Clinical Chemistry 44:3
622–631 (1998)

Laboratory
Management

A robust approach to reference interval estimation and evaluation

Paul S. Horn,[1*] Amadeo J. Pesce,[2] and Bradley E. Copeland[2]

We propose a new methodology for the estimation of reference intervals for data sets with small numbers of observations or for those with substantial numbers of outliers. We propose a prediction interval that uses robust estimates of location and scale. The SAS software can be readily modified to do these calculations. We compared four reference interval procedures (nonparametric, transformed, robust with a nonparametric lower limit, and transformed robust) for sample sizes of 20, 40, 60, 80, 100, and 120 from χ^2 distributions of 1, 4, 7, and 10 *df*. χ^2 distributions were chosen because they simulate the skewness of distributions often found in clinical chemistry populations. We used the root mean square error as the measure of performance and used computer simulation to calculate this measure. The robust estimator showed the best performance for small sample sizes. As the sample size increased, the performance values converged. The robust method for calculating upper reference interval values yields reasonable results. In two examples using real data for haptoglobin and glucose, the robust estimator provides slightly smaller upper reference limits than the other procedures. Lastly, the robust estimator was compared with the other procedures in a population where 5% of the values were multiplied by a factor of 5. The reference intervals were calculated with and without outlier detection. In this case, the robust approach consistently yielded upper reference interval values that were closer to those of the true underlying distributions. We propose that robust statistical analysis can be of great use for determinations of reference intervals from limited or possibly unreliable data.

The concept of a reference interval in medicine is based on determining a set of values within which some percentage, 95% for example, of the values of a particular analyte in a healthy population would fall. This interval is then used for medical decisionmaking. Recommendations on how to obtain such reference intervals have focused on the types of statistics best used to calculate such a reference interval. These parameters then determine the number of individual specimens required to describe the reference interval with a high degree of confidence. Laboratories are mandated by the College of American Pathologists and the Joint Commission for the Accreditation of Health Organizations and Health Care Finance Administration to determine reference intervals for the populations they serve. Currently, NCCLS guidelines recommend samples of 120 individuals for parametric and 200 individuals for nonparametric interval determination. Very often it is not possible to obtain the suggested number of 120 individuals of a specific group to define a reference interval. In some cases, only 20 or 40 individuals in a particular group may be available for study, which is not an ideal population because of the potential for large errors in the resulting estimates. In our own experience, we found it virtually impossible to obtain sufficient numbers to determine the reference interval for the metabolism of the drug lidocaine into its metabolite methylxylidide because it was difficult to get volunteers who were both healthy and willing to have lidocaine injected into them. In addition, the actual cost per each individual test result was on the order of $100 or more for the test reagents. To use this assay to make life-or-death decisions in liver transplant patients, it was very important that we obtain reliable decisionmaking results with a limited number of test specimens. The question then arises as to the best statistical method to calculate the reference interval when limited sample numbers are available. The purpose of this presentation is to show the usefulness of robust statistical analysis for obtaining a good estimate of reference intervals with a small number of samples. In this presentation, we present the theoretical background for

Departments of [1] Mathematical Sciences and [2] Pathology and Laboratory Medicine, University of Cincinnati, Cincinnati, OH 45221.

[*]Address correspondence to this author at: Department of Mathematical Sciences, University of Cincinnati, PO Box 210025, Cincinnati, OH 45221-0025. Fax 513-556-3417; e-mail paul.horn@uc.edu.

Received September 12, 1997; revision accepted December 12, 1997.

using this approach. We look at cases where 20 samples are available.

CURRENT APPROACHES

There are two traditional approaches to the derivation of reference intervals. The first is nonparametric and is based on the sample quantiles. For example, if the (central) 90% reference interval is required, then the 5th and 95th sample quantiles are used. A better approach is to use the distribution-free quantile estimators described by Harrel and Davis *[1]*. This estimator is essentially a bootstrapped *[2]* version of the traditional sample quantile. The estimator of the p^{th} quantile is as follows:

$$Q_p = \sum_{i=1}^{n} W_{n,i} \cdot x_{(i)}$$

where

$$W_{n,i} = I_{i/n}\{p(n+1), (1-p)(n+1)\} - I_{(i-1)/n}\{p(n+1), (1-p)(n+1)\}$$

where $I_x(\alpha,\beta)$ is the incomplete beta function and $x_{(1)} \leq \ldots \leq x_{(n)}$ are the observed order statistics. The Harrel and Davis estimator has been recommended as the nonparametric method of choice for the derivation of reference intervals *[3]*. Therefore, it is this version of the nonparametric method that will be examined in this study.

The second approach to deriving reference intervals is based on transforming the data to achieve normality, computing the appropriate quantile estimators using normal theory, and back-transforming to the original scale. The transformation used in this study is described by Harris and Boyd *[3]*. Briefly, an initial transformation removes skewness:

$$y = \begin{cases} (x^{\lambda} - 1)/\lambda \text{ for } \lambda \neq 0 \\ log(x + c) \text{ for } \lambda = 0 \end{cases}$$

This transformation was introduced by Box and Cox *[4]*. Here, the maximum likelihood estimator of λ, $\hat{\lambda}$, is computed from the original x-data. If $|\lambda| < 0.10$, then the $\log(x + c)$ transformation is used, and $\hat{c}$, the maximum likelihood estimator of c, is then computed.

