stats

MDPI

Article

Nonparametric Limits of Agreement for Small to Moderate Sample Sizes: A Simulation Study

Maria E. Frey [1], Hans C. Petersen [2] and Oke Gerke [3,4,*]

[1] Department of Toxicology, Charles River Laboratories Copenhagen A/S, 4623 Lille Skensved, Denmark; mariafrey93@hotmail.com
[2] Department of Mathematics and Computer Science, University of Southern Denmark, 5230 Odense M, Denmark; hcpetersen@sdu.dk
[3] Department of Nuclear Medicine, Odense University Hospital, 5000 Odense C, Denmark
[4] Department of Clinical Research, University of Southern Denmark, 5000 Odense C, Denmark
* Correspondence: oke.gerke@rsyd.dk

Received: 7 August 2020; Accepted: 25 August 2020; Published: 28 August 2020

Abstract: The assessment of agreement in method comparison and observer variability analysis of quantitative measurements is usually done by the Bland–Altman Limits of Agreement, where the paired differences are implicitly assumed to follow a normal distribution. Whenever this assumption does not hold, the 2.5% and 97.5% percentiles are obtained by quantile estimation. In the literature, empirical quantiles have been used for this purpose. In this simulation study, we applied both sample, subsampling, and kernel quantile estimators, as well as other methods for quantile estimation to sample sizes between 30 and 150 and different distributions of the paired differences. The performance of 15 estimators in generating prediction intervals was measured by their respective coverage probability for one newly generated observation. Our results indicated that sample quantile estimators based on one or two order statistics outperformed all of the other estimators and they can be used for deriving nonparametric Limits of Agreement. For sample sizes exceeding 80 observations, more advanced quantile estimators, such as the Harrell–Davis and estimators of Sfakianakis–Verginis type, which use all of the observed differences, performed likewise well, but may be considered intuitively more appealing than simple sample quantile estimators that are based on only two observations per quantile.

Keywords: agreement; Bland-Altman plot; coverage; limits of agreement; method comparison; quantile estimation; repeatability; reproducibility

1. Introduction

The classical Bland–Altman Limits of Agreement (BA LoA) define a range within which approximately 95% of normally distributed differences between paired measurements are expected to lie [1–3]. In cases of non-normally distributed differences, the use of empirical quantiles has been proposed as a robust alternative [2,4,5]; however, extensive research endeavors in the past have suggested the application of nonparametric quantile estimation to the assessment of 2.5% and 97.5% percentiles as nonparametric LoA. We performed a simulation study on 15 nonparametric quantile estimators to derive nonparametric prediction intervals and assessed their performance by means of the coverage probability for one newly generated observation. The aim of this study was to suggest a nonparametric and robust alternative to the classical BA LoA when the normality assumption does not hold and/or the sample sizes are small to moderate. Our findings are illustrated by an application to data from a previously published clinical study on coronary artery calcification measured by the Agatston score [6].

2. Methods

Let the differences of paired observations be independent observations of a random variable X with a cumulative distribution function (CDF) $F : \mathbf{R} \to [\mathbf{0}, \mathbf{1}]$. If F is continuous from the right, then the *quantile function* of X is continuous from the left: $Q(u) := \inf\{x : F(x) \geq u\}, u \in (0,1)$; hence, at least $100u$ percent of the values of X are below $Q(u)$ [7,8]. In the following, we seek to estimate the 2.5% and 97.5% percentiles of F, corresponding to $u = 0.025, 0.975$, by different methods of nonparametric quantile estimation.

Databases, such as JSTOR (Journal Storage), ScienceDirect, the online journal platform "Taylor & Francis Online", and those maintained by PubMed/Medline in the NCBI (National Center for Biotechnology Information) were searched for quantile estimator, nonparametric quantile estimator, nonparametric kernel quantile estimator, subsampling quantile estimator, and new quantile estimator. Fifteen nonparametric quantile estimators were chosen, three of which are sample quantile estimators, four are subsampling quantile estimators, two are kernel quantile estimators, and six are other quantile estimators. Hence, we chose very different types of nonparametric quantile estimators, which use one, two, or all the observations in a sample.

2.1. Sample Quantile Estimators

The simplest way to estimate quantiles nonparametrically is by using sample quantile estimators. A random sample $X_1, \ldots X_n$ of size n is sorted in increasing order $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$; the symbols here denote the *order statistics* of the random sample. The CDF of F can then be estimated by the step function

$$\tilde{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[x_i, \infty)}(x), \tag{1}$$

where $I_A(x)$ is the indicator function that takes the value 1 if $x \in A$ and 0 if $x \notin A$; the sample quantile function is defined as $\tilde{Q}(u) := \inf\{x : \tilde{F}(x) \geq u\}, u \in (0,1)$ by Cheng [7]. The quantile estimator

$$SQ_{p1} = \begin{cases} X_{(np)} & \text{if } [np] = np \\ X_{([np]+1)} & \text{if } [np] < np \end{cases} \tag{2}$$

is based on a single order statistic, where $[x]$ is the greatest integer that is less than or equal to x [9,10]. SQ_{p1} is the smallest observation for which at least p percent of the observed values in the sample are smaller than or equal to SQ_{p1}.

The second sample quantile estimator SQ_{p2} is a weighted average of the two order statistics that are closest to including p percent of all the observations in the sample:

$$SQ_{p2} = (1 - \alpha)X_{(r)} + \alpha X_{(r+1)} \tag{3}$$

with $\alpha = p(n+1) - r$ and $r = [p(n+1)]$ [9,11].

Finally, we considered a weighted average of $X_{[np+0.5]}$ and $X_{[np+0.5]+1}$:

$$SQI_p = (-np + 0.5 + i)X_{(i)} + (np + 0.5 - i)X_{(i+1)} \tag{4}$$

with $i = [np + 0.5]$ and $0.5 \leq np \leq (n - 0.5)$ [9,12].

