Available online at www.sciencedirect.com
ScienceDirect
Clinica Chimica Acta 379 (2007) 71 – 80
www.elsevier.com/locate/clinchim

A more flexible parametric estimation of univariate reference intervals: A new method based on the GS-distribution

Javier Trujillano, José M Muiño, Jaume March, Albert Sorribas *

Grup de Bioestadística i Biomatemàtica, Departament de Ciències Mèdiques Bàsiques, Institut de Recerca Biomèdica de Lleida (IRBLLEIDA), Universitat de Lleida, Montserrat Roig, 2, 25008-Lleida, Spain

Received 25 October 2006; received in revised form 11 December 2006; accepted 11 December 2006
Available online 22 December 2006

Abstract

Background: Reference interval estimation is an important issue in clinical laboratories. Present methods are based either on data transformation or on non-parametric approaches.
Methods: We present a new technique based in a family of statistical distributions known as GS-distributions that provide a suitable model for continuous unimodal variables. We compare, both by simulation studies an on actual data, the reference intervals estimated by using non-parametric methods and data transformations suggested by the IFCC and those obtained by fitting a GS-distribution. Simulated data are generated from various distributions to evaluate the accuracy of these methods. In each case, confidence intervals for the resulting reference intervals are obtained by bootstrap.
Results: In all the cases, the GS-distribution based method provides comparable or more accurate results than the non-parametric methods. In most cases, the proposed method produces better results than those obtained by transforming the original data.
Conclusions: Our results suggest that the method for computing reference intervals based on GS-distribution is a valid alternative for the current non-parametric methods.

Keywords: Reference intervals; Statitical distributions; Reference range

1. Introduction

Reference intervals (RI) play an important role in clinical practice as they are required for assessing the health status of patients. Furthermore, they are a basic tool of clinical laboratories, both in quality control and in providing reference values according to the protocols used in each case. An RI is typically defined as the range comprised between the 2.5 and 97.5 percentiles of the data distribution from a given reference population. Accordingly this interval estimates the expected values that would contain the 95% of the subjects of the considered population. Guidelines for appropriately estimating RI include rules for subject selection, data validation, outlier detection, and indications on the appropriate statistical computations [1]. In particular, the target reference population should be clearly defined. In general, the reference population is any population defined according to precise inclusion criteria and it does not always correspond strictly to a healthy population [2]. For instance, in a given clinical application the target population could be those patients of a given age range that present a severe status of a given disease.

From a statistical point of view, the available approaches for RI estimation include non-parametric methods [3–5], robust methods [6–8], transformation methods [9–11], and different variants of these basic methodologies (see Ref. [12] for a review). From a practical point of view, it is common to follow the NCCLS (National Committee for Clinical Laboratory Standards) recommendations and obtain nonparametric reference intervals using a sample size of at least 120 subjects [1].

The development of the different methods indicated above arises from the lack of information on the underlying

Abbreviations: GSD, GS-distribution; TST, Two-stages transformation method; NP, non-parametric method; HDM, Harrell–Davis method; RMSE, Root Mean Square Error.

* Corresponding author. Fax: +34 973702426.
E-mail address: albert.sorribas@cmb.udl.es (A. Sorribas).

0009-8981/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cca.2006.12.010

distribution. Otherwise, the computation of the corresponding RI would be straightforward once the appropriated parameters for the distribution were obtained from the sample data. When no reasonable information is available on the underlying distribution, one may consider using a transformation that provides a new variable with know distribution. For instance, an appropriate Box–Cox transformation can convert the actual variable in a new variable with normal distribution [10,11]. RI would then be estimated on the transformed variable; a reverse transformation would provide the required RI on the original variable. In many cases, a simple logarithmic transformation is used [13,14]. However, this may not be appropriate for most situations and alternative transformations should be considered. Although this is a suitable technique, there is no guarantee that this transformation exists for a given set of data [15]. Furthermore, even when the appropriate transformation can be found, the back transformation may be impossible, due to out-of-bounds problems with interpolated values. Non-parametric methods provide a highly recommended alternative in those cases.

The problem of estimating a RI would be greatly simplified if a general parametric model could be defined. Then, the problem would be reduced to obtain the appropriate parameters according to the data and to assess the goodness-of-fit. Once the particular instance of the distribution is fitted, RI would be obtained by a simple computation. With this possibility in mind, we developed the GS-distribution (GSD) [16]. This is a family of distributions defined as:

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \alpha F(x)^g (1-F(x)^k)^\gamma \qquad F(x_0) = 0.5 \tag{1}$$

where $F(x)$ is the cumulative. This family has three parameters that account for the shape of the resulting distribution (g, k, γ). Thus, these parameters are responsible for the skewness and kurtosis of the resulting distribution. Parameter α is related to the spread of the distribution, and x_0 corresponds to the median of the distribution and fixes the initial conditions of the differential equation. For simplicity, we shall indicate a given GSD as GSD[x_0, α, g, k, γ]. Some examples of the flexibility of the GSD are shown in Fig. 1.