Once the initial transformation is fit, a second transformation is derived to remove any remaining kurtosis. The y values in the previous equation are standardized to have zero mean and unit variance. Then a constant, K, is determined so that:

$$z = sign(y) \cdot |y|^{K}$$

has kurtosis = 0. The power of the transform that is actually used is $(K + 1)/2$ *[5]*. The z-data are then tested for normality by using the Anderson–Darling statistic at significance level 0.15 *[6]*. If the null hypothesis of normality is not rejected, then the traditional normal quantile estimates are used on the z-data, namely:

$$\bar{z} \pm z(1 - \alpha/2) \cdot s_z \tag{1}$$

where $\bar{z}$ and s_z are the sample mean and SD of the z-data, and $z(1 - \alpha/2)$ is the appropriate standard normal quantile. Therefore, for a 90% reference interval, $\alpha = 0.10$, and $z(0.95) = 1.645$ are used. These two estimates are then back-transformed to the y-data scale and, finally, the original x-data scale. On the other hand, if normality is rejected, then the nonparametric (Harrel and Davis) reference interval is used.

We end this section by noting that the reference interval can be viewed as a prediction interval based on the random sample $X_1, \ldots, X_n$ for the next observation, X_{n+1}. If the underlying population is normal, then the random variable:

$$\frac{X_{n+1} - \bar{X}}{s\sqrt{1 + 1/n}}$$

has a Student's t-distribution with $(n - 1)$ *df*. Thus, the appropriate $(1 - \alpha)$ 100% reference interval is equal to:

$$\bar{x} \pm t_{n-1}(1 - \alpha/2)\, s\sqrt{1 + 1/n} \tag{2}$$

where $t_{n-1}(1 - \alpha/2)$ is the appropriate quantile from a Student's t-distribution with $(n - 1)$ *df*.

Clearly, for large samples the reference intervals defined in Eqs. 1 and 2 are approximately equal. The 90–95% reference intervals defined by Eq. 2 are ~8% wider than those defined by Eq. 1 for $n = 20$ and only ~1% wider for $n = 100$. However, we will use the reference interval defined by Eq. 1 on the transformed data, even for small samples, because it is more prevalent in the clinical chemistry literature. We did examine the reference interval based on Eq. 2, as well as the interval based on the uniform minimum variance unbiased estimators of the quantiles. Neither of these performed well enough to replace the interval defined by Eq. 1 for this study.

ROBUST PREDICTION INTERVALS AND QUANTILE ESTIMATORS

As noted in the previous section, the $(1 - \alpha)$ 100% prediction interval for the next observation, X_{n+1} given an observed random sample $X_1 = x_1, \ldots, X_n = x_n$ has the form given by Eq. 2. Horn *[7]* pointed out that Eq. 2 can be written as:

$$\bar{x} \pm t_{n-1}(1 - \alpha/2)\sqrt{s^2 + s^2/n} \tag{3}$$

In this way, the two components on variation are s^2, the variance of the unknown observation X_{n+1}, and s^2/n, the variance of $\bar{x}$, the estimated center of the interval.

Horn *[7]* proposed a prediction interval replacing the three estimates, $\bar{x}$, s^2, and s^2/n, by robust estimates of location and scale. Specifically, the $(1 - \alpha)$ 100% biweight prediction interval for symmetric populations is defined as follows:

$$T_{bi}(c_1) \pm t_{n-1}(1 - \alpha/2)\,[S_T^2(c_1) + s_{bi}^2(c_2)]^{1/2} \tag{4}$$

where $T_{bi}(c_1)$ is the biweight location estimator with tuning constant c_1, $S_T^2(c_1)$ is the biweight estimator of the variability of $T_{bi}(c_1)$, and $s_{bi}(c_2)$ is the biweight estimator of spread with tuning constant c_2 *[8]*. Briefly, T_{bi} is the solution to the equation:

$$\sum_{i=1}^{n} \psi(u_{\mathrm{i}}) = 0 \tag{5}$$

where

$$\psi(u) = \begin{cases} u(1-u^2)^2, & \text{for } |u| < 1 \\ 0, & \text{elsewhere} \end{cases}$$

$$u_i = \frac{(x_i - T_{bi})}{c \cdot s*}$$

$$s* = \text{estimate of spread}$$

$$c = \text{tuning constant}$$

The term $\psi(u)$ may be rewritten as $\psi(u) = u \cdot w(u)$, where $w(\cdot)$ is a weight function. Making this substitution in Eq. 5 and solving yields $T_{bi} = \Sigma_i[x_i\, w(u_i)]/\Sigma w(u_i)$. Thus, T_{bi} is a weighted mean with weights that decrease as u_i goes from 0 to ±1; equivalently, as an observation x_i goes from the center T_{bi} to $T_{bi} \pm c \cdot s*$, its weight decreases. If an observation is more than $c \cdot s*$ from the center T_{bi}, it gets weight zero. For example, if $c = 6$ and $s* = s$, the sample SD, then observations >6 SD from the center get weight zero.

Equation 5 defines a class of different estimators; each estimator is the solution based on a specific ψ function. For example, if $\psi(u) = u$, all observations get equal weight, and the solution is $T = \bar{x}$. However, in the case of biweight $\psi(\cdot)$, the solution T_{bi} is computed iteratively, starting with the sample median. The iteration is necessary because T_{bi} is a weighted mean with weights that depend on (a previously computed) T_{bi}.

A popular class of estimators of spread is based on variance estimates of robust estimators of location. For estimators based on Eq. 5, the asymptotic variance is simply $E(\psi^2)/[E(\psi')]^2$, where $E(\cdot)$ denotes mathematical expectation. The variance estimate, S^2_{ψ}, simply replaces mathematical expectation with empirical averaging. For the biweight ψ function, we use $S^2_T(c)$ to denote this estimate of $Var[T_{bi}(c)]$ based on the tuning constant, c. Because this variance estimate is essentially a standard error squared, it goes to 0 by order n. Thus, a reasonable estimate of spread is $S(c) = \sqrt{n}$ times the square root of variance estimate of $T_{bi}(c)$.

