Statistica Sinica Preprint No: SS-2016-0061R2	
Title	Estimation of quantiles from data with additional measurement errors
Manuscript ID	SS-2016-0061R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0061
Complete List of Authors	Matthias Hansmann and Michael Kohler
Corresponding Author	Matthias Hansmann
E-mail	hansmann@mathematik.tu-darmstadt.de

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

Statistica Sinica

ESTIMATION OF QUANTILES FROM DATA WITH ADDITIONAL MEASUREMENT ERRORS

Matthias Hansmann and Michael Kohler

Technische Universität Darmstadt

Abstract: In this paper we study the problem of estimating quantiles from data that contain additional measurement errors. The only assumption on these errors is that the average absolute measurement error converges to zero for sample size tending to infinity with probability one. In particular we do not assume that the measurement errors are independent with expectation zero. We show that the empirical measure based on the data with measurement errors leads to an estimator which approaches the quantile set asymptotically. Provided the quantile is uniquely determined, this implies that this quantile estimate is strongly consistent for the true quantile. If this assumption does not hold, we also show that we can construct estimators for the limits of the quantile set if the average absolute measurement error is bounded by a given sequence, that tends to zero for sample size tending to infinity with probability one. But if such a sequence, which upper bounds the measurement errors, is not given, we show that there exists no estimator that is consistent for every distribution of the underlying random variable and all data containing the measurement errors. We derive the rate of convergence of our estimator and show that the derived rate of conver-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

gence is optimal. The results are applied in simulations and in the context of experimental fatigue tests.

Key words and phrases: Consistency, experimental fatigue tests, quantile estimation, rate of convergence.

1. Introduction

Let X be a real-valued random variable with cumulative distribution function (cdf.) F, i.e., $F(x) = \mathbf{P}\{X \leq x\}$. For $\alpha \in (0,1)$ let

$$Q_{X,\alpha} := \{z \in \mathbb{R} : \mathbf{P}\left(X \leq z\right) \geq \alpha \quad \text{and} \quad \mathbf{P}\left(X \geq z\right) \geq 1-\alpha\}$$

be the set of all α-quantiles of X. More precisely, we have

$$Q_{X,\alpha} = \left[q_{X,\alpha}^{[low]}, q_{X,\alpha}^{[up]}\right],$$

where $q_{X,\alpha}^{[low]} := \min\left\{z \in \mathbb{R} : F\left(z\right) \geq \alpha\right\}$ is the lower α-quantile and $q_{X,\alpha}^{[up]} := \sup\{z \in \mathbb{R} : F\left(z\right) \leq \alpha\}$ is the upper α-quantile. The estimation of this set, or its limits $q_{X,\alpha}^{[low]}$ and $q_{X,\alpha}^{[up]}$, is well-researched in the literature. For example, a simple idea to estimate $q_{X,\alpha}^{[low]}$ from a sample $X_1, \ldots, X_n$ of X is to use $X_1, \ldots, X_n$ to compute the empirical cdf.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}} \tag{1.1}$$

and to estimate the quantile by the corresponding plug-in estimate

$$\hat{q}_{X,n,\alpha} = \min\{z \in \mathbb{R} : F_n(z) \geq \alpha\}, \tag{1.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

which is in fact an order statistics (Arnold, Balakrishnan, and Nagaraja (1992)).

In this paper we assume that we have available only data $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \tag{1.3}$$

We do not assume anything on the measurement errors $\bar{X}_{i,n} - X_i$ $(i = 1, \ldots, n)$, in particular, these errors do not need to have expectation zero. They also do not need to be random and, in case that they are random, they do not need to be independent or identically distributed. Particularly, it is not assumed that these errors are independent of the i.i.d. data or that their distribution is known, so estimates for convolution problems (see, e.g., Meister (2009) and the literature cited therein) are not applicable in our context. Note also that our set-up is triangular.

The consideration of additional measurement errors is motivated by experimental fatigue tests from the Collaborative Research Center 666 at the Technische Universität Darmstadt, where we have to use measured data from other similar materials to estimate quantiles of number of cycles until failure for a certain material (cf., Section 3 below).

Measurement errors of this type have been recently considered in the context of distribution estimation (cf., Bott, Devroye, and Kohler (2013)),

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

nonparametric regression with random design (cf., Kohler (2006)), and nonparametric regression with fixed design (cf., Furer, Kohler, and Krzyżak (2013), Furer and Kohler (2015)).

