ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 610–615

Journal of Clinical Epidemiology

A method is presented to plan the required sample size when estimating regression-based reference limits

Carine A. Bellera*, James A. Hanley

Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H2T 1H3, Canada

Accepted 16 September 2006

Abstract

Background and Objective: Reference limits are widely used in anthropometry, the behavioral sciences, medicine, and clinical chemistry. They describe the distribution of a quantitative variable in a healthy population, and are often a smooth function of age or another determinant. Thus, instead of estimating reference limits separately for several age groups, it is more economical and parsimonious to use regression methods to estimate reference limits as a function of age. Although the variability of regression-based reference limits has been addressed previously, the available methods to determine the sample sizes needed to estimate them are neither transparent nor user-friendly.

Methods: We propose a simple and intuitive formula using margins of error, to project the sample sizes required to achieve a given degree of precision, for different sampling strategies.

Results: We present two examples for the calculation of the sample size required to estimate a specific reference limit using various age distributions.

Conclusion: We provide a simple formula to calculate the sample size needed to estimate a specific reference limit to a specified degree of precision. The structure of the formula can easily accommodate different age-sampling strategies.

Keywords: Reference limits; Regression; Precision; Sample size; Study design; Sampling strategy

1. Introduction

Reference limits are widely used in anthropometry, the behavioral sciences, medicine, and clinical chemistry. Just like the mean or the median, these parameters are used to describe the distribution of a characteristic, measured on a continuous scale, in a healthy population. The 100p% reference limit, where $0 \leqslant p \leqslant 1$, is the value below which 100p% of the values fall. For example, the median is equivalent to the 50% reference limit. Reference limits are also called reference values, percentiles, or quantiles. The terms reference limits and reference values are mostly used in the life sciences, and percentile and quantile in the statistical literature. In this paper, we have chosen to use the term *reference limit*.

The reference *limit* should be distinguished from the $100(1-\beta)$% reference *range*, where $0 \leqslant \beta \leqslant 1$, and which corresponds to the interval between two predetermined reference limits centered around the median and encompassing $100(1-\beta)$% of all values.

Suppose we have a continuum of distributions indexed by a covariate. For example, assume that we are studying blood pressure (BP) in a group of adults with different ages. We might be interested in the BP distribution, more particularly in the 5% reference limit of the BP distribution for various ages. How many adults should we sample in order to have a precise estimate of this reference limit? One way that this question can be answered is by applying linear regression techniques to estimate the reference limit as a function of age. Sample size issues for *regression-based reference limits* have been discussed in the literature; however, the proposed estimation techniques are neither complete nor transparent enough. We present a simple formula for the minimum sample size required to estimate a regression-based reference limit, given a specific margin of error. We illustrate its practicality through simple examples, assuming several age distributions. Once the sampling strategy has been selected, we show that the sample size formula can be easily adapted to estimate the required sample size at the desired reference limit and

* Corresponding author. Unité de Recherche Epidémiologique et Clinique, Institut Bergonié, Centre Régional de Lutte Contre le Cancer de Bordeaux et du Sud-Ouest, 229 Cours de l' Argonne, 33076 Bordeaux, France. Tel. +33-05-56-33-33-33.

E-mail address: bellera@bergonie.org (C.A. Bellera).

0895-4356/07/$ – see front matter
doi: 10.1016/j.jclinepi.2006.09.004

covariate value. Mathematical details are provided in the Appendix.

2. Background

We first discuss some of the methods available for the estimation of a reference limit and its precision as a function of an underlying covariate. A first simple method involves categorizing the covariate, such as age, into a small number of groups, and estimating the reference limit within each age group separately. Within each age group, one can apply nonparametric or parametric methods to estimate the reference limit, calculate its precision and the required sample size. Nonparametric approaches have been studied extensively [1–4]. Because they involve order statistics and thus ranking the observations, the confidence intervals for the reference limits are data driven and thus ''after-the-fact''. For that reason, they do not allow us, when planning the study, to project how wide the confidence interval will be. If one is willing to make distributional assumptions, parametric methods are more flexible, and allow one to evaluate the sample size for a fixed degree of (relative) precision ahead of time. We assume that the variable of interest Y follows a normal distribution at any given age. Then given the random variable Y has a cumulative distribution function F_Y, that is, $Y \sim F_Y(y)$, the 100p% reference limit is Q_p, the value below which 100p% of the Y values fall:

$$F_Y(Q_p) = p, 0 \leqslant p \leqslant 1.$$

For a random sample of size n, the sample mean and sample standard deviation are given by $\bar{y}$ and s, respectively. Then, the parametric estimator of Q_p, the 100p% reference limit, is $\bar{y} + z_p s$, where z_p is the standard normal deviate corresponding to p. From the variance of this estimator, one can then project the sample size requirements for each group and sum these over the age groups to obtain the overall sample size. Once within-group reference limits are estimated, it is possible to join them across groups, but the resulting curve is usually rough and often nonmonotonic. An alternative smoothing approach consists in calculating reference limits in overlapping age groups [5].

Although these methods can be easily implemented, they do not take advantage of the smooth structure of the age-specific distributions. To take this characteristic into account, we can apply linear regression techniques, assuming the linearity and normality assumptions can be fulfilled (the latter possibly after an appropriate transformation of the data) to estimate the reference limit as a function of the covariate distribution. Sample size issues for such regression-based reference limits have been considered, however, the proposed methods are neither complete nor transparent enough for statisticians, and much less so for nonstatistical end-users [6–8].

3. Methods

Our objective is to provide a simple formula for the minimum sample size required to estimate a regression-based reference limit, given a specific margin of error. We assume that the mean value of the response variable of interest (e.g., BP) varies linearly with the covariate (e.g., age), and that the response values are approximately normally distributed about this mean. In addition, the variability of the response values is assumed to remain constant across all values of the covariate. Finally, the following sample size formulas are based on large sample results and should be applied in large enough data sets only. Given the large sample sizes typically envisaged, the large sample behavior is likely to be appropriate.

When calculating the sample size required to estimate a regression-based reference limit, we need to specify several parameters:

1. The 100p% reference limit of interest and the corresponding standard normal deviate, z_p. Recall that z_p is the value below which 100p% of the standard normal distribution lies. For example, if we are interested in the 95% reference limit, the standard normal deviate is $z_{0.95} = 1.645$ (one sided).
2. The 100(1−α)% confidence interval for the reference limit of interest, and its corresponding standard normal deviate, $z_{1-(\alpha/2)}$, where $0 < \alpha < 1$. For example, if we want the two-sided 95% confidence interval, then $\alpha = 0.05$ and $z_{1-(\alpha/2)} = z_{0.975} = 1.96$.
3. The 100(1−β)% reference range, which encompasses 100(1−β)% of the y values (e.g. BP) as well as its corresponding standard normal deviate, $z_{1-(\beta/2)}$, where $0 < \beta < 1$. For example, if we want the 100(1−β)% = 95% reference range, then $\beta = 0.05$ and $z_{1-(\beta/2)} = z_{0.975} = 1.96$.
4. The relative margin of error Δ. It is defined as the ratio of the width of the 100(1−α)% confidence interval for the reference limit to the width of the 100(1−β)% reference range [9,10]. This means that we want a sample size large enough so that the width of the 100(1−α)% confidence interval for our reference limit is small when compared to the width of the 100(1−β)% reference range (we usually take $\alpha = \beta$). For example, suppose that in a given sample, 95% of the systolic BPs at age 30 are between 100 mmHg and 140 mmHg, then the 95% reference range has a span of 40 mmHg. Then, if we want a margin of error of $\Delta = 10\%$, we should measure the 100p% reference limit such that its confidence interval has a width less than 4 mmHg, that is, 10% of the span of 40 mmHg.
5. The design of the study, that is, the distribution of the covariate (e.g., age) in the sample investigated, which will also impact the computation of the sample size.