2.2. Subsampling Quantile Estimators

The abovementioned sample quantile estimators are based on only one or two order statistics whereas subsampling, kernel, and other quantile estimators employ linear combinations of all the available order statistics, weighting them according to their relative closeness to the target percentile.

Based on the sample quantile function $\tilde{Q}$, linear smooth nonparametric estimators of $Q(u)$ can be written as

$$Q(u) = \int_0^1 \tilde{Q}(t) d_t G(u;t), \tag{5}$$

where $G(u;\cdot)$ is a CDF with support on the unit interval. Many distributions $G(u;\cdot)$ have been proposed, the choice of which depends on the sample size and, typically, a smoothing parameter. Depending on the choice of $G(u;\cdot)$, there are two major classes of quantile estimators according to Cheng [7]: subsampling quantile estimators and kernel quantile estimators. Both can be given as L-statistics, which is, $\sum_{j=1}^{n} W_j \cdot X_{(j)}$, where W_j and $X_{(j)}$ is the weight for the j-th order statistic and the j-th order statistic itself, respectively [13]. For subsampling quantile estimators, a discrete distribution is chosen for $G(u;\cdot)$, interpreted as a resampling distribution from the set of the observed order statistics $X_{(j)}, j = 1, \ldots, n$ [7].

The Harrell-Davis estimator is given by

$$HD_p = \sum_{i=1}^{n} W_i X_{(i)} \tag{6}$$

with weight function

$$\begin{aligned} W_i &= \frac{1}{\beta\left\{(n+1)p, (n+1)(1-p)\right\}} \int_{(i-1)/n}^{i/n} y^{(n+1)p-1}(1-y)^{(n+1)(1-p)-1} dy \\ &= I_{i/n}\left\{p(n+1), (1-p)(n+1)\right\} - I_{(i-1)/n}\left\{p(n+1), (1-p)(n+1)\right\}, \end{aligned}$$

where $I_{i/n}\{a, b\}$ is the incomplete beta function [10,14,15].

The quantile estimator of Kaigh and Lachenbruch

$$KL_p = \left[\sum_{j=r}^{r+n-k} \binom{j-1}{r-1}\binom{n-j}{k-r} \Big/ \binom{n}{k}\right] X_{(j)}, \tag{7}$$

where $r = [(k+1)p]$, is obtained by averaging a subsample quantile estimate over all $\binom{k}{n}$ subsamples of size $k, 1 \leq k \leq n$, which are sampled without replacement. The subsample size k is an arbitrary smoothing (or reduction) parameter, and Kaigh and Lachenbruch proposed choosing k, so as to minimize the mean squared error (MSE), which is, $MSE = E(KL_p - \varepsilon_p)^2$, where ε_p is the true value of the p-th quantile [7,16,17].

Kaigh and Cheng [18] proposed the quantile estimator

$$KC_p = \sum_{j=1}^{n} \left[\binom{r+j-2}{r-1}\binom{n-j+k-r}{k-r} \Big/ \binom{n+k-1}{k}\right] X_{(j)} \tag{8}$$

with $r = \lceil kp \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x [7]. The value of k is again determined by minimizing the MSE of the estimator.

Finally, the Bernstein polynomial quantile estimator is given by

$$BP_p = \sum_{j=1}^{n} \binom{n-1}{j-1} p^{j-1}(1-p)^{n-j} X_{(j)} \tag{9}$$

according to Cheng [7,19].

2.3. Kernel Quantile Estimators

Like subsampling quantile estimators, kernel quantile estimators can be written in the form of Equation (5). Here, $G(u;.)$ is a location-scale family CDF with density function K, location parameter u, and scale parameter $h(n)$:

$$G(u;t) = \frac{1}{h(n)} K\left(\frac{t-u}{h(n)}\right).$$

Subsequently, Equation (5) becomes the kernel quantile estimator introduced by Parzen [8]:

$$\hat{Q}(u) = \int_0^1 \tilde{Q}(t) \cdot \frac{1}{h(n)} \cdot K\left(\frac{t-u}{h(n)}\right) dt$$

with its L-statistic representation being

$$\hat{Q}(u) = \sum_{j=1}^{n} \left[\int_{(j-1)/n}^{j/n} \frac{1}{h(n)} \cdot K\left(\frac{t-u}{h(n)}\right) dt \right] X_{(j)}.$$

The density K, called the kernel, is symmetric around zero and it has a chosen *bandwidth* $h(n)$, which satisfies $h(n) \to \infty$ when $n \to \infty$ [7,20]. Yang [21] proposed a discretized version, which we used as the first of the two kernel quantile estimators in our study due to its closed form:

$$KQ_{p1} = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h(n)} \cdot K\left(\frac{(j/n)-p}{h(n)}\right) X_{(j)}. \tag{10}$$

As $K\left(\frac{1/n-p}{h(n)}\right), \ldots, K\left(\frac{n/n-p}{h(n)}\right)$ do not generally provide a (discrete) probability distribution on $[0,1]$, monotonicity, translation and scale equivariance, and the symmetry relation do not hold. Translation and scale equivariance would imply that KQ_{p1} applied to $(X_1+c, \ldots, X_n+c)$ is equal to c plus KQ_{p1} applied to $(X_1, \ldots, X_n)$; the asymmetry means that KQ_{p1} applied to $(-X_1, \ldots, -X_n)$ is not equal to $-KQ_{p1}$ applied to $(X_1, \ldots, X_n)$. The Nadaraya–Watson type estimator

$$KQ_{p2} = \sum_{j=1}^{n} \frac{K\left(\frac{((j-0.5)/n)-p}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{((i-0.5)/n)-p}{h(n)}\right)} X_{(j)} \tag{11}$$

overcomes these drawbacks and was used as the second of the two kernel quantile estimators in this study [7,22,23]. Its name originates from the Nadaraya–Watson estimator in kernel regression [24,25].

