Journal of Multivariate Analysis 134 (2015) 50–60

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Central tolerance regions and reference regions for multivariate normal populations

Xiaoyu Dong, Thomas Mathew*

Department of Mathematics and Statistics, University of Maryland, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 6 April 2013
Available online 7 November 2014

AMS 2000 subject classifications:
primary 62F25
secondary 62H99

Keywords:
Central tolerance factor
Content
Multivariate regression
Prediction region
Wishart distribution

ABSTRACT

Reference intervals and regions are widely used to identify the measurement range expected from a reference population. Such regions capture the central part of the population, and have potential applications in the field of laboratory medicine. Furthermore, the uncertainty in an estimated reference region can be assessed using a central tolerance region, namely, a region that will contain the population reference region, with a specified confidence level. The construction of a central tolerance region is investigated in this article for a multivariate normal population, and also for a multivariate normal linear regression model. A theoretical framework is developed that will facilitate the numerical computation of the tolerance factor. The performance of a prediction region is also evaluated, in terms of capturing the central part of the population, and the prediction region is found to be unsatisfactory. Some examples from laboratory medicine are used to illustrate the results.

1. Introduction

In the field of laboratory medicine, reference intervals and reference regions represent a range of values that a physician can use in order to interpret the test results of a patient. For a given test, the reference interval (or region) is obtained based on test results from healthy individuals. A test result outside the reference region may be indicative of a disease. In the case of univariate test results from a healthy population, a 95% reference interval is simply the interval from the 2.5th to the 97.5th percentiles of the distribution of the test results; see the guidance document of the National Committee for Clinical Laboratory Standards [22], and the books by Horn and Pesce [10] and by Harris and Boyd [7]. In other words, the reference interval captures 95% of the central part of the distribution. When covariates (age, for example) are present, regression based reference limits are very often used; see, for example, [26,28,29]. Typically, the percentiles that capture 95% of the central part of the distribution are unknown, and have to be estimated using a random sample of test measurements from healthy individuals. Clearly, if point estimates of the percentiles are used, the coverage of the resulting interval is random, and could be less than the desired 95%. Furthermore, the issue of quantifying the uncertainty in an estimated reference interval should be of obvious interest. In the literature, this is sometimes done by computing separate confidence intervals for the lower and upper limits of the reference interval, where each confidence interval has a specified confidence level, say 95%; see [7]. Another option is to compute an interval that contains the population reference interval with 95% confidence level. In the case of normally distributed endpoints, this has been done by Owen [23]. The resulting interval that is expected to include the population reference interval is referred to as a central tolerance interval; see also [17, Chapter 3]. The use of tolerance intervals in the context of estimating reference intervals is indeed pointed out in the literature; see [1, p. 55] and

* Correspondence to: Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA.

E-mail addresses: hhdong12@gmail.com (X. Dong), mathew@umbc.edu (T. Mathew).

http://dx.doi.org/10.1016/j.jmva.2014.10.009
0047-259X/

[9,13]. In [13, Section 4.1], the authors argue that separate confidence intervals for the lower and upper limits of the reference interval, or a joint confidence region for them, are both inappropriate for assessing the uncertainty in the estimated reference interval; rather, a tolerance interval must be used. Similar to the reference interval for a univariate normal distribution, a reference region can be defined for a multivariate normal distribution. Computation of such a region, and the assessment of its uncertainty are then of obvious interest. This article addresses these issues for multivariate normal distributions and multivariate regression models.

2. Background

Applications where reference regions are required in the multivariate context are available in the laboratory medicine literature; see for example, the books by Albert and Harris [1] and Harris and Boyd [7]. An example given in [1] deals with the assessment of kidney function using the amount of urea, uric acid and creatinine in blood. Here it is required to construct a reference region based on data on these analytes from 284 healthy subjects. An application where a reference region is required in the multivariate linear regression context is given in [21]. The problem addressed by the authors deals with the construction of reference regions for the serum concentrations of insulin-like growth factor I (S-IGF-I), insulin-like growth factor-binding protein 2 (S-IGFBP-2) and insulin-like growth factor-binding protein 3 (S-IGFBP-3) among healthy individuals, where age, gender and body mass index are covariates. For individuals in certain disease states, the serum concentrations of S-IGF-I, S-IGFBP-2 and S-IGFBP-3 are expected to be different from those among healthy individuals. Consequently, such disease states can be identified using a reference region based on data from healthy individuals.

2.1. Univariate reference intervals versus a multivariate region

The articles by Boyd [2] and by Trost [25] underscore the need to have multivariate reference regions, since different tests on the same individual could be correlated, and the univariate reference limits do not contain any information about the cross-correlations. Numerical results, given for example in [25, Section 4.1.1], show that using several 95% univariate reference intervals does not result in a 5% false positive probability; the probability can be significantly higher. The issue of univariate intervals as opposed to a multivariate region is also addressed in Section 3.4.

2.2. Prediction region or tolerance region?

Prediction intervals (in the univariate case) and prediction regions (in the multivariate case) have been recommended in the literature to be used as reference intervals and regions, respectively. In his work on multivariate reference regions, Trost [25, Section 2.1] mentions that "Reference intervals referred to in this document are arguably the closest to prediction intervals since we want exactly 95% of the future observations from reference individuals to fall inside the bounds". This statement implies that reference bounds, once obtained, will be used repeatedly to decide whether or not future measurements come from the reference population. The condition that is used to derive a prediction interval does not capture this *repeated use* scenario of the *same* reference interval. Rather, it is the tolerance interval that takes care of the repeated use aspect. Under the normality assumption for a univariate reference population, the central part is the interval from the 2.5th to the 97.5th percentiles, and is typically taken as the population reference interval. An interval, based on a random sample, that includes the central part of the reference population is the tolerance interval; to be precise, a central tolerance interval. In fact, while commenting on the criterion to be used to construct a reference interval, Albert and Harris [1, p. 55] comment that "It would seem that the statistical tolerance interval is what clinical chemists have in mind when they speak of a reference range derived from a sample of individuals representing some defined population". Thus a central tolerance interval captures the central part of the reference population, with a given confidence level, and it can be used to quantify the uncertainty in an estimated reference interval. A prediction region is certainly narrower compared to a central tolerance region, and the net result is that the use of a central tolerance region will result in a smaller false positive rate, at the cost of increasing the false negative rate. We recall that the false positive rate is the probability that an individual from the reference population is declared to be outside the population. (A false negative rate is similarly defined.) In terms of capturing the central part of a population, the inappropriateness of a prediction region is also noted in Section 3.3.

Similar to a central tolerance interval in the univariate case, a central tolerance region can be defined for a multivariate normal distribution. Such a central tolerance region is actually defined in [12], and the definition is as follows. Let $\boldsymbol{y}$ be a $q \times 1$ vector following the multivariate normal distribution $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Consider the region $\mathcal{R}$ given by

$$\mathcal{R} = \left\{\boldsymbol{y} : (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \leq \chi^2_q(p)\right\}, \tag{1}$$

where $\chi^2_q(p)$ denotes the pth percentile of a chi-square distribution with df $= q$. It is clear that $\mathcal{R}$ is the central $100p\%$ region of the multivariate normal distribution $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A region constructed based on a random sample, that contains $\mathcal{R}$ with a given confidence level, is referred to as a central tolerance region for $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The proportion p in (1) is referred to as the *content* of the central tolerance region. A central tolerance region can be similarly defined for a multivariate regression model, under the assumption of multivariate normality.