Since we do not assume anything on the nature of the measurement errors besides being asymptotically negligible in the sense that (1.3) holds, it seems natural to ignore them completely and to try to use the same estimates as in the case that an independent and identically distributed sample is given. We investigate whether the corresponding quantile estimates are still consistent in this situation and how their rate of convergence depends on

$$\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_{i,n}|.$$

But first, consider results of quantile estimation with i.i.d. data, without additional measurement errors. If the quantile is uniquely determined, $\hat{q}_{X,n,\alpha} \to q_{X,\alpha}^{[low]}$ $a.s.$ (cf., e.g., Theorem 2.2. in Puri and Ralescu (1986)). We show that this result also holds if data with the above mentioned measurement error is used instead of the i.i.d. data (see Corollary 1 below).

In case the quantile is not uniquely determined, $\hat{q}_{X,n,\alpha}$ is no longer a strong consistent estimate of $q_{X,\alpha}^{[low]}$ (cf., e.g., Theorem 1 in Feldman and Tucker (1966)), but it is possible to find a suitable sequence α_n, such that $\hat{q}_{X,n,\alpha_n}$ is a strong (or weak) consistent estimator for $q_{X,\alpha}^{[low]}$ for all distri-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

butions of the random variable X (cf. Theorem 4 (or 5) in Feldman and Tucker (1966)). If we use data with measurement errors for the quantile estimation, one cannot find a sequence α_n such that $\hat{q}_{\bar{X},n,\alpha_n}$ is a strong consistent estimator of $q_{X,\alpha}^{[low]}$ for all distributions of X and all corresponding data with measurement error fullfilling (1.3). There does not even exist a general estimator that is strongly consistent for all distributions of X and all corresponding data with measurement error fullfilling (1.3) (Theorem 3). Should we know an upper bound on the average measurement error that tends to zero almost surely for sample size tending to infinity, it is possible to find sequences α_n and β_n, such that $\hat{q}_{\bar{X},n,\alpha_n}$ and $\hat{q}_{\bar{X},n,\beta_n}$ are strongly consistent estimators of $q_{X,\alpha}^{[low]}$ and $q_{X,\alpha}^{[up]}$, respectively (Theorem 2).

The rate of convergence of quantile estimates with i.i.d. data can be derived from the asymptotic theory of order statistics (cf., e.g., Mosteller (1946), Smirnov (1952), and Bahadur (1966)). Then if the cdf. F of X is continuous and differentiable at $q_{X,\alpha}^{[low]}$ with derivative greater than zero we have

$$\sqrt{n} \cdot F'(q_{X,\alpha}^{[low]}) \cdot \frac{\hat{q}_{X,n,\alpha} - q_{X,\alpha}^{[low]}}{\sqrt{\alpha \cdot (1-\alpha)}} \to \mathcal{N}(0,1) \quad \text{in distribution} \tag{1.4}$$

(cf., e.g., Theorem A on page 77 in Serfling (1980)). Reiss (1974) investigated the accuracy of this normal approximation. Since (1.4) holds, we

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

have

$$|\hat{q}_{X,n,\alpha} - q_{X,\alpha}^{[low]}| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{1.5}$$

where we write $X_n = O_{\mathbf{P}}(Y_n)$ if the nonnegative random variables X_n and Y_n satisfy $\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\{X_n > c \cdot Y_n\} = 0$. We investigate how additional measurement errors influence the rate of convergence of our quantile estimates. In Theorem 4 it is shown that if the average additional measurement error is bounded above by some $\eta_n \geq 0$, then our estimate achieves a rate of convergence of order

$$\frac{1}{\sqrt{n}} + \sqrt{\eta_n}. \tag{1.6}$$

We show in Theorem 5 that it is in general not possible to derive a better rate of convergence.

Throughout this paper the following notation is used: The sets of positive natural numbers and real numbers are denoted by $\mathbb{N}$ and $\mathbb{R}$, respectively. For a real number x, we denote by $\lfloor x \rfloor$ and $\lceil x \rceil$ the largest integer less than or equal to x and the smallest integer larger than or equal to x, respectively. We write $\to^{\mathbf{P}}$ as an abbreviation for convergence in probability and I_A for the indicator function on the set A.

The outline of the paper is as follows: The main results are formulated in Section 2 and proven in the supplementary materials. In Section 3 we

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

illustrate the finite sample size performance of our estimates by applying them to simulated data, and we describe an application of our estimates in the context of experimental fatigue tests.

2. Main results

Let $\bar{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\bar{X}_{i,n}\le x\}}$ be the empirical cumulative distribution function corresponding to $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$, and let $\hat{q}_{\bar{X},n,\alpha} = \min\{z \in \mathbb{R} : \bar{F}_n(z) \ge \alpha\}$ be the corresponding plug-in quantile estimate.

2.1. Strong consistency

First we investigate whether the estimator $\hat{q}_{\bar{X},n,\alpha}$ approaches the quantile set $Q_{X,\alpha}$ asymptotically.