In the following, we chose the standard Gaussian kernel for K and used the value of the bandwith $h(n)$ that minimized the MSE to find k for both KL_p in Equation (7) and KC_p in Equation (8).

2.4. Other Quantile Estimators

The kernel quantile estimators that are presented in the previous section are all based on the usual empirical distribution (1) with equal weights $1/n$ assigned to each observation. To improve the performance of quantile estimators, Huang and Brill [26] proposed using a weighted empirical distribution, for instance, the level crossing empirical distribution function

$$F_{lc}(x) = \sum_{i=1}^{n} w_{i,n} I_{(-\infty,x]}(X_{(i)}) \tag{12}$$

with weight function

$$w_{i,n} = \begin{cases} \frac{1}{2}\left[1 - \frac{n-2}{\sqrt{n(n-1)}}\right] & \text{if } i = 1, n \\ \frac{1}{\sqrt{n(n-1)}} & \text{if } i = 2, 3..., n-1. \end{cases} \tag{13}$$

One such kernel quantile estimator using level crossing empirical distributions is

$$KQ_{plc} = \sum_{i=1}^{n} \left[n^{-1} \frac{1}{h(n)} K \left(\frac{\sum_{j=1}^{i} w_{j,n} - p}{h(n)} \right) \right] X_{(i)}. \tag{14}$$

Huang [27] modified the Harrell–Davis estimator (6) by applying a weighted empirical distribution function instead of the empirical distribution with equal weights $1/n$. The Harrell–Davis estimator using a level crossing empirical distribution function can be written as

$$\begin{aligned} HD_{plc} &= \frac{1}{\beta((n+1)p, (n+1)q)} \int_0^1 F_{lc}^{-1}(y) y^{(n+1)p-1} (1-y)^{(n+1)q-1} dy \\ &= \sum_{i=1}^{n} \left[\int_{q_{i-1,n}}^{q_{i,n}} \frac{1}{\beta((n+1)p, (n+1)q)} y^{(n+1)p-1} (1-y)^{(n+1)q-1} dy \right] X_{(i)}, \end{aligned} \tag{15}$$

where $\beta(\cdot, \cdot)$ is the beta function, $q = 1 - p$, $F_{lc}(\cdot)$ is given by (12), $q_{i,n} = \sum_{j=1}^{i} w_{j,n}$, $i = 1, \ldots, n$, with $w_{j,n}$, as defined in (13), and $q_{0,n} \equiv 0$.

Sfakianakis and Verginis [28] proposed a group of estimators, motivated by the fact that nonparametric quantile estimation of extreme quantiles close to 0 and 1 requires large samples for sufficient accuracy. These three quantile estimators are supposed to better estimate quantiles in the tails of a distribution when using small samples and they employ the Binomial probability of observing exactly i out of n events with an event probability of p, $B(i; n, p)$:

$$\begin{aligned} SV_{p1} = {} & \frac{2B(0; n, p) + B(1; n, p)}{2} X_{(1)} + \frac{B(0; n, p)}{2} X_{(2)} - \frac{B(0; n, p)}{2} X_{(3)} \\ & + \sum_{i=2}^{n-1} \frac{B(i; n, p) + B(i-1; n, p)}{2} X_{(i)} \\ & - \frac{B(n; n, p)}{2} X_{(n-2)} + \frac{B(n; n, p)}{2} X_{(n-1)} + \frac{2B(n; n, p) + B(n-1; n, p)}{2} X_{(n)}, \end{aligned} \tag{16}$$

$$SV_{p2} = \sum_{i=0}^{n-1} B(i; n, p) X_{(i+1)} + (2X_{(n)} - X_{(n-1)}) B(n; n, p), \tag{17}$$

$$SV_{p3} = \sum_{i=1}^{n} B(i; n, p) X_{(i)} + (2X_{(1)} - X_{(2)}) B(0; n, p). \tag{18}$$

Finally, Navruz and Özdemir [29] introduced a new quantile estimator, which is a weighted average of all order statistics:

$$NO_p = (B(0;n,p)2p + B(1;n,p)p)X_{(1)} + B(0;n,p)(2-3p)X_{(2)} - B(0;n,p)(1-p)X_{(3)} + \sum_{i=1}^{n-2}(B(i;n,p)(1-p) + B(i+1;n,p)p)X_{(i+1)} - B(n;n,p)pX_{(n-2)} + B(n;n,p)(3p-1)X_{(n-1)} + (B(n-1;n,p)(1-p) + B(n;n,p)(2-2p))X_{(n)}. \quad (19)$$

2.5. Simulation Setup

We contrasted nonparametric LoA as constructed with the 15 abovementioned quantile estimators by comparing their coverage probabilities for the next paired difference under the given distributional assumption. Here, we employed the standard normal distribution (ND), a standard normal distribution with 1%, 2%, and 5% outliers (ND 1%, ND 2%, and ND 5%, respectively), an exponential distribution (ED) with a rate of 1, and a lognormal distribution (LND) with meanlog = 0 and sdlog = 1. For normal distributions comprising outliers, simulated data were replaced with a probability of 1%, 2%, and 5% by data sampled from a normal distribution with a mean of 0 and a standard deviation of 3. To examine small to moderate sample sizes, the sample size was set to 30, 50, 80, 100, and 150. For each combination of distribution, sample size, and nonparametric quantile estimator, 20,000 simulated trials of size $(n+1)$ were generated with R (the code is available as Supplemental Material S1). Here, a seed was set in order to use the same simulated data for each combination of distribution and sample size across nonparametric quantile estimators. The first n observations in each simulated trial were used to derive nonparametric LoA, to which the last observation was compared. The coverage probability was then the proportion of cases out of the 20,000 trials where the nonparametric LoA included the last observation. All of the figures were generated with Stata/MP 16.1 (College Station, TX 77845, USA).