The GSD is a parametric family that results from a generalization of the S-distribution [17–19] and it is more flexible than classical parametric models and hence is better for modeling data observed in practice [16]. Using this family, we can fit a GSD to unimodal data without further assumptions on the actual underlying distribution. As any continuous unimodal distribution can be accurately represented as a GSD, this assures that we can always obtain an estimated distribution that fit the data. In most cases, the resulting fit is comparable to the one we would obtain if the true distribution was known. In that sense, the GSD is a practical tool for obtaining a distribution that explains the

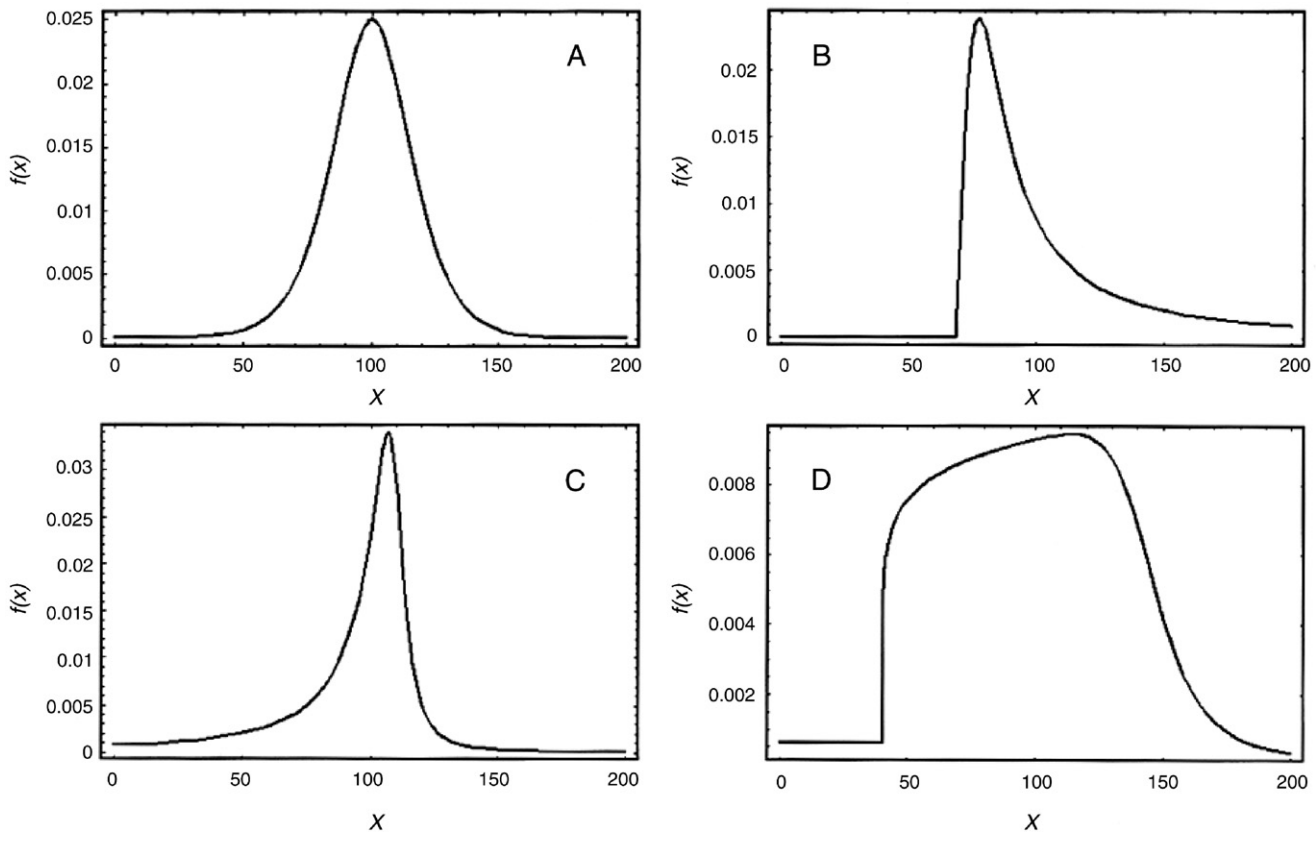

Fig. 1. Examples of GSD. In all cases the median is equal to 100. Parameters α, g, k, γ are: (A) 0.1, 1.0, 1.0, 1.0; (B) 0.1, 0.5, 1.0, 3.0; (C) 0.1, 2.0, 5.0, 2.0; (D) 0.01, 0.1, 12.0, 2.0.