Once we have specified the above parameters, we can estimate the sample size. Assume, first, that we choose our sample so that the covariate follows a uniform distribution. It can be shown after some mathematical derivations (see Appendix), that the minimum sample size, n, required to estimate the $100(1-\alpha)\%$ confidence interval for the $100p\%$ reference limit, with a margin of error of Δ, when compared to the $100(1-\beta)\%$ reference range, is given by:

$$n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}\left(4+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}. \tag{1}$$

Equation (1) can also be modified to suit *other sampling strategies* (see Appendix). Instead of a uniform age distribution, one might take one-third of the sample at one age extreme, one-third at the midpoint, and one-third at the other age extreme. In this study design, the sample size requirement becomes:

$$n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}\left(\frac{5}{2}+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}. \tag{2}$$

Similarly, we can also expect that the age distribution in the sample will follow a normal distribution. If we assume that the range of X is approximately 4 times the standard deviation of X, then we show that the sample size requirement becomes:

$$n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}\left(5+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}. \tag{3}$$

Note that equations (1)–(3) were all derived under the ''worst-case'' scenario, that is, assuming that we are interested in estimating the reference limit at the extreme end of age, where the variability is highest, and thus the largest sample size is needed. If on the other hand, one is interested in the $100p\%$ reference limit at the *average age value*, then the sample size formula is reduced as follows, whatever the age distribution:

$$n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}\left(1+\frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2}. \tag{4}$$

Additional mathematical details are provided in the Appendix. In particular, we provide an insight into why the equations (1)–(3) are very similar and differ only by the first factor in the parenthesis, which equals 4, 5/2, or 5, depending on the design of the study.

4. Results

As an illustration, suppose that we are interested in estimating a specific BP reference limit as a function of age. Specifically, we wish to produce a 95% confidence interval ($z_{1-(\alpha/2)}=1.96$) for the 95% BP reference limit ($z_p = 1.645$), with a relative margin of error of 10% (Δ = 10%) when compared with the 95% reference range ($z_{1-(\beta/2)}=1.96$). If age is uniformly distributed in the sample, then applying equation (1):

$$n \geqslant \frac{1.96^2\left(4+\frac{1.645^2}{2}\right)}{1.96^2 \times 0.10^2} = 536,$$

i.e., we would need at least 536 observations to obtain this precise an estimate of the 95% BP reference limit at *any place in the age range*.

If on the other hand, one is interested in the $100p\%$ reference limit only at the average age value, or in a homogeneous population not indexed by a covariate, then we apply equation (4), and obtain a sample size of at least 236 observations for the same settings as above.

As a second example, consider a study investigating heights in young children. Suppose we are interested in estimating the 80% reference limit ($z_p = 0.845$) with a 95% confidence interval ($z_{1-(\alpha/2)}=1.96$). We also want a relative margin of error of 10% (Δ = 10%) when compared with the 95% reference range ($z_{1-(\beta/2)}=1.96$), that is if 95% of heights at 5 years of age are between 100 cm and 120 cm—a span of 20 cm—then we should try to measure the 80% reference limit with a total width less than 2 cm, that is, 10% of the span of 20. Applying equations (1)–(3), we have estimated the sample size in Table 1, for several age distributions (uniform, normal, or equally distributed in three groups at the midpoint and extreme ranges). The sample size is smallest when using three age subgroups (286 subjects) and largest in the case of normally distributed ages (536 subjects). If on the other hand, one is interested in the 80% reference limit at the average age value, then using equation (4), 136 observations are required, whatever the age distribution.

