DOI: https://doi.org/10.4038/sjhs.v3i1.45

Received: 22 October 2022 *Accepted: 24 December 2022*

Reference Ranges and Control Limits that are Resistant to Baseline Outliers

Amaratunga, D.

damaratung@yahoo.com
Princeton Data Analytics LLC, USA.

Abstract

Reference ranges and control limits are used in many settings – for example, to assess a person's health or to monitor the stability of a manufacturing process. Such ranges are established based on a baseline sample of what is considered normal data, but it is not possible to always avoid a few outliers being present even in this sample. If, as is common, the range is calculated using statistics, such as the mean and standard deviation, which could be influenced by outliers, then the use of such a range could adversely affect the decisions made. This can be avoided by constructing the reference range using statistics that are resistant to outliers. In this paper, we studied possible approaches and found two methods that had superior performance overall: one based on MM-estimation and one based on a form of Winsorization.

Keywords: M-estimation, Outliers, Quality control, Reference range, Robust statistics.

Introduction

Reference ranges and control limits are used in various settings. In medicine, a reference range (often also called a normal range) is an interval of values that is deemed normal for a healthy person; a value outside the reference range limits could indicate a potential health issue. In manufacturing, control limits define the acceptable range of results for a process; a value outside the control limits could be indicating that the process being monitored is out of control. In agriculture, control limits have been used to monitor harvest yields. In environmental studies, control limits have been used to monitor air pollution levels.In general, a laboratory or a manufacturer will establish a reference range for a variable by testing a random sample of individuals from the "normal" population (we will refer to this sample as the "baseline") and then using these baseline values to derive an interval which is such that the probability, β, that a random observation from the "normal" population would be large (usually β = 95% or 99%). Such β-level reference ranges are usually constructed using the mean, $\bar{x}$, and standard deviation, s, of the baseline values. Some references for estimation procedures are Solberg (1984), Harris and Boyd (1995), Amaratunga (1997), Linnet (2000) and Wright and Royston (1999).In practice it is difficult to totally avoid a few atypical values or outliers being present in the baseline sample used to

construct a reference range. If that happens, then this can impair the development of the reference range if it was constructed using statistics such as the mean and standard deviation which can be unduly affected by outliers. In such situations, it is useful to consider reference ranges constructed using statistics that are resistant to sampling from contaminated distributions. Thus, for example, outlier-removal techniques and M-estimators have been recommended for deriving clinical reference ranges (Horn et al, 1998, Horn et al, 2001) and M-estimates of dispersion have been suggested for constructing manufacturing process control charts (Shahriari et al., 2009). The value of considering such approaches in data analysis in general was pointed out by Tukey (1960), in what is regarded as the seminal paper on robust statistics. In this paper, we will study some possible approaches for this particular problem – the construction of reference ranges.

Materials and Methods

The primary purpose of a reference range is to identify atypical values. Let $f(.)$ be the probability distribution function of the typical values and let x be a new observation. If an interval $\hat{I} = (a,b)$ is to be used as a reference range, it would clearly be desirable for $\hat{I}$ to be such that $\text{Prob}(x \epsilon \hat{I}) = \int_a^b f(u)du$ is large if x is a typical value and small if x is an atypical value. Hence, we use the "positive rate",

$$PR = 1 - E[\int_a^b f(u)du] \quad (1)$$

as a statistic to assess the performance of potential reference ranges. *PR* would be small for typical values and large for atypical values. We explore five methods (labeled RR1 to RR5) for computing reference ranges.