The actual value used for s_{bi} in the iteration of T_{bi} is slightly different from that given above. We follow the modified formula given by Kafadar *[8]*. Specifically, s_{bi} is computed by using the biweight function, $\psi(\cdot)$, but with the sample median used for location and MAD/0.6745 (the median absolute deviation about the median) used as an estimate for scale. (The factor of 0.6745 is included so that MAD/0.6745 is consistent for σ in the gaussian case.) The s_{bi} used in Eq. 4 is computed in the same manner; the only difference is the value of the tuning constant. For details, see Horn *[7]* and Kafadar *[8]*. Simple SAS code, which can be modified for most languages, is included in the *Appendix* to this report.

The tuning constant c_1 is set equal to 3.7, which means that, for the purposes of location estimation, observations are down-weighted (smoothly) the further they lie from the center (i.e., the current value of T_{bi} in the iteration procedure). Any observations that are more than ~3.7 SD from the center get zero weight. The tuning constant c_2, on the other hand, is a function of the value of the prediction interval $(1 - \alpha)$. Specifically, $c_2 = [0.58173 - 0.607227(1 - \alpha)]^{-1}$ for $0.05 \leq \alpha \leq 0.5$. Thus, for 90% reference intervals, $\alpha = 0.10$ and $c_2 = 28.4$, and for 95% reference intervals, $\alpha = 0.10$ and $c_2 = 205.4$ *[7]*.

We intend to examine the performance of this robust prediction interval after the Box–Cox transformation to symmetry. Because it was designed to accommodate possibly heavy-tailed distributions, the power transform to remove any residual kurtosis is not required.

Another candidate for a robust reference interval uses the robust quantile estimator for skewed populations as its upper endpoint *[9]*. This quantile estimator is based on the robust prediction described above. The idea is to examine only data points greater than the sample median. Then a symmetric pseudo-sample is created by including all data points greater than the sample median *and* their pseudo-values that are equidistant *less* than the median. For example, if $n = 20$, and the data are ordered $x_1 < \ldots < x_{20}$, then the median, $M = (x_{10} + x_{11})/2$ and the symmetric pseudo-sample, is:

$$2M - x_{20} < 2M - x_{19} < \ldots < 2M - x_{11}$$
$$(= x_{10}) < x_{11} < \ldots < x_{20}$$

From this sample, the appropriate *symmetric* prediction interval is computed as before, and the upper endpoint is used as the upper limit (quantile) on the reference interval. See Horn *[9]* for details.

The analogous lower robust quantile is not used, because in most cases the underlying populations are positively skewed, and thus the median will be greater than the mode. Reflected pseudo-samples in these cases, although symmetric, will be indicative of underlying bimodal populations *[9]*. Thus, for the lower endpoint, we will use the nonparametric estimator (Harrel and Davis) because it also does not require transformation of the data.

simulation and assessment

To evaluate the four reference interval procedures (nonparametric, transformed, robust with nonparametric lower limit, and transformed robust), a simulation study was run. Random samples of size 20, 40, 60, 80, 100, and 120 were generated from each of four χ^2 distributions with *df* 1, 4, 7, and 10, respectively. The usual measure of performance is the root mean square error (RMSE) for each of the endpoints that constitute the reference interval. Specifically, the RMSE of the upper endpoints of a particular (1 − α) 100% reference interval, for example, is as follows:

$$\text{RMSE} = \{E[\hat{F}^{-1}(1 - \alpha/2) - F^{-1}(1 - \alpha/2)]^2\}^{1/2}$$

where $\hat{F}^{-1}(1 - \alpha/2)$ is the estimate of the true endpoint (quantile) $F^{-1}(1 - \alpha/2)$. This value is estimated via simulation by:

$$\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{F}_i^{-1}[(1 - \alpha/2) - F^{-1}(1 - \alpha/2)]^2\right\}^{1/2}$$

where $\hat{F}_i^{-1}$ (1 − α/2) is the endpoint of the reference range derived from the i^{th} random sample, and N is the number of random samples in the simulation; here $N =$ 1000.

The RMSEs of the lower and upper endpoints of 90% and 95% reference intervals are given in Tables 1 and 2, respectively. For the upper endpoint of 90% reference intervals, the robust quantile estimator (untransformed) achieves the smallest RMSE, especially for smaller sample sizes and the more skewed populations (fewer *df*). However, it is essentially equal in performance to the transformed traditional procedure for $n \geq 40$. For the lower endpoint, the RMSEs of the two transformed procedures (traditional and robust) are about equal and slightly better than the RMSE of the nonparametric, which is also used by the untransformed robust procedure. For 95% reference intervals, however, the robust procedure is clearly best, especially for the smaller sample sizes ($n \leq 40$). For the larger sample sizes, the transformed procedures again are about equal, with the robust slightly better for the more skewed populations and the traditional (normal theory) procedure slightly better for the more symmetric populations (more *df*). For the lower endpoints, the transformed procedures again are about equal and only slightly (5–10%) better than the nonparametric procedure.

Traditionally, assessment of reference interval limits has focused on the RMSE of the interval endpoints as described above. This certainly makes sense if the interval endpoints are used as targets for treatment. For example, suppose the endpoint of the 95% reference interval for creatine kinase for middle-aged women is 192 U/L, as derived by a particular laboratory. Physicians who use this laboratory may evaluate their patients who have

Table 1. RMSE of 90% reference interval endpoints.