3. Results

For $n = 30$ (Table 1), none of the estimators reached the nominal coverage probability of 0.95. The coverage probability of SQ_{p1} was closest to 0.95, ranging from 0.934 to 0.938. Note that, for sample sizes of up to $n = 40$ observations, the smallest and largest difference are used as nonparametric quantile estimates for the 2.5% and 97.5% percentiles, respectively. For SQI_p, HD_p, HD_{plc}, SV_{p1}, SV_{p2}, and SV_{p3}, the coverage probabilities were at least 0.921, 0.911, 0.910, 0.920, 0.914, and 0.923, respectively. Neither SQ_{p2} nor KL_p are defined for $n < 40$.

Table 1. Coverage probabilities for nonparametric Limits of Agreement ($n = 30$). Neither SQ_{p2} nor KL_p are defined for $n < 40$.

Estimator	ND	ND 1%	ND 2%	ND 5%	ED	LND
SQ_{p1}	0.937	0.938	0.937	0.937	0.934	0.937
SQ_{p2}	-	-	-	-	-	-
SQI_p	0.926	0.927	0.927	0.928	0.921	0.926
HD_p	0.911	0.923	0.923	0.924	0.916	0.920
KL_p	-	-	-	-	-	-
KC_p	0.900	0.890	0.877	0.851	0.885	0.880
BP_p	0.905	0.908	0.908	0.909	0.897	0.904
KQ_{p1}	0.904	0.882	0.865	0.832	0.936	0.922
KQ_{p2}	0.912	0.893	0.880	0.857	0.916	0.906
KQ_{plc}	0.915	0.893	0.874	0.838	0.923	0.919
HD_{plc}	0.916	0.917	0.917	0.919	0.910	0.915
SV_{p1}	0.924	0.925	0.925	0.926	0.920	0.924
SV_{p2}	0.925	0.927	0.926	0.927	0.914	0.919
SV_{p3}	0.925	0.926	0.925	0.926	0.923	0.929
NO_p	0.813	0.814	0.817	0.821	0.808	0.834

SQ_{p2} was the only estimator with coverage probabilities oscillating closely around 0.95 for all the investigated sample sizes $n \geq 50$ (Tables 2–5); for $n = 50$, SQ_{p2} was the only one to do so.

For $n = 80$ (Table 3), the coverage probabilities of HD_p, SV_{p1}, SV_{p2}, and SV_{p3} fluctuated closely around the nominal level except for the simulations with an ED (0.94). For $n \geq 100$ (Tables 4 and 5), these estimators performed close to the 0.95 nominal level for all of the investigated distributions.

The coverage probabilities of the recently proposed NO_p estimator varied between 0.945 and 0.950 for $n = 200$ and they were very close to 0.95 for $n = 250$ (results not shown here).

Table 2. Coverage probabilities for nonparametric Limits of Agreement ($n = 50$). Bold figures indicate coverage probabilities exceeding the nominal level of 0.95.

Estimator	ND	ND 1%	ND 2%	ND 5%	ED	LND
SQ_{p1}	0.919	0.918	0.920	0.919	0.919	0.919
SQ_{p2}	**0.951**	**0.952**	**0.953**	**0.955**	**0.951**	**0.951**
SQI_p	0.934	0.935	0.937	0.938	0.931	0.934
HD_p	0.939	0.940	0.942	0.945	0.935	0.938
KL_p	0.938	0.932	0.930	0.917	0.937	0.929
KC_p	0.924	0.906	0.898	0.879	0.915	0.912
BP_p	0.922	0.925	0.927	0.931	0.919	0.922
KQ_{p1}	0.919	0.902	0.896	0.861	0.931	0.933
KQ_{p2}	0.924	0.909	0.900	0.881	0.924	0.916
KQ_{plc}	0.924	0.912	0.901	0.867	0.944	0.930
HD_{plc}	0.933	0.935	0.936	0.940	0.928	0.931
SV_{p1}	0.940	0.941	0.942	0.945	0.935	0.939
SV_{p2}	0.937	0.939	0.939	0.943	0.933	0.934
SV_{p3}	0.938	0.941	0.942	0.944	0.938	0.941
NO_p	0.839	0.843	0.845	0.855	0.845	0.870

Table 3. Coverage probabilities for nonparametric Limits of Agreement ($n = 80$). Bold figures indicate coverage probabilities exceeding the nominal level of 0.95.

Estimator	ND	ND 1%	ND 2%	ND 5%	ED	LND
SQ_{p1}	0.939	0.939	0.940	0.938	0.934	0.939
SQ_{p2}	**0.950**	**0.951**	**0.951**	0.949	0.945	**0.950**
SQI_p	0.941	0.941	0.942	0.941	0.935	0.941
HD_p	**0.950**	**0.952**	**0.954**	**0.955**	0.940	0.949
KL_p	0.943	0.939	0.935	0.933	0.938	0.939
KC_p	0.936	0.923	0.916	0.895	0.925	0.929
BP_p	0.937	0.939	0.941	0.942	0.929	0.937
KQ_{p1}	0.934	0.926	0.921	0.887	0.934	0.940
KQ_{p2}	0.936	0.925	0.917	0.897	0.930	0.931
KQ_{plc}	0.933	0.928	0.922	0.890	0.942	0.939
HD_{plc}	0.943	0.945	0.947	0.948	0.934	0.943
SV_{p1}	**0.951**	**0.952**	**0.954**	**0.955**	0.940	0.949
SV_{p2}	**0.950**	**0.951**	**0.953**	**0.953**	0.940	0.948
SV_{p3}	0.949	**0.951**	**0.952**	**0.954**	0.940	**0.950**
NO_p	0.888	0.893	0.896	0.901	0.891	0.910

Table 4. Coverage probabilities for nonparametric Limits of Agreement ($n = 100$). Bold figures indicate coverage probabilities exceeding the nominal level of 0.95.