Method RR1

First, we will assume that the typical values are Normally distributed or have been transformed to Normality, i.e., they have a $N(\mu,\sigma^2)$ distribution, while atypical values are such that they have a $N(\mu+\Delta,\sigma^2)$ distribution where $\Delta \neq 0$. In this case, it is natural to seek a reference range of the form: $\hat{I}_k = (\bar{x} \pm ks)$. Large values of k will result in small values of *PR* if $\Delta = 0$ and large values of *PR* if $\Delta \neq 0$; in fact, the value of *PR* will increase monotonically with Δ. Making k too small will result in too many atypical values being judged typical while making k too large will result in the opposite. In order to balance this, we could choose k to be such that $PR = \beta$ for a reasonable value of β when $\Delta = 0$. This produces a β -expectation tolerance interval and, as it turns out, also a β -level prediction interval (Guttman, 1970). The corresponding value of k is

$$k = t_{((1-\beta/2),(n-1))}\sqrt{(1+1/n)} \quad (2)$$

where $t_{a,n}$ denotes the a-th quantile of a t distribution with n degrees of freedom (Wilks, 1941). We call this reference range RR1.

Method RR2

Since both $\bar{x}$ and s can be influenced by outliers, we can replace them in RR1 by the Median and MAD (median absolute deviation) respectively since they are highly resistant to outliers – they both have breakdown point 50% meaning that they can tolerate up to about 50% of the sample being outliers. This is reference range RR2. A drawback however is that the Median and MAD have low efficiency at the Normal if the baseline sample size is small.

Method RR3

Another possibility to avoid the reference range being overly influenced by baseline outliers is to use a robust method. Robust estimates strive to provide good efficiency at the nominal Normal distribution as well as resistance to outliers. Horn et al (1998) suggested using M-estimates and showed that doing this, besides providing resistance to outliers, also protects against imperfect transformations to Normality. Here we use MM-estimation, an extension of M-estimation, that has somewhat better performance (Yohai, 1987) – it produces estimators with breakdown point 50% and high efficiency at the Normal. This is reference range RR3.

Method RR4

A common exploratory approach for detection and management of outliers is to run an outlier test and to eliminate any outliers found. For example, we could calculate standardized residuals using the median, M, and MAD, i.e., calculate $u_i=(x_i-M)/MAD$, and designate as outliers any values whose $u_i>k$, where k is the value used in method RR1. However, instead of eliminating such values, which could lead to bias, we can instead set such values equal to the cutoff, k, a form of Winsorizing (Dixon, 1960) We can then use method RR1 on the "cleaned-up" data as it now will not contain any outliers. This is reference range RR4.

Method RR5

In the event that, normality or a transformation to normality is not tenable, nonparametric reference ranges of the form $\hat{I}_r = (x_{(r)}, x_{(n-r+1)})$, where $x_{(r)}$ denotes the r^{th} order statistic ($r<n/2$), can be considered. Setting r small (i.e., setting the limits of the reference range close to the sample extremes) will result in small values of PR if $\Delta=0$, but could also result in small values of PR if $\Delta\neq0$. In analogy with the discussion related to method RR1, it seems best to select r to be as large as possible but with PR set to its prespecified value of β when $\Delta=0$. This is obtained by setting $h=(n+1)(1-\beta)/2$ (Guttman, 1970) and, if h is an integer, letting $r=h$. If h is not an integer and $h=s+f$, where s is its integer part and f is its fractional part, the left endpoint of $\hat{I}$ is determined by linearly interpolating between the s^{th} and $(s+1)^{th}$ order statistics: $(1-f)x_{(s)} + fx_{(s+1)}$, with the right endpoint of $\hat{I}$ determined analogously (Hall and Rieck, 2011). We shall refer to this as reference range RR5. These five methods are summarized in Table 1.

Table 1.
List of methods for deriving a reference range.

Method	Formula	Notation details
RR1	$(\bar{x} \pm ks)$	$\bar{x}$ are the mean and standard deviation
RR2	$(M \pm kMAD)$	M and MAD are the median and median absolute deviation
RR3	$(\bar{x}_{MM} \pm ks_{MM})$	$\bar{x}_{MM}$ and s_{MM} are the MM-estimates of the mean and standard deviation
RR4	$(\bar{x}_{WIN} \pm ks_{WIN})$	$\bar{x}_{WIN}$ and s_{WIN} are the mean and standard deviation after replacing large values at the high and low ends with cutoff values
RR5	$(x_{(r)}, x_{(n-r+1)})$	$x_{(r)}$ is the rth order statistic, where r is defined in the Methods section

Note: The value of k is the same in methods RR1 to RR4 and is given in Equation (2).