	Nonparametric		Robust		Transformed		Transformed robust	
Sample size	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
χ^2 with 1 *df*								
20	0.024	1.721	0.024	1.393	0.020	1.766	0.015	2.922
40	0.012	1.140	0.012	0.877	0.009	0.996	0.007	0.967
60	0.007	0.986	0.007	0.778	0.006	0.892	0.005	0.815
80	0.006	0.832	0.006	0.669	0.005	0.815	0.005	0.690
100	0.004	0.729	0.004	0.604	0.004	0.723	0.004	0.617
120	0.004	0.659	0.004	0.549	0.004	0.660	0.004	0.565
χ^2 with 4 *df*								
20	0.379	2.188	0.379	1.938	0.391	1.987	0.353	2.800
40	0.238	1.701	0.238	1.393	0.247	1.366	0.243	1.531
60	0.192	1.311	0.192	1.112	0.194	1.104	0.196	1.161
80	0.172	1.197	0.172	1.014	0.168	0.987	0.170	1.028
100	0.153	1.002	0.153	0.827	0.147	0.851	0.152	0.820
120	0.140	0.968	0.140	0.827	0.132	0.809	0.136	0.791
χ^2 with 7 *df*								
20	0.696	2.685	0.696	2.414	0.696	2.351	0.667	3.070
40	0.468	2.022	0.468	1.690	0.476	1.682	0.459	1.987
60	0.398	1.568	0.398	1.342	0.389	1.339	0.378	1.450
80	0.338	1.329	0.338	1.136	0.324	1.116	0.322	1.155
100	0.316	1.271	0.316	1.087	0.293	1.079	0.295	1.092
120	0.275	1.103	0.275	0.944	0.253	0.955	0.256	0.954
χ^2 with 10 *df*								
20	0.970	2.783	0.970	2.517	0.963	2.409	0.929	3.061
40	0.679	2.180	0.679	1.828	0.663	1.795	0.647	2.009
60	0.552	1.780	0.552	1.523	0.519	1.532	0.517	1.606
80	0.484	1.512	0.484	1.302	0.452	1.267	0.455	1.324
100	0.430	1.390	0.430	1.206	0.384	1.184	0.387	1.229
120	0.406	1.202	0.406	1.012	0.363	1.007	0.361	1.022

Table 2. RMSE of 95% reference interval endpoints.

Sample size	Nonparametric		Robust		Transformed		Transformed robust	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
χ^2 with 1 *df*								
20	0.017	2.075	0.017	1.802	0.013	3.615	0.011	8.961
40	0.006	1.605	0.006	1.192	0.005	1.537	0.004	1.609
60	0.004	1.462	0.004	1.078	0.003	1.355	0.003	1.282
80	0.003	1.281	0.003	0.944	0.002	1.199	0.002	0.971
100	0.002	1.124	0.002	0.887	0.002	1.081	0.002	0.883
120	0.002	1.022	0.002	0.825	0.002	1.009	0.002	0.830
χ^2 with 4 *df*								
20	0.435	2.637	0.435	2.315	0.396	2.895	0.347	5.366
40	0.231	2.276	0.231	1.703	0.240	1.878	0.235	2.304
60	0.177	1.891	0.177	1.385	0.186	1.508	0.189	1.662
80	0.150	1.777	0.150	1.268	0.158	1.368	0.163	1.457
100	0.136	1.393	0.136	1.082	0.140	1.118	0.148	1.105
120	0.128	1.377	0.128	1.075	0.130	1.100	0.137	1.069
χ^2 with 7 *df*								
20	0.821	3.243	0.821	2.880	0.755	3.248	0.734	5.184
40	0.516	2.640	0.516	2.008	0.510	2.322	0.491	2.993
60	0.420	2.292	0.420	1.638	0.418	1.839	0.402	2.060
80	0.343	1.951	0.343	1.394	0.338	1.503	0.336	1.595
100	0.321	1.764	0.321	1.347	0.308	1.423	0.311	1.471
120	0.278	1.564	0.278	1.204	0.267	1.290	0.272	1.311
χ^2 with 10 *df*								
20	1.169	3.403	1.169	2.989	1.059	3.231	1.041	5.070
40	0.754	2.827	0.754	2.178	0.727	2.402	0.711	2.926
60	0.599	2.527	0.599	1.852	0.571	2.047	0.571	2.219
80	0.514	2.187	0.514	1.562	0.489	1.693	0.494	1.800
100	0.465	1.963	0.465	1.466	0.426	1.580	0.429	1.679
120	0.437	1.646	0.437	1.270	0.402	1.339	0.398	1.380

concentrations in excess of this value to determine the cause. In this case, clearly, the value of the endpoint of the reference interval itself is vital, and its accuracy (RMSE) is vital for assessment of a procedure.

On the other hand, the reference interval is designed to include (or exclude) a specified percentage of the underlying population. It could be argued that, in fact, it is this percentage that should be evaluated. Specifically, we will now consider the RMSE of the percentage as estimated by the lower and upper endpoints of the reference interval. Here, the RMSE of the upper probabilities, for example, is as follows:

$$\text{RMSE of probability} = (\mathrm{E}\{\mathrm{F}[\hat{\mathrm{F}}^{-1}(1-\alpha/2)] - (1-\alpha/2)\}^2)^{1/2}$$

which is estimated from the simulation by:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\{F[\hat{F}_{\mathrm{i}}^{-1}(1-\alpha/2)] - (1-\alpha/2)\}^2\right)^{1/2},$$

where $F[\hat{F}_i^{-1}(1-\alpha/2)]$ is the actual (unknown, in practice) proportion of the population less than the upper limit of the reference interval from the i^{th} simulated random sample.