Estimator	ND	ND 1%	ND 2%	ND 5%	ED	LND
SQ_{p1}	0.941	0.941	0.941	0.942	0.941	0.941
SQ_{p2}	**0.952**	**0.952**	**0.952**	**0.954**	**0.953**	**0.952**
SQI_p	0.941	0.941	0.941	0.942	0.941	0.941
HD_p	**0.950**	**0.951**	**0.953**	**0.957**	0.948	**0.950**
KL_p	0.947	0.941	0.939	0.938	0.946	0.941
KC_p	0.940	0.929	0.920	0.902	0.935	0.935
BP_p	0.940	0.942	0.944	0.948	0.937	0.940
KQ_{p1}	0.937	0.935	0.921	0.896	0.942	0.943
KQ_{p2}	0.939	0.930	0.920	0.903	0.940	0.935
KQ_{plc}	0.938	0.935	0.932	0.905	0.940	0.941
HD_{plc}	0.944	0.946	0.948	**0.951**	0.942	0.944
SV_{p1}	**0.950**	**0.952**	**0.954**	**0.959**	0.948	**0.950**
SV_{p2}	**0.950**	**0.952**	**0.953**	**0.957**	0.948	0.949
SV_{p3}	**0.950**	**0.951**	**0.953**	**0.959**	0.947	**0.950**
NO_p	0.911	0.914	0.917	0.924	0.913	0.924

Table 5. Coverage probabilities for nonparametric Limits of Agreement ($n = 150$). Bold figures indicate coverage probabilities exceeding the nominal level of 0.95.

Estimator	ND	ND 1%	ND 2%	ND 5%	ED	LND
SQ_{p1}	0.945	0.945	0.944	0.946	0.946	0.945
SQ_{p2}	0.949	0.949	0.948	**0.951**	**0.950**	0.949
SQI_p	0.942	0.942	0.941	0.943	0.944	0.942
HD_p	0.948	0.949	**0.950**	**0.954**	0.948	0.948
KL_p	0.945	0.941	0.939	0.936	0.947	0.945
KC_p	0.941	0.934	0.930	0.906	0.939	0.940
BP_p	0.942	0.944	0.943	0.947	0.941	0.942
KQ_{p1}	0.940	0.936	0.928	0.903	0.943	0.943
KQ_{p2}	0.940	0.932	0.923	0.908	0.941	0.937
KQ_{plc}	0.939	0.935	0.928	0.904	0.941	0.943
HD_{plc}	0.944	0.946	0.945	0.949	0.944	0.944
SV_{p1}	0.949	**0.951**	**0.952**	**0.957**	0.949	0.949
SV_{p2}	0.949	**0.952**	**0.952**	**0.956**	0.948	**0.950**
SV_{p3}	0.949	**0.950**	**0.951**	**0.954**	0.947	0.949
NO_p	0.934	0.937	0.937	0.943	0.936	0.940

4. Example

Diederichsen et al. [6] compared coronary artery calcification measurements using the Agatston score with the measurements using Framingham Heart Score in Danes of 50 and 60 years of age. Of 1825 randomly sampled citizens, 1257 consented to participation in the study, and 1156 of them were eligible. Agatston scores were independently reanalyzed for 129 randomly chosen study participants, and the agreement measures were the proportions of agreement and the kappa statistics for dichotomized calcification status (absence vs. presence) to assess intra- and inter-rater agreement. In the following, the intra-rater differences are used for exemplification purposes.

Approximately half of the 129 participants had an Agatston score of 0. The paired intra-rater differences ranged from -683 to 130, with a first, second, and third quartile being equal to 0; the 5th, 10th, 90th, and 95th percentiles were -23, -12, 1.1, and 5, respectively. The empirical distribution of the paired differences was, therefore, characterized by its denseness around 0 and a single, comparatively extreme outlier, clearly indicating the inappropriateness of the normality assumption in this setting (see also a histogram including an approximating normal distribution as Supplemental Material S2).

Using SQ_{p2}, HD_p, and SV_{p1}, the nonparametric, asymmetric, and robust LoA are -61.5, 12.8; -96.2, 26.7; and, -122.1, 30.6, respectively, whereas the symmetric BA LoA of -129.8, 116.1 are equidistant from the estimated mean difference of -6.9 (Figure 1). The upper LoA for HD_p and SV_{p1} are similar, but the respective lower LoA are differently affected by the single outlier (3942.5, -683). SQ_{p2} appears to be most robust to few outliers due to its definition. The `R` source code for the derivation of these nonparametric LoA as well as the example data can be found as Supplemental Material S3 and S4, respectively.