5. Discussion

Sample size issues for regression-based limits have been considered [6–8], but the proposed methods are neither flexible nor complete. Royston provided the standard error (from which the required sample size can be deduced) for a reference limit, but only at the mean value $\bar{x}$ of the covariate [6]. However, the author did not consider other values along the covariate range where the uncertainty in

Table 1
Minimum sample size (n) required to estimate the 80% reference limit depending on the age distribution of those sampled

Age distribution	Formula	n
Uniform	1	436
1/3 at each end, 1/3 at the midpoint	2	286
Normal distribution ($\text{range}_X \approx 4\sigma_X$)	3	536

the estimated limit is greater. Virtanen et al. considered the complete covariate range but the method was overly complex and not transparent [7]. Specifically, the authors suggested that one should ensure that the value $v=1/n+(x_0-\bar{x})^2/\sum_i(x_i-\bar{x})^2$ did not exceed 0.1 at the minimum and maximum covariate values, where x_0 is the value of the covariate at which we want to compute the reference limit. Then if this condition is satisfied, the authors concluded that reference limits with sufficiently narrow confidence intervals could be produced by regression analysis with a sample size of about 70. Note that this approach does not allow the user to plan for a desired level of precision nor to specify where along the covariate range he/she is interested in. Moreover, it seems counterintuitive that a sample size of 70 will allow one to estimate extreme reference limits precisely. Finally, Elveback and Taylor reviewed the variability of the point estimator of the reference limit but did not address precision or sample size requirements [8].

Previous formulas did not allow the end-user to take specific requirements into account, such as the anticipated level of precision or at which value of the covariate the reference limit is calculated. Our approach is more flexible as it allows the user to compute the minimum sample size after specifying the desired level of precision and the sampling strategy. In addition, we have provided formulas for specific distributions of the covariate (uniform, normal, or three groups), but other distributions are easily accommodated by adjusting the intermediate equations detailed in the Appendix.

If there is some nonlinearity in the covariate, such as for example a quadratic relationship, the formula can also be accommodated by adjusting the point estimator of the reference limit of interest (see equation (5), Appendix).

Similarly, one may also need to allow for heteroscedasticity of the variable of interest across the covariate, such as when dealing with growth curves. In this case, regression techniques can be used to model the standard deviation as a function of the mean, and our formula can still be used as a rough guide for sample size planning.

Finally, it is important to notice that the variation σ to be used in planning (see Appendix) includes both the true interindividual variability and the variability of the measuring instruments used; measurement tools with differing precisions will provide different sample size estimates.

6. Conclusion

We provide a simple method to calculate the sample size needed to estimate a regression-based reference limit with a specific degree of precision. The sample sizes provided are estimated at the extreme range of the covariate where the variability is the most important, but the formula can be easily adapted to estimate reference limits at other locations of the covariate, such as at the mean value of the covariate, as illustrated. Finally, our formula can also be reversed, and solved for the relative margin of error or the confidence limit if the sample size is already fixed.

Acknowledgments

We wish to thank Dr. Marie Lambert (Department of Pediatrics, Hôpital Sainte-Justine and Université de Montréal), as well as Dr. Antonio Ciampi and Dr. Robert Platt (Department of Epidemiology and Biostatistics, McGill University) for their useful comments.

Support: This work was supported by an operating grant to JH from the Natural Sciences and Engineering Research Council of Canada and by a team grant from Fonds de la Recherche sur la Nature et Technologies du Québec.

Appendix

Mathematical derivations

In this section, we review the intermediate steps necessary to produce our sample size formula, as given by equation (1). Let X denote the covariate of interest, such as age. Assume that at any given value x_0 of age, the mean value of interest, such as BP, is an approximate linear function of X, and the individual BP values are normally distributed around this mean (the latter eventually after a suitable transformation) with constant variance:

$$Y|x_0 \sim N(\beta_0+\beta_1 x_0,\sigma^2).$$

The $100p\%$ reference limit for Y at this specific age point x_0 is given by:

$$\begin{aligned} Q_0 &= \beta_0+\beta_1 x_0+z_p\sigma \\ &= \mu_0+z_p\sigma \end{aligned}$$

where z_p is the standard normal deviate corresponding to the $100p\%$ reference limit of interest.