The RMSE for the lower and upper probabilities of 90% and 95% reference interval limits are not presented here because all procedures achieved roughly the same RMSE, although that of the robust upper probability limit was slightly smaller for $n = 20$. One particularly interesting fact is that the transformed robust procedure, which appeared to perform poorly (especially for small samples) as an upper endpoint estimator, should do so well in terms of the RMSE of the probability. This phenomenon may be explained by the first few terms of the Taylor expansion of the mean square error (MSE) of the probability. Specifically, if we expand *MSE* $\{F[\hat{F}^{-1}(p)]\}$ ($p = 1-\alpha/2$ for brevity) about the true quantile F^{-1}, we get the following:

$$E\{F[\hat{F}^{-1}(p)] - p\}^2 = E[\hat{F}^{-1}(p) - F^{-1}(p)]^2 \cdot f^2[F^{-1}(p)] + E[\hat{F}^{-1}(p) - F^{-1}(p)]^3 \cdot f[F^{-1}(p)] \cdot f'[F^{-1}(p)],$$

where $f(\cdot) = F'(\cdot)$, the underlying population density.

If we examine only the first (nonzero) term of the Taylor expansion, we see that the MSE (and thus the RMSE) for the probability is proportional to that of its corresponding endpoint. However, the second term shows that the upper limit estimators, which are positively skewed with respect to the true quantile, will benefit in terms of MSE for the probability. This is because, in general, $f'[F^{-1}(p)] < 0$ for the upper limits. (These results are not surprising because the probability contained between an upper quantile and a one-unit shift to the right is less than that of a one-unit shift to the left.)

Although rewarding upper limits that tend to be skewed toward larger values may be the conservative thing to do statistically (e.g., if we state "$p\% < x$", we want at least p%), it could be disastrous in the context of a medical reference interval, where large values of the analyte in question are indicative of a possibly adverse health condition. To equalize the MSE loss, we introduce a factor to be multiplied by the difference $\{F[\hat{F}_i^{-1}(1 - \alpha/2)] - (1 - \alpha/2)\}$ before squaring for those samples where $[\hat{F}_i^{-1}(1 - \alpha/2)] > 1 - \alpha/2$. We will use as this factor the ratio of probabilities to the left and right of the upper quantile. Thus, for the upper limit, this factor on the true difference in probabilities will be $(1 - \alpha/2)/(\alpha/2)$, when the probability contained by the upper limit of the reference interval exceeds the nominal, target value. The same factor is used for lower limits when their true probabilities are *less* than the target value.

The weighted RMSEs of the probabilities for 90% reference intervals, where the above factors premultiply the differences before the squaring operation, are presented in Table 3. Essentially, all of the procedures are equivalent, with a slight edge going to the robust approach for the most skewed population and to the transformed traditional approach for the others. One fact to note is that the transformed robust is worst for $n = 20$, as it was for the interval endpoints. The results for 95% reference intervals are provided in Table 4. In this situation, however, the robust procedure for the upper limit is clearly superior in every case. Of particular interest is that the robust method does very well compared with the other methods for larger sample sizes. This indicates that the robust upper limit is a reasonable procedure for large as well as small samples.

EXAMPLES

As a first example, we examine the haptoglobin data as given in Harris and Boyd *[3]*. For these 100 values, the 95% reference intervals (i.e., the 2.5 and 97.5 percentiles)

Table 3. Weighted RMSE of 90% reference interval: upper and lower probabilities.

	Nonparametric		Robust		Transformed		Transformed robust	
Sample Size	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
χ^2 with 1 *df*								
20	0.137	0.385	0.137	0.328	0.342	0.335	0.378	0.411
40	0.149	0.346	0.149	0.239	0.376	0.246	0.420	0.244
60	0.147	0.305	0.147	0.221	0.347	0.232	0.399	0.202
80	0.150	0.279	0.150	0.203	0.295	0.247	0.344	0.192
100	0.150	0.247	0.150	0.182	0.257	0.226	0.308	0.169
120	0.144	0.230	0.144	0.172	0.190	0.220	0.237	0.167
χ^2 with 4 *df*								
20	0.283	0.371	0.283	0.349	0.273	0.327	0.378	0.446
40	0.277	0.350	0.277	0.275	0.252	0.251	0.330	0.313
60	0.253	0.298	0.253	0.232	0.231	0.209	0.293	0.250
80	0.224	0.279	0.224	0.223	0.204	0.196	0.258	0.225
100	0.210	0.237	0.210	0.180	0.193	0.162	0.243	0.176
120	0.194	0.234	0.194	0.186	0.178	0.158	0.225	0.170
χ^2 with 7 *df*								
20	0.322	0.370	0.322	0.362	0.279	0.325	0.384	0.444
40	0.283	0.352	0.283	0.290	0.224	0.272	0.285	0.342
60	0.246	0.302	0.246	0.241	0.197	0.224	0.243	0.274
80	0.227	0.266	0.227	0.213	0.189	0.195	0.229	0.229
100	0.223	0.251	0.223	0.203	0.183	0.187	0.218	0.211
120	0.195	0.228	0.195	0.180	0.162	0.167	0.195	0.186
χ^2 with 10 *df*								
20	0.325	0.363	0.325	0.360	0.285	0.303	0.377	0.432
40	0.302	0.345	0.302	0.280	0.236	0.254	0.289	0.320
60	0.265	0.291	0.265	0.235	0.205	0.219	0.248	0.260
80	0.243	0.271	0.243	0.222	0.188	0.203	0.226	0.237
100	0.215	0.251	0.215	0.207	0.167	0.192	0.199	0.220
120	0.206	0.219	0.206	0.170	0.159	0.156	0.186	0.182