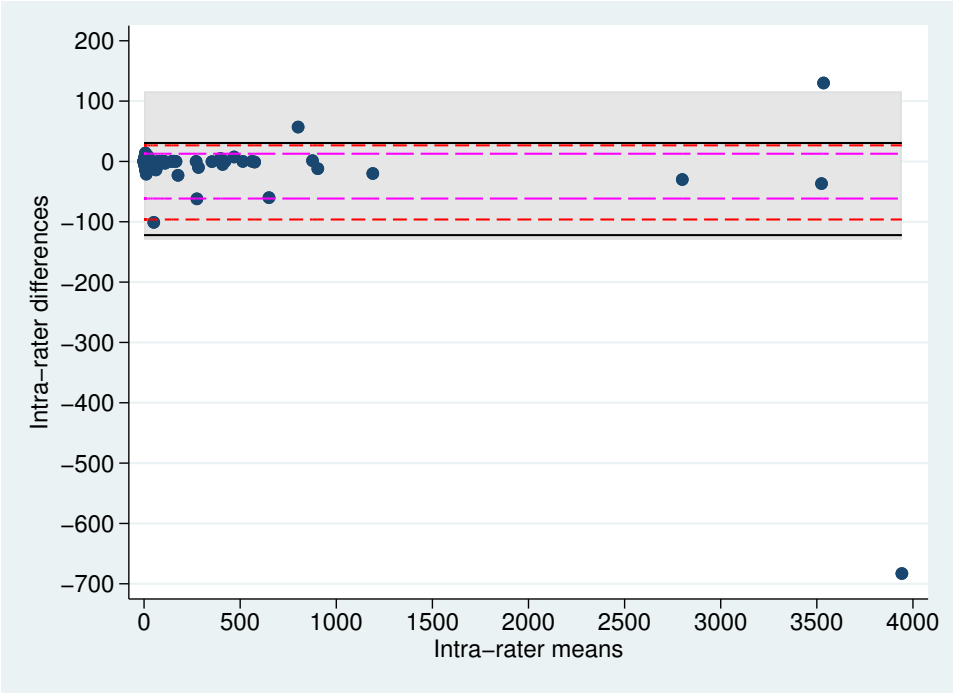

Figure 1. SQ_{p2} (magenta, long dashes), HD_p (red, short dashes) and SV_{p1} (black, solid lines) contrasted with classical BA LoA (shaded area).

The sensitivity of the classical BA LoA to outliers becomes crystal-clear when excluding the single outlier here. Subsequently, the symmetric BA LoA are -37.4 and 34.3, and the estimated mean difference reduces to -1.6 (results not shown here). In practice, outliers are, though, kept in the analysis dataset if there is no reasonable explanation for an exclusion. This underlines the importance of robust alternatives to the BA LoA.

5. Discussion

5.1. Statement of Principal Findings

The simple sample quantile estimators that are based on one and two order statistics performed closest to the nominal level in terms of the coverage probability for the next observation across six distributional scenarios for $n = 30$ and $n = 50, 80, 100, 150$. The Harrell–Davis subsampling estimator and estimators of the Sfakianakis–Verginis type followed closely for sample sizes of at least $n = 80$ and may be considered intuitively more appealing, as they use the entire sample, whereas more simple and outlier-robust sample quantile estimators are only based on a few observations from the sample.

5.2. Strengths and Limitations of The Study

The choice of distributions for the simulation study was motivated by our own experience with agreement assessments in clinical studies, especially roughly normal distributions with a few percent outliers. We investigated a wide range of quantile estimators, comprising sample, subsampling,

and kernel quantile estimators as well as other methods for quantile estimation with sample sizes between 30 and 150. As a measure of performance, we considered the coverage probability of nonparametric LoA for the next observation, interpreting the nonparametric LoA as a prediction interval, as the lower and upper LoA need to be simultaneously assessed. Therefore, we did not pursue evaluations using, for instance, mean squared errors.

5.3. Strengths and Limitations in Relation to other Studies

A peculiarity of LoA is the sole focus on the 2.5th and 97.5th quantiles, two extreme quantiles. Dielman, Lowry and Pfaffenberger [9] investigated 0.02 and 0.98, but only for small samples ($n = 10, 15, 25, 30$). Others examined 0.05 and 0.95 [11,14,16,21,22,29], whereas Kaigh and Cheng [17,18] assessed 0.1 and 0.9. Only Huang and Brill [26,27] also targeted 0.025 and 0.975, but only for samples of maximum size $n = 30$. Sfakianakis and Verginis [28] analyzed 0.01 and 0.99 as well as 0.05 and 0.95 in various sample sizes.

When compared to the usual number of 2000 iterations, the chosen number of 20,000 iteration runs translated for a given nonparametric estimator and sample size into a reduced range of the coverage probabilities across distributions by approximately 0.005 and is deemed appropriately accurate. However, the increased number of iterations did implicate considerably longer running times in creating the data for one Table (12 as opposed to 2 h). The abovementioned studies employed between 1000 and 10,000 iterations [16,22].

Harrell and Davis [14] did not recommend HD_p for small n and extreme p, and Dielman, Lowry and Pfaffenberger [9] concluded that there was not one best estimator across scenarios, based on maximum sample sizes of 30 and 60; however, Dielman, Lowry and Pfaffenberger [9] suggested that HD_p performs well in a wide range of cases, except when $p = 0.02, 0.98$. Our findings for HD_p are in line with these former conclusions, but extend to larger sample sizes of $n = 80, 100, 150$, in which HD_p appears to be a preferable choice for estimating extreme quantiles.

5.4. Meaning of the Findings: Possible Mechanisms and Implications

Our findings suggest using SQ_{p1} in small samples with approximately $n = 30$ but SQI_p, SV_{p1}, or SV_{p3} may be preferential alternatives as SQ_{p1} simply reduces to the smallest and largest observations as estimates for the 2.5% and 97.5% quantiles, respectively. The latter is, in turn, unfortunate in the case of outliers due to their unabated impact on the estimates. SQ_{p2} performed closest to the nominal level for all samples with $n \geq 50$ and appeared to be less prone to the single outlier in our clinical example than HD_p and SV_{p1}. However, the latter two estimators do involve all the observations. SQ_{p2} can, therefore, be considered the first choice for samples of approximately $n = 50$, but, for larger n, both HD_p and Sfakianakis–Verginis type estimators are equally applicable and actually preferable if the researcher seeks to include the entire dataset in quantile estimation and not only pairs of order statistics.

The normality assumption of the paired differences may often be considered to be reasonable in the planning stage; however, alternative quantile estimators should be equally specified in the planning stage as empirical distributions may deviate notably from ideal assumptions. Moreover, our investigation suggests several beneficial nonparametric alternatives to BA LoA instead of the simple percentile estimators that currently seem to prevail.

