results if one set of results represents the true values. The error must be eliminated or minimized by attention to issues such as training of personnel, standardization of conditions of measurement, and proper calibration and maintenance of instruments (i.e., verification by comparison with a known standard).

Although this explanation of error has centered on laboratory measurements, the same concepts apply even if interest is focused on other forms of measurement,

Table 1
Summary of procedures for agreement analysis.

Number of methods to compare	Variable		Procedure
2	Categorical	2 categories	Cohen's kappa
			McNemar's test
		>2 ordered categories	Weighted kappa
	Numerical		Intraclass correlation coefficient
			Lin's concordance correlation coefficient
			Bland and Altman diagram
			Paired *t*-test
			British Standards reproducibility coefficient
>2	Consult an appropriate advanced text or a statistician		