Table 4. Weighted RMSE of 95% reference interval: upper and lower probabilities

Sample size	Nonparametric		Robust		Transformed		Transformed robust	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
χ^2 with 1 *df*								
20	0.270	0.963	0.270	0.925	0.886	1.100	0.914	1.340
40	0.353	1.084	0.353	0.633	1.002	0.934	0.895	0.813
60	0.377	1.087	0.377	0.568	0.979	0.920	0.866	0.644
80	0.420	1.069	0.420	0.496	0.831	0.958	0.827	0.532
100	0.440	0.979	0.440	0.436	0.770	0.911	0.651	0.439
120	0.440	0.924	0.440	0.387	0.671	0.888	0.634	0.395
χ^2 with 4 *df*								
20	0.760	0.923	0.760	1.042	0.974	1.144	1.443	1.592
40	0.894	1.082	0.894	0.784	0.962	0.889	1.321	1.171
60	0.868	1.071	0.868	0.631	0.912	0.749	1.200	0.952
80	0.796	1.079	0.796	0.602	0.844	0.724	1.099	0.849
100	0.768	0.926	0.768	0.454	0.823	0.568	1.062	0.666
120	0.727	0.921	0.727	0.475	0.783	0.583	1.003	0.644
χ^2 with 7 *df*								
20	0.857	0.933	0.857	1.113	0.990	1.125	1.453	1.585
40	0.906	1.105	0.906	0.863	0.825	0.971	1.118	1.261
60	0.889	1.088	0.889	0.700	0.753	0.828	0.973	1.044
80	0.855	1.035	0.855	0.593	0.724	0.721	0.919	0.884
100	0.849	0.966	0.849	0.556	0.723	0.678	0.895	0.799
120	0.750	0.896	0.750	0.482	0.642	0.630	0.805	0.728
χ^2 with 10 *df*								
20	0.872	0.894	0.872	1.124	0.971	1.055	1.391	1.577
40	0.956	1.083	0.956	0.848	0.841	0.916	1.106	1.203
60	0.964	1.064	0.964	0.694	0.776	0.799	0.987	0.980
80	0.913	1.065	0.913	0.645	0.720	0.760	0.903	0.905
100	0.843	0.981	0.843	0.592	0.656	0.714	0.813	0.847
120	0.797	0.872	0.797	0.456	0.627	0.601	0.755	0.723

were computed. For each point estimator of a percentile, a 90% confidence interval is provided. The confidence interval for the transformed procedure made use of the formula, (percentile estimate) $\pm u_{(1 + \beta)/2}[(2 + c_1^2 - \alpha)s_x^2/2N]^{1/2}$, where s_x is the sample SD of the transformed data, N is the sample size, α defines the quantiles of interest, and β defines the confidence level of the interval for each of the point estimators *[10]*. In our case, $\alpha = 0.025$ and $\beta = 0.90$.

The confidence intervals for the other methods were derived by using the bootstrap methodology *[2]*. Here, 200 samples were drawn with replacement (i.e., resampled from the observed data), yielding 200 reference intervals for each methodology. From these values, the observed 5th and 95th quantiles were used as a 90% confidence interval.

The results for the haptoglobin data are given in the top of Table 5. All of the methods are reasonably consistent. The transformed methods have a lower quantile estimator about two units larger than that of the nonparametric. The confidence interval for the upper endpoint based on the bootstrapped robust method is ~1% tighter than that based on the transformation approach.

As a second example, we compute similar statistics for blood glucose concentrations (mmol/L) in samples obtained in our laboratory from 46 men, ≥80 years of age. The data are as follows:

3.520 3.905 4.070 4.070 4.290 4.345 4.400 4.455 4.565
4.620 4.620 4.675 4.840 4.840 4.895 4.895 4.950 4.950
5.115 5.115 5.225 5.225 5.225 5.335 5.335 5.390 5.390
5.390 5.455 5.555 5.610 5.665 5.720 5.775 5.830 5.830
5.885 5.885 6.215 7.095 7.205 8.140 9.900 10.890 11.605
12.045

The results are given in the bottom of Table 5. We note that, in this case, no suitable transformation to normality was found; therefore, only the nonparametric and robust procedures appear. From Table 5, we see again that the upper quantile estimator provided by the robust procedure is tighter than that of the nonparametric. Note that the confidence intervals of both upper quantile estimators lie entirely in the range defined as diabetic (>7.7 mmol/L or 1.4 g/L) by the American Diabetic Association *[11]*.

Table 5. 95% reference interval endpoints (with 90% confidence intervals).

Analyte	Lower endpoint	Upper endpoint
Haptoglobin (n = 100)		
Nonparametric	23.3 (17.0–33.9)	194.6 (178.1–214.1)
Robust	23.3 (17.0–33.9)	190.5 (173.1–206.8)
Transformed	25.8 (19.6–32.9)	191.1 (174.5–208.4)
Transformed robust	25.1 (18.4–34.2)	192.6 (175.6–210.4)
Glucose (n = 46)		
Nonparametric	3.7 (3.5–4.2)	11.6 (9.2–12.0)
Robust	3.7 (3.5–4.2)	9.7 (7.7–11.0)

OUTLYING OBSERVATIONS

Until now, we have assumed that all of the data come from a homogeneous population and that any large aberrant values are also part of that population. However, in practice, real data are subject to contamination from a variety of sources, such as human error or the presence of disease in an individual. Simulation results for the upper limit of 90% and 95% reference intervals in the presence of outliers are given in Table 6. In this case, the outliers comprise 5% of the sample, and they are derived by multiplying a valid observation by a factor of 5. Except for a few isolated cases, the robust methods are best with respect to RMSE for the upper interval endpoint; the transformed robust method is slightly better for $n \geq 60$. All of the methods “broke down”, however, in the sense that their RMSEs were at least an order of magnitude larger than those with uncontaminated data (Tables 1 and 2). Nevertheless, the robust methods were more resistant. (Note that, although the RMSEs generally decrease as the sample size increases, the decrease is not exactly monotone, as was the case without outliers. This is because the contamination of the samples introduced more noise to the simulation.)