5.5. Unanswered Questions and Future Research

In the case of normally distributed paired differences, Bland and Altman [1,2] have already proposed approximate confidence intervals for the BA LoA. Recently, Vock [30] emphasized that only a tolerance interval or the outer confidence limits for BA LoA can provide a range that will contain a specified percentage of future differences with a known certainty. Carkeet and Goh [31,32] proposed exact confidence intervals for BA LoA, while using two-sided tolerance factors for a normal distribution.

In the case of any given distribution for the paired differences, several approaches for the construction of nonparametric confidence intervals for quantiles have been proposed over

half a century [33–39]. For both HD_p and KL_p, confidence intervals for quantiles have been proposed [10,14,16]. In the context of nonparametric LoA, future research will naturally lie in the proposal and evaluation of confidence intervals for the 2.5% and 97.5% quantiles in small-to-moderate samples, especially with regard to SQ_{p2}, SQI_p, HD_p, and Sfakianakis–Verginis type estimators.

Regression procedures for method comparison analysis have not been considered here [40–43]. Robust methods designed for data configurations with outliers, such as S- or MM-estimation, Least Trimmed Squares, or the Forward Search, are of interest in this context [44–48].

Supplementary Materials: The following are available online at http://www.mdpi.com/2571-905X/3/3/22/s1, Code S1: `R` source code for generating Tables 1–5. Figure S2: Histogram for the data of the clinical example in Section 4, including an approximating normal distribution. Code S3: `R` source code for generating Limits of Agreement for the clinical example in Section 4. Data S4: Dataset of the clinical example in Section 4.

Author Contributions: Conceptualization, O.G.; methodology, H.C.P. and O.G.; software, M.E.F.; validation, M.E.F. and O.G.; formal analysis, M.E.F., H.C.P. and O.G.; writing—original draft preparation, M.E.F. and O.G.; writing—review and editing, M.E.F., H.C.P. and O.G.; visualization, O.G.; supervision, H.C.P. and O.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Axel Diederichsen (Odense University Hospital, Denmark) for the permission to reanalyze fully anonymized data from the DanRisk study, three anonymous reviewers for very helpful comments on earlier versions of the manuscript, and Editage (www.editage.com) for English language editing.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307–310, doi:10.1016/S0140-6736(86)90837-8.
2. Bland, J.M.; Altman, D.G. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135–160, doi:10.1191/096228099673819272.
3. Rosner, B. *Fundamentals of Biostatistics*, 8th ed.; Cengage Learning: Boston, MA, USA, 2015.
4. Schmitz, S.; Krummenauer, F.; Henn, S.; Dick, H.B. Comparison of three different technologies for pupil diameter measurement. *Graefe's Arch. Clin. Exp. Ophthalmol.* **2003**, *241*, 472–477, doi:10.1007/s00417-003-0669-x.
5. Twomey, P.J. How to use difference plots in quantitative method comparison. *Ann. Clin. Biochem.* **2006**, *43*, 124–129, doi:10.1258/000456306776021616.
6. Diederichsen, A.C.; Sand, N.P.; Nørgaard, B.; Lambrechtsen, J.; Jensen, J.M.; Munkholm, H.; Aziz, A.; Gerke, O.; Egstrup, K.; Larsen, M.L.; et al. Discrepancy between coronary artery calcium score and HeartScore in middle-aged Danes: The DanRisk study. *Eur. J. Prev. Cardiol.* **2012**, *19*, 558–564, doi:10.1177/1741826711409172.
7. Cheng, C. On Estimation of Quantiles and Quantile Density Functions. Ph.D. Thesis, Texas A & M University, College Station, TX, USA, 1993.
8. Parzen, E. Nonparametric statistical data modeling. *J. Am. Stat. Assoc.* **1979**, *74*, 105–121, doi:10.1080/01621459.1979.10481621.
9. Dielman, T.; Lowry, C.; Pfaffenberger, R. A comparision of quantile estimators. *Commun. Stat. Simul. Comput.* **1994**, *23*, 355–371, doi:10.1080/03610919408813175.
10. Steinberg, S.M. Confidence Intervals for Functions of Quantiles Using Linear Combinations of Order Statistics. Ph.D. Thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1983.
11. Parrish, R.S. Comparision of quantile estimators in normal sampling. *Biometrics* **1990**, *46*, 247–257, doi:10.2307/2531649.
12. Hyndman, R.J.; Fan, Y. Sample quantiles in statistical packages. *Am. Stat.* **1996**, *50*, 361–365, doi:10.1080/00031305.1996.10473566.
13. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1980.
14. Harrell, F.E.; Davis, C.E. A new distribution-free quantile estimator. *Biometrika* **1982**, *69*, 635–640, doi:10.1093/biomet/69.3.635.