Given n selected individuals with data points $\{(x_i, y_i), i=1,\ldots, n\}$, a point estimator for the $100p\%$ reference limit is given by:

$$\hat{Q}_0=\hat{\mu}_0+z_p s_{Y|X} \quad (5)$$

where $\hat{\mu}_0=\hat{\beta}_0+\hat{\beta}_1x_0$, $\hat{\beta}_0$, and $\hat{\beta}_1$ are obtained by least-squares estimation of the regression coefficients β_0 and β_1, and $s_{Y|X}$ is the observed root mean square error. Note that $s_{Y|X}$ is not perfectly unbiased for σ, but with the large sample size envisaged, it can be used as a close-to-unbiased estimator for σ, with an approximate variance of $\sigma^2/2n$ (with more than 60 observations, $s_{Y|X}$ is almost unbiased for σ [9]). Therefore, $\hat{Q}_0$ is a close-to-unbiased estimator for Q_0.

We first assume that in the sample the covariate X is *uniformly* distributed over its range $range_X$, and therefore its variance is:

$$\sigma_X^2 = \frac{(x_{\max} - x_{\min})^2}{12} = \frac{\text{range}_X^2}{12}.$$

Next, to be conservative, we compute the variance of the estimator at one of the extreme ends of $range_X$. This is the ''worst-case'' scenario, since of course, the variance will be greatest at the extreme values of the covariate. In this special case, $x_0 = x_{\min}$ or $x_0 = x_{\max}$, and therefore $|x_0 - \bar{x}| = range_X/2$. Finally, exploiting the independence of $\hat{\mu}_0$ and $s_{Y|X}$ in the case of normality, the variance of the estimator is given by:

$$\begin{aligned} \text{var}(\hat{Q}_0) &= \text{var}(\hat{\mu}_0 + z_p s_{Y|X}) \\ &= \text{var}(\hat{\mu}_0) + \text{var}(z_p s_{Y|X}) \\ &\approx \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) + \left(\frac{\sigma^2}{2n}\right) z_p^2 \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n\sigma_X^2} + \frac{z_p^2}{2n}\right) \\ &= \frac{\sigma^2}{n}\left(4 + \frac{z_p^2}{2}\right) \end{aligned} \tag{6}$$

The width of the 100(1−α)% confidence interval for the 100p% reference limit at the *extreme* of age is therefore:

$$2 z_{1-\frac{\alpha}{2}} \sigma \frac{\sqrt{4 + \frac{z_p^2}{2}}}{\sqrt{n}} \tag{7}$$

Assume we want a relative error of Δ, defined as the ratio of the width of the 100(1−α)% confidence interval for the reference limit to the width of the 100(1−β)% reference range. The width of the 100(1−β)% reference range is given by:

$$2 z_{1-\frac{\beta}{2}} \sigma \tag{8}$$

Thus, if we want the ratio of (7) to (8) to be smaller than the relative error Δ, we require

$$\frac{z_{1-\frac{\alpha}{2}} \sqrt{4 + \frac{z_p^2}{2}}}{z_{1-\frac{\beta}{2}} \sqrt{n}} \leqslant \Delta$$

that is:

$$n \geqslant \frac{z_{1-\frac{\alpha}{2}}^2 \left(4 + \frac{z_p^2}{2}\right)}{z_{1-\frac{\beta}{2}}^2 \Delta^2}$$

It is instructive to work through other sampling strategies. For example, one might take *one-third of the sample at one age extreme, one-third at the midpoint, and one-third at the other age extreme*. In this study design, we have:

$$\begin{aligned} \sum_i (x_i - \bar{x})^2 &= \frac{n}{3}(x_{\min} - \bar{x})^2 + \frac{n}{3}(\bar{x} - \bar{x})^2 + \frac{n}{3}(x_{\max} - \bar{x})^2 \\ &= \frac{n}{3}\left(\frac{1}{4} range_X^2 + 0 + \frac{1}{4} range_X^2\right) \\ &= n \frac{range_X^2}{6} \end{aligned}$$

If we then modify equation (6) accordingly, we have $\text{var}(\hat{Q}_0) = \sigma^2/n(5/2 + z_p^2/2)$, and therefore

$$n \geqslant z_{1-\alpha/2}^2 \left(5/2 + z_p^2/2\right) \Big/ z_{1-\beta/2}^2 \Delta^2.$$