Note that γ-content β-level tolerance intervals have been suggested in the literature for use

as reference ranges (see e.g., Liu et al., 2021). One rationale for this suggestion is that if a reference range based on a prediction interval is used repeatedly, it will, due to multiplicity, falsely flag too many typical values as atypical - a reference range based on a tolerance interval is less likely to do this. However, intervals based on tolerance intervals will be too wide (unless γ and/or β are set low) and many truly atypical values would miss being detected (Wellek & Jennen-Steinmetz, 2022). In practice, the balance desired between the rates of correct and incorrect decisions should be considered carefully before a method is implemented. Since β-level prediction intervals are directly consistent with the objective of constructing an interval I such that a fresh random observation from the baseline distribution falls into I with a specific probability, only such intervals were considered in this paper.

Results and Discussion

We ran a series of simulations to compare the above 5 reference range construction methods. The simulations were carried out as follows. A sample $\{z_i\}$ of size n was drawn from a $N(0,1)$ distribution and a sample $\{d_i\}$ of size n was drawn from a Bernoulli(p_0) distribution. Then we set $x_i = z_i+\Delta_0 d_i$. Here p_0 and Δ_0 are respectively the probability and mean shift of an atypical value in the baseline sample. Reference ranges were calculated using each of the methods, RR1 to RR5. The probability, $p(\Delta; n, p_0, \Delta_0)$, of an observation from a $N(\mu+\Delta,s^2)$ distribution being declared atypical was calculated for several different values of Δ using methods RR1 to RR5 with β=95%. This was repeated 500 times and the mean, $PR(\Delta; n, p_0, \Delta_0)$, and standard deviation, $SDPR(\Delta; n, p_0, \Delta_0)$, of the $p(\Delta; n, p_0, \Delta_0)$ values were recorded.

The above simulation was carried out for several different values of n, p_0, Δ_0 and Δ. Table 2a shows the results for situations in which the baseline sample does not contain any atypical values, so that $p_0=\Delta_0=0$.

Table 2.

The simulation results for situations in which the baseline sample (a) does not contain any atypical values, so that $p_0=\Delta_0=0$ and (b). contains on average 10% atypical values (i.e., $p_0=0.1$) with mean 3 (i.e., $\Delta_0=3$).

(a).

	Mean p_0=0 Δ_0=0 Δ=0	SD p_0=0 Δ_0=0 Δ=0	Mean p_0=0 Δ_0=0 Δ=1	Mean p_0=0 Δ_0=0 Δ=3	Mean p_0=0 Δ_0=0 Δ=5
N=10					
RR1	0.048	*0.062*	0.131	0.382	0.703
RR2	0.096	*0.128*	0.194	0.438	0.711
RR3	0.065	*0.096*	0.150	0.385	0.674
RR4	0.064	*0.086*	0.161	0.420	0.725
RR5	0.173	*0.113*	0.325	0.648	0.893
N=30					
RR1	0.048	*0.030*	0.154	0.468	0.807
RR2	0.069	*0.059*	0.179	0.490	0.808
RR3	0.056	*0.045*	0.163	0.474	0.804
RR4	0.060	*0.041*	0.179	0.504	0.827
RR5	0.066	*0.044*	0.179	0.490	0.810
N=50					
RR1	0.050	*0.024*	0.163	0.492	0.829
RR2	0.063	*0.047*	0.181	0.505	0.827
RR3	0.055	*0.035*	0.170	0.496	0.827
RR4	0.062	*0.033*	0.187	0.526	0.848
RR5	0.058	*0.031*	0.172	0.496	0.824