15. Steinberg, S.M.; Davis, C.E. Comparison of nonparametric point estimators for interquantile differences in moderate sized samples. *Commun. Stat. Theory Methods* **1987**, *16*, 1607–1616, doi: 10.1080/03610928708829457.
16. Kaigh, W.D.; Lachenbruch, P.A. A generalized quantile estimator. *Commun. Stat. Theory Methods* **1982**, *11*, 2217–2238, doi:10.1080/03610926208828383.
17. Kaigh, W.D. Quantile interval estimation. *Commun. Stat. Theory Methods* **1983**, *12*, 2427–2443, doi:10.1080/03610928308828610.
18. Kaigh, W.D.; Cheng, C. Subsampling quantile estimators and uniformity criteria. *Commun. Stat. Theory Methods* **1991**, *20*, 539–560, doi:10.1080/03610929108830514.
19. Cheng, C. The Bernstein polynomial estimator of a smooth quantile function. *Stat. Probab. Lett.* **1995**, *24*, 321–330, doi:10.1016/0167-7152(94)00190-J.
20. Delampady, M.; Ghosh, J.K.; Samanta, T. *An Introduction to Bayesian Analysis Theory and Methods*; Springer: New York, NY, USA, 2006.
21. Yang, S.S. A smooth nonparametric estimator of a quantile function. *J. Am. Stat. Assoc.* **1985**, *80*, 1004–1111, doi:10.1080/01621459.1985.10478217.
22. Sheather, S.J.; Marron, J.S. Kernel quantile estimators. *J. Am. Stat. Assoc.* **1990**, *85*, 410–416, doi:10.1080/01621459.1990.10476214.
23. Zelterman, D. Smooth nonparametric estimation of the quantile function. *J. Stat. Plan. Inference* **1990**, *26*, 339–352, doi:10.1016/0378-3758(90)90136-I.
24. Nadaraya, E.A. Smooth regression analysis. *Sankhyā Indian J. Stat.* **1964**, *26*, 359–372.
25. Watson, G.S. On estimating regression. *Theory Probab. Appl.* **1964**, *9*, 141–142.
26. Huang, M.L.; Brill, P. A level crossing quantile estimation method. *Stat. Probab. Lett.* **1999**, *45*, 111–119, doi:10.1016/s0167-7152(99)00049-8.
27. Huang, M.L. On a distribution-free quantile estimator. *Comput. Stat. Data Anal.* **2001**, *37*, 477–486, doi:10.1016/S0167-9473(01)00020-2.
28. Sfakianakis, M.E.; Verginis, D.G. A new family of nonparametric quantile estimators. *Commun. Stat. Simul. Comput.* **2008**, *37*, 337–345, doi:10.1080/03610910701790491.
29. Navruz, G.; Özdemir, A.F. A new quantile estimator with weights based on a subsampling approach. *Br. J. Math. Stat. Psychol.* **2020**, *73*, doi:10.1111/bmsp.12198.
30. Vock, M. Intervals for the assessment of measurement agreement: Similarities, differences, and consequences of incorrect interpretations. *Biom. J.* **2016**, *58*, 489–501, doi:10.1002/bimj.201400234.
31. Carkeet, A. Exact parametric confidence intervals for Bland-Altman limits of agreement. *Optom. Vis. Sci.* **2015**, *92*, e71–e80, doi:10.1097/OPX.0000000000000513.
32. Carkeet, A.; Goh, Y.T. Confidence and coverage for Bland-Altman limits of agreement and their approximate confidence intervals. *Stat. Methods Med. Res.* **2018**, *27*, 1559–1574, doi:10.1177/0962280216665419.
33. Chu, J.T. Some uses of quasi-ranges. *Ann. Math. Stat.* **1957**, *28*, 173–180.
34. Campbell, M.J.; Gardner, M.J. Calculating confidence intervals for some non-parametric analyses. *Br. Med. J.* **1988**, *296*, 1454–1456, doi:10.1136/bmj.296.6634.1454.
35. Beran, R.; Hall, P. Interpolated nonparametric prediction intervals and confidence intervals. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 643–652.
36. Hutson, A.D. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *J. Appl. Stat.* **1999**, *26*, 343–353, doi:10.1080/02664769922458.
37. Hutson, A.D. 'Exact' bootstrap confidence bands for the quantile function via Steck's determinant. *J. Comput. Graph. Stat.* **2002**, *11*, 471–482, doi:10.1198/106186002760180626.
38. Zielinski, R.; Zielinski, W. Best exact nonparametric confidence intervals for quantiles. *Statistics* **2005**, *39*, 67–71, doi:10.1080/02331880412331329854.
39. Balakrishnan, N.; Li, T. Confidence intervals for quantiles and tolerance intervals based on ordered ranked set samples. *Ann. Inst. Stat. Math.* **2006**, *58*, 757–777, doi:10.1007/s10463-006-0035-y.
40. Cornbleet, P.J.; Gochman, N. Incorrect least-squares regression coefficients in method-comparison analysis. *Clin. Chem.* **1979**, *25*, 432–438, doi:10.1093/clinchem/25.3.432.
41. Passing, H.; Bablok, W. A new biometrical method for testing the equality of measurements from two different analytical methods. *Clin. Chem. Lab. Med.* **1983**, *21*, 709–720, doi:10.1515/cclm.1983.21.11.709.
42. Passing, H.; Bablok, W. Comparison of several regression procedures for method comparison studies and determination of sample size. *Clin. Chem. Lab. Med.* **1984**, *22*, 431–445, doi:10.1515/cclm.1984.22.6.431.

43. Payne, R.B. Method comparison: Evaluation of least squares, Deming and Passing/Bablok regression procedures using computer simulation. *Ann. Clin. Biochem.* **1997**, *34*, 319–320, doi:10.1177/000456329703400317.
44. Rousseeuw, P.J. Least median of squares regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880, doi:10.1080/01621459.1984.10477105.
45. Yohai, V.J.; Zamar, R. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J. Am. Stat. Assoc.* **1988**, *83*, 406–413, doi:10.1080/01621459.1988.10478611.
46. Riani, M.; Cerioli, A.; Atkinson, A.C.; Perrotta, D. Monitoring robust regression. *Electron. J. Stat.* **2014**, *8*, 646–677, doi:10.1214/14-EJS897.
47. Rousseeuw, P.; Perrotta, D.; Riani, M.; Hubert, M. Robust monitoring of time series with application to fraud detection. *Econom. Stat.* **2019**, *9*, 108–121, doi:10.1016/j.ecosta.2018.05.001.
48. Riani, M.; Atkinson, A.C.; Corbellini, A.; Perrotta, D. Robust regression with density power divergence: Theory, comparisons, and data analysis. *Entropy* **2020**, *22*, 399, doi:10.3390/E22040399.