The use of outlier detection is not routinely recommended for reference interval analysis because large values from a skewed population may be mislabeled as outliers *[11]*. However, for completeness, Table 7 presents results based on the same data as Table 6 but with a simple outlier detection method on the original data; any value >3.5 SD away from the mean is ignored. One thing is clear from Table 7—the drastic improvement of all the

Table 6. RMSE of upper 90% and 95% reference interval endpoints: 5% outliers (×5).

Sample size	Nonparametric		Robust		Transformed		Transformed robust	
	90%	95%	90%	95%	90%	95%	90%	95%
χ^2 with 1 *df*								
20	4.298	5.281	2.178	4.004	3.697	7.089	5.012	17.144
40	3.038	5.378	1.616	3.166	1.996	3.839	1.967	4.178
60	2.230	5.346	1.449	3.228	1.805	4.091	1.553	3.191
80	1.760	4.567	1.336	2.994	1.547	3.870	1.290	2.800
100	1.480	3.950	1.225	2.915	1.367	3.412	1.162	2.481
120	1.354	3.480	1.178	2.810	1.308	3.125	1.149	2.475
χ^2 with 4 *df*								
20	10.976	13.775	6.042	10.527	6.631	11.878	8.379	23.683
40	8.429	15.006	5.181	9.148	5.432	10.018	5.384	10.276
60	6.660	14.571	4.940	8.793	4.903	9.561	4.707	8.470
80	5.563	13.332	4.822	8.476	4.572	9.147	4.452	7.777
100	4.669	12.110	4.603	8.087	4.161	8.647	4.140	7.282
120	4.507	11.659	4.664	8.254	4.137	8.767	4.101	7.445
χ^2 with 7 *df*								
20	18.562	23.619	10.299	18.049	10.886	18.660	11.751	27.552
40	14.899	25.195	9.101	15.709	9.453	16.879	9.024	16.930
60	11.515	23.994	8.287	14.589	8.487	17.000	7.911	14.523
80	10.000	22.833	8.195	14.258	8.156	17.164	7.621	13.767
100	9.090	21.863	8.109	14.092	7.930	17.214	7.453	13.365
120	8.500	21.188	8.076	13.984	7.664	17.240	7.377	13.285
χ^2 with 10 *df*								
20	24.180	30.996	13.155	23.549	13.146	22.134	13.644	28.918
40	20.855	34.828	11.933	21.821	13.088	23.299	11.167	21.467
60	16.874	33.509	11.190	20.413	12.323	24.183	10.269	19.339
80	15.465	33.646	11.620	20.822	12.540	25.847	10.639	19.645
100	13.722	31.621	11.328	20.180	11.762	25.456	10.336	19.056
120	12.627	30.379	11.003	19.650	11.243	25.345	10.121	18.577

Table 7. RMSE of upper 90% and 95% reference interval endpoints: 5% outliers (×5)—with outlier detection.

Sample	Nonparametric		Robust		Transformed		Transformed robust	
size	90%	95%	90%	95%	90%	95%	90%	95%
χ^2 with 1 *df*								
20	1.780	2.083	1.600	1.985	2.006	4.116	3.581	11.169
40	1.212	1.687	0.980	1.411	1.059	1.577	1.014	1.667
60	1.063	1.461	0.856	1.222	0.980	1.388	0.841	1.263
80	0.903	1.252	0.726	1.064	0.880	1.191	0.729	1.069
100	0.833	1.193	0.683	1.018	0.818	1.146	0.674	0.997
120	0.746	1.100	0.608	0.930	0.752	1.096	0.613	0.908
χ^2 with 4 *df*								
20	3.474	3.758	2.969	3.490	3.031	4.650	4.506	8.893
40	2.909	4.116	2.127	2.518	2.080	3.141	2.549	4.187
60	2.525	4.245	1.940	2.324	1.891	2.989	2.152	3.367
80	2.282	3.990	1.862	2.142	1.752	2.711	1.936	2.921
100	1.999	3.648	1.709	1.936	1.638	2.521	1.792	2.701
120	2.065	3.857	1.840	2.099	1.741	2.933	1.860	2.772
χ^2 with 7 *df*								
20	4.702	5.247	3.889	4.684	3.767	5.554	5.166	9.036
40	4.800	7.918	3.488	4.518	3.503	5.823	4.129	6.840
60	3.987	7.456	3.247	4.048	3.110	5.144	3.496	5.539
80	3.605	7.312	3.206	3.992	3.040	5.292	3.328	5.181
100	3.397	7.285	3.270	4.052	3.000	5.371	3.200	4.844
120	3.030	6.668	3.087	3.760	2.805	4.968	3.028	4.632
χ^2 with 10 *df*								
20	5.554	6.282	4.531	5.473	4.253	6.308	5.811	9.921
40	6.503	11.826	4.666	6.588	4.658	8.269	5.116	8.538
60	5.464	11.268	4.584	6.195	4.385	8.087	4.768	7.687
80	4.933	11.123	4.707	6.247	4.417	8.514	4.810	7.576
100	4.497	10.649	4.652	6.053	4.175	8.001	4.580	7.099
120	4.003	10.053	4.420	5.768	3.959	8.089	4.402	6.798

methods. Nevertheless, the robust method maintains its superiority in virtually every situation.