We can also consider the case where the covariate follows a *normal* distribution: $X \sim N$ with variance $\sigma^2{}_X$. In such case, if we assume that the range of the covariate X is approximately $4\sigma_X$, then $\sigma_X^2 \approx range_X^2/16$. Going back to equation (6), the variance of the estimator is then given by:

$$\begin{aligned} \text{var}(\hat{Q}_0) &= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) + \left(\frac{\sigma^2}{2n}\right) z_p^2 \\ &= \sigma^2\left(\frac{1}{n} + \frac{4}{n} + \frac{z_p^2}{2n}\right) \\ &= \frac{\sigma^2}{n}\left(5 + \frac{z_p^2}{2}\right) \end{aligned}$$

Thus, the minimum sample size is given by:

$$n \geqslant \frac{z_{1-\frac{\alpha}{2}}^2 \left(5 + \frac{z_p^2}{2}\right)}{z_{1-\frac{\beta}{2}}^2 \Delta^2}.$$

We can be more conservative, and assume that the approximate range of the covariate is $range_X = {}_X\sigma$ instead of ${}_X\sigma$. In this setting, the minimum sample size requirement becomes:

$$n \geqslant \frac{z_{1-\frac{\alpha}{2}}^2 \left(10 + \frac{z_p^2}{2}\right)}{z_{1-\frac{\beta}{2}}^2 \Delta^2}.$$

Finally, note that we have provided sample size formulas by computing the variance of our estimator at the *extreme value* for the covariate. If on the other hand, one is interested in the 100p% reference limit at the *average value* $\bar{x}$ of the covariate, one can easily see that the term $(x_0 - \bar{x})^2$ in equation (6) vanishes and the sample size requirement becomes:

$$n \geqslant \frac{z_{1-\frac{\alpha}{2}}^2 \left(1 + \frac{z_p^2}{2}\right)}{z_{1-\frac{\beta}{2}}^2 \Delta^2} \tag{9}$$

whatever the covariate distribution. Requirement (9) corresponds to the formula given by Royston [6], and to the usual formulas presented for the ''single-age'' case [9].

Table 2
Values of *factor* according to the sampling strategy (see equation 10)

Distribution of the covariate	*Factor*
1/3 at each end, 1/3 at the midpoint	5/2
Uniform	4
Normal distribution assuming $\text{range}_X \approx 4\sigma_X$	5
Normal distribution assuming $\text{range}_X \approx 6\sigma_X$	10

For the designs investigated, the sample size requirement can be summarized by the following formula:

$$n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}\left(factor + \frac{z_p^2}{2}\right)}{z^2_{1-\frac{\beta}{2}}\Delta^2} \qquad (10)$$

where the variable *factor* depends on the sampling strategy, as summarized in Table 2. We observe that, of the three sampling strategies, selecting a normal distribution for the covariate provides the largest sample size (largest *factor*), whereas the three-subgroup strategy requires the smallest sample.

References

[1] Harrel F. A new distribution-free quantile estimator. Biometrika 1982;69:635–40.
[2] Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46:867–9.
[3] Hutson AD. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. J Appl Stat 1999;26:343–53.
[4] Sheater SJ, Marron JS. Kernel quantile estimators. J Am Stat Assoc 1990;85:410–6.
[5] Healy MJR, Rasbah J, Yang M. Distribution-free estimation of age-related centiles. Ann Hum Biol 1988;15:17–22.
[6] Royston P. Constructing time-specific reference ranges. Stat Med 1991;10:675–90.
[7] Virtanen A, Kairisto V, Uusipaikka E. Regression-based reference limits: determination of sufficient sample size. Clin Chem 1998;44:2353–8.
[8] Elveback LR, Taylor WF. Statistical methods of estimating percentiles. Ann NY Acad Sci 1969;161:538–48.
[9] Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker; 1995.
[10] Linnet K. Two-Stage transformation systems for normalization of reference distributions evaluated. Clin Chem 1987;33:381–6.