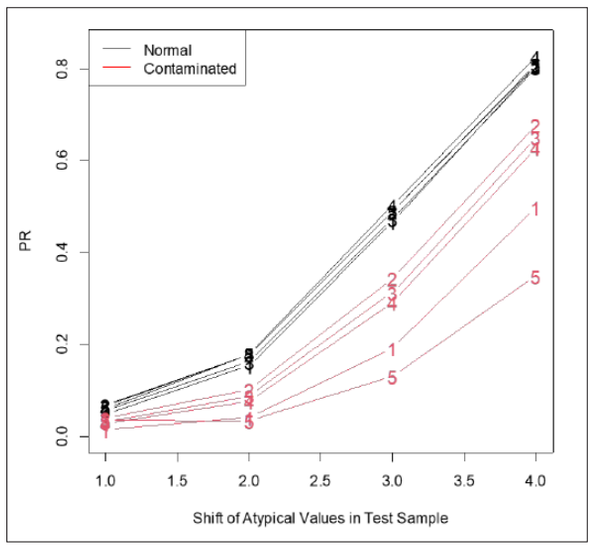
(b).

	Mean p_0=0.1 Δ_0=3 Δ=0	SD p_0=0.1 Δ_0=3 Δ=0	Mean p_0=0.1 Δ_0=3 Δ=1	Mean p_0=0.1 Δ_0=3 Δ=3	Mean p_0=0.1 Δ_0=3 Δ=5
N=10					
RR1	0.027	*0.049*	0.067	0.210	0.447
RR2	0.073	*0.102*	0.144	0.360	0.631
RR3	0.046	*0.077*	0.104	0.289	0.553
RR4	0.040	*0.062*	0.096	0.277	0.542
RR5	0.143	*0.110*	0.167	0.358	0.604
N=30					
RR1	0.016	*0.017*	0.042	0.193	0.498
RR2	0.040	*0.046*	0.103	0.345	0.678
RR3	0.032	*0.034*	0.088	0.314	0.652
RR4	0.029	*0.027*	0.077	0.292	0.629
RR5	0.036	*0.032*	0.034	0.131	0.349
N=50					
RR1	0.014	*0.013*	0.035	0.182	0.500
RR2	0.036	*0.031*	0.100	0.351	0.698
RR3	0.031	*0.023*	0.087	0.328	0.680
RR4	0.029	*0.021*	0.077	0.303	0.654
RR5	0.034	*0.025*	0.016	0.081	0.283

Simulation results are reported for n=10, 30, 50; and Δ=0, 1, 3, 5. Note that the PR when Δ=0 is an estimate of the false positive rate (FPR) of the method. Table 2b shows the results for situations in which the baseline sample contains on average 10% atypical values (i.e., p_0=0.1) with mean 3 (i.e., Δ_0=3). Here, too, simulation results are reported for n=10, 30, 50; and Δ=0, 1, 3, 5. Figure 1 is a scatterplot of PR vs Δ when n=30 for both the normal and contaminated situations.

Figure 1.

Plot of the simulation results for N=30. The black lines show the PR values for situations in which the baseline sample does not contain any atypical values, so that p_0=Δ_0=0, while the red lines show the PR values for situations in which the baseline sample contains on average 10% atypical values (i.e., p_0=0.1) with mean 3 (i.e., Δ_0=3). The plotting character refers to the reference range construction method (RR1 to RR5).



When there were no atypical values in the baseline data, i.e., when (p_0, Δ_0)=(0,0), the conventional method RR1 and the robust methods RR3 and RR4 all performed equally well - they had FPR of 5% as desired and had high PR when atypical values were present in the test data. On the other hand, method RR2 had a somewhat higher FPR when the sample size was small and it had a higher variability than all the other methods due to the lower efficiency of the Median and MAD. Method RR5 also had high FPR and high variability that was too high at small sample sizes.