Although not presented here, results for the weighted RMSE of the upper probabilities do not contradict the above results. Without outlier detection, all methods are essentially the same, with the robust method having a slight advantage for the most skewed population. With outlier detection, all methods improve, but the robust method becomes clearly the best in every situation.

Discussion

The need to derive reference ranges from samples where the number of observed data values is small, for example, $20 \leq n \leq 60$, clearly exists. We show by the simulation study presented here that the RMSE calculated by the robust quantile estimator was the smallest for upper endpoints calculated on small sample sizes. However, when evaluating the upper and lower probabilities for the 90% and 95% reference intervals, we showed that the losses in over- vs underestimating should not be treated symmetrically. To equalize the MSE loss, a weighting factor was introduced. In this case, the robust statistic was superior for estimating the upper limit of the 95% reference interval and about equal to the transformed traditional interval for estimating the 90% reference interval. When real serum haptoglobin data were examined in this fashion, the robust estimator of the 97.5 percentile limit was smaller than that of the nonparametric estimator and comparable with the estimator based on transformation. A second example, using glucose data from 46 elderly men, showed a similar result. Thus, it is reasonable to propose that robust estimators can provide relevant reference intervals when only small numbers of samples are available. Furthermore, if it is suspected that outliers may exist, then the robust method should do as well as, if not better than, other methods, whether or not outlier detection is used. However, because none of the procedures did particularly well when confronted with severe contamination, we cannot overstate the importance of ensuring the quality of the data and the data collection process when determining reference intervals.

In summation, we recommend that nonparametric, robust, and normal theory (on transformed data) reference intervals be computed in practice. If the methods are in agreement, then any one will do reasonably well for reporting purposes. However, if the methods disagree, then we believe that the tightest interval should be used. The reason we recommend this is that, given the choice between reasonable, though disparate, reference intervals, we would prefer to err on the side of more false positives, rather than false negatives, thus forcing the clinician to further evaluate the patient. Finally, if the sample size is so small that it precludes reasonable nonparametric confidence intervals for the limits, or if a suitable transformation to achieve normality is not possible, then the proposed robust method should be used, at least for the upper endpoint of the reference interval.

Appendix

**** The ordered data are in the ARRAY Y; the sample size, N, is even. (Minor modifications are necessary for when N is odd.)

```
M=N/2;
M1=M+1;
C=3.7;
MEDIAN=(Y(M)+Y(M1))/2;
IF MOD(M,2)=0 THEN MAD=((Y(M+M/2)+Y(M+1+
  M/2))/2)-MEDIAN; ELSE MAD=Y(M+(M+1)/2-
  MEDIAN; MAD=MAD/.6745;
S=MAD;
S1=0; S2=0;
DO J=M1 TO N;
U2=((Y(J)-MEDIAN)/(C*S))**2;
IF U2<1 THEN DO;
  S1=S1+2*U2*((1-U2)**4);
  S2=S2+2*(1-U2)*(1–5*U2); END;
END;
S3=S2–1; IF S3<1 THEN S3=1;
IF S2>00001 THEN S=C*S*SQRT((N*S1)/(S2*S3));
SBI=S; S1=0; S2=0;
DO J=M1 TO N;
U2=((Y(J)-MEDIAN)/(C*SBI))**2;
IF U2<1 THEN DO;
  S1=S1+2*U2*((1-U2)**4);
  S2=S2+2*(1-U2)*(1–5*U2);
END;
END;
S3=S2–1; IF S3<1 THEN S3=1;
IF S2>00001 THEN S=C*S*SQRT(S1/(S2*S3));
ST2=S**2;
S=MAD;
S1=0; S2=0;
C=28.4; *** C=205.4 for 95% reference intervals;
DO J=M1 TO N;
U2=((Y(J)-MEDIAN)/(C*S))**2;
IF U2<1 THEN DO;
  S1=S1+2*U2((1-U2)**4);
  S2=S2+2*(1-U2)*(1–5*U2); END;
END;
S3=S2–1; IF S3<1 THEN S3=1;
IF S2>00001 THEN S=C*S*SQRT((N*S1)/(S2*S3));
SBI2=S**2;
rob95=MEDIAN+TINV(.95,(N-1))*SQRT(ST2+SBI2);
*** TINV(.975,(N-1)) for 95% reference intervals;
```

References

1. Harrel FE, Davis CE. A new distribution free quantile estimator. Biometrika 1982;69:635–70.
2. Efron B. The jackknife, the bootstrap, and other resampling plans. CBMS-NSF regional conference series in applied mathematics. Philadelphia: Society for Industrial and Applied Mathematics, 1982:29–36.
3. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker, 1995:1–61.
4. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc 1964;B26: 211–52.
5. Boyd JC, Lacher DA. A multi-stage gaussian transformation algorithm for clinical laboratory data. Clin Chem 1982;28:1735–41.
6. Linnet K. Two-stage transformation systems for normalization of reference distributions evaluated. Clin Chem 1987;33:381–6.
7. Horn PS. A biweight prediction interval for random samples. J Am Stat Assoc 1988;83:249–56.
8. Kafadar K. A biweight approach to the one-sample problem. J Am Stat Assoc 1982;77:416–24.
9. Horn PS. Robust quantile estimators for skewed populations. Biometrika 1990;77:631–6.
10. International Federation of Clinical Chemistry. Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. J Clin Chem Clin Biochem 1987;25:645–56.
11. American diabetes data group classifications and diagnosis of diabetes and other categories of glucose intolerance. Diabetes 1979;28:1039–57.