When there were atypical values in the baseline data, the performance of the conventional method RR1 deteriorated substantially, both in terms of FPR and its ability to detect atypical values when they were present in the test data. RR5 also had overall weak performance in this case. The methods based on resistant statistics, RR2, RR3 and RR4, all performed well except that RR2 had somewhat high FPR when the sample size was small and high variability.The findings from the simulation are consistent with the properties of the

statistics used to derive them. The mean and standard deviation are not at all resistant to outliers having zero breakdown point. The median and MAD are highly resistant to outliers having 50% breakdown point, but have high variability and are inefficient at the normal especially when the baseline sample size is small. MM-estimators and Winsorized estimators are robust; they have high efficiency at the normal and also when there are outliers in the data or when there is some kurtosis in the data (in the event of skewness, the data could be transformed to symmetry). Therefore, constructing reference ranges using these latter estimators is likely to lead to better performance overall.

Conclusions

We considered five methods for constructing reference ranges and compared their performance both in the absence as well as in the presence of atypical values in the sample on which the ranges were based. While the conventional method for constructing reference ranges based on mean and standard deviation worked well when there were no atypical values in the baseline sample, its performance deteriorated when there were some. On the other hand, methods based on robust statistics (methods RR3 and RR4) performed well in both situations. Therefore, we recommend using a robust approach, in particular either RR3 (based on MM-estimates) or RR4 (based on a form of Winsorized statistics), when constructing reference ranges.

References

Amaratunga, D. (1997). Reference ranges for screening preclinical drug safety data. *Journal of Biopharmaceutical Statistics*, 7, 417- 422.

Dixon, W. J. (1960). Simplified estimation from censored normal samples. *Annals of Mathematical Statistics*, 31, 385-391.

Guttman, I. (1970). Construction of β-content tolerance regions at confidence level γ for large samples from the k-variate normal distribution. *Annals of Mathematical Statistics*, 41, 376-400.

Hall, P., & Rieck, A. (2001). Improving coverage accuracy of nonparametric prediction intervals. *Journal of the Royal Statistical Society*, *Series B,* 63, 717-726.

Harris, E. K., & Boyd, J. C. (1995). Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker.

Horn, P. S., Feng, L., Li, Y., & Pesce, A. J. (2001). Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47, 2137-2145.

Horn, P. S., Pesce, A. J., & Copeland, B. E. (1998). A robust approach to reference interval estimation and evaluation. *Clinical Chemistry*, 44, 622-631.

Linnet, K. (2000). Nonparametric estimation of reference intervals by simple and

bootstrap-based procedures. *Clinical Chemistry*, 46, 867-869.

Liu, W., Bretz, F., & Cortina-Borja, M. (2021). Reference range: which statistical intervals to use? *Statistical Methods in Medical Research*, 30, 523-534.

Shahriari, H., Maddahi, A., & Shokouhi, A. (2009). A robust dispersion control chart based on M-estimate, *Journal of Industrial and Systems Engineering*, 2:297-307.

Solberg, H. (1984). International Federation of Clinical Chemistry, Scientific Committee, Clinical Section. Expert Panel on Theory of Reference Values (EPTRV). The theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *Clinica chimica acta; International Journal of Clinical Chemistry*, 137, 97-114.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling (I. Olkin et al., eds.), 448-485. Stanford University Press.

Wellek, S., & Jennen-Steinmetz, C. (2022). Reference ranges: Why tolerance intervals should not be used. Comment on Liu, Bretz and Cortina-Borja, Reference range: Which statistical intervals to use? *Statistical Methods in Medical Research*. 31, 2255-2256.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12, 91-96.

Wright, E. M., & Royston, P. (1999). Calculating reference intervals for laboratory measurements. *Statistical Methods in Medical Research*, 8, 93-112.

Yohai, V. J. (1987). High breakdowns point and high efficiency robust estimates for regression. *Annals of Statistics.* 15, 642-656.