Theorem 1. *Let $X, X_1, X_2 \ldots$ be independent and identically distributed real-valued random variables and let $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ be random variables that satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \tag{2.1}$$

If $\alpha \in (0,1)$ is arbitrary, then dist $\left(\hat{q}_{\bar{X},n,\alpha}, Q_{X,\alpha}\right) \to 0$ a.s., where dist $(x, A) := \inf_{a\in A} |x - a|$ for $x \in \mathbb{R}$ and a set $A \subset \mathbb{R}$.

Corollary 1. *Under the assumptions of Theorem 1 assume the α-quantile is uniquely determined. Then $\hat{q}_{\bar{X},n,\alpha} \to q_{X,\alpha}^{[low]}$ a.s.*

Proof. The uniqueness of the α-quantile implies $q_{X,\alpha}^{[up]} = q_{X,\alpha}^{[low]}$ and therefore

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

$Q_{X,\alpha} = \{q_{X,\alpha}^{[low]}\}$. The assertion follows directly by Theorem 1. □

Remark 1. The uniqueness of the α-quantile is necessary for obtaining $\hat{q}_{\bar{X},n,\alpha} \to q_{X,\alpha}^{[low]} \quad a.s.$ Without it, the case $q_{X,\alpha}^{[low]} < q_{X,\alpha}^{[up]}$ with $F(x) = \alpha$ for $x \in \left[q_{X,\alpha}^{[low]}, q_{X,\alpha}^{[up]}\right)$ is possible. In this case we get for i.i.d. data without measurement errors $\mathbf{P}\left(\hat{q}_{X,n,\alpha} \leq q_{X,\alpha}^{[low]} \ i.o.\right) = \mathbf{P}\left(\hat{q}_{X,n,\alpha} \geq q_{X,\alpha}^{[up]} \ i.o.\right) = 1$, where i.o. means infinitely often (cf., e.g., Theorem 1 in Feldman and Tucker (1966)). This implies that $\hat{q}_{\bar{X},n,\alpha} \to q_{X,\alpha}^{[low]} \quad a.s.$ cannot hold in this case.

Theorem 1 tells us under which conditions $\hat{q}_{\bar{X},n,\alpha}$ converges $a.s.$ towards the set $Q_{X,\alpha}$. Estimating the lower bound $q_{X,\alpha}^{[low]}$ of this set by $\hat{q}_{\bar{X},n,\alpha}$ is only possible under a suitable condition on F. It is possible to drop this condition, if we replace α by an appropriate sequence α_n, and if we know an upper bound η_n of the average absolute measurement error, that tends to zero almost surely as n tends to infinity. This approach extends Theorem 4 in Feldman and Tucker (1966) to data that contains additional measurement errors.

Theorem 2. *Let $X, X_1, X_2 \ldots$ be independent and identically distributed real-valued random variables with cdf. F and let $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ be random*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

variables which satisfy

$$\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}_{i,n}|\leq\eta_n \quad a.s. \tag{2.2}$$

for some $\eta_n\geq 0$ *satisfying* $\eta_n\to 0$ *a.s. Let* $\alpha\in(0,1)$ *be arbitrary. With*

$$\alpha_n=\alpha-2\sqrt{\frac{2\log\left(\log\left(n/2\right)\right)}{n}}-\sqrt{\eta_n} \quad \textit{and} \quad \beta_n=\alpha+2\sqrt{\frac{2\log\left(\log\left(n/2\right)\right)}{n}}+\sqrt{\eta_n},$$

$$\hat{q}_{\bar{X},n,\alpha_n}\to q_{X,\alpha}^{[low]} \quad a.s. \quad \textit{and} \quad \hat{q}_{\bar{X},n,\beta_n}\to q_{X,\alpha}^{[up]} \quad a.s.$$

Remark 2. The term $2\sqrt{\frac{2\log(\log(n/2))}{n}}$ in the definition of the sequences α_n and β_n in Theorem 2 can be replaced by any c_n satisfying $c_n\to 0$ as $n\to\infty$ and

$$c_n\geq(1+\nu)\sqrt{\frac{2\log\left(\log\left(n/2\right)\right)}{n}}$$

for some $\nu>0$.

It is natural to ask whether there exists a sequence α_n such that $\hat{q}_{\bar{X},n,\alpha_n}$ is a strong consistent estimator of $q_{X,\alpha}^{[low]}$ for all distributions of X and all random variables $\bar{X}_{1,n},...,\bar{X}_{n,n}$ satisfying (2.1). The answer is no, even if the sample with measurement errors does not change each time when the sample size changes, i.e., if we have given data $\bar{X}_1,...,\bar{X}_n$.

Theorem 3. *Let* $\alpha\in(0,1)$ *be arbitrary. There does not exist a sequence* $\left(\hat{q}_{n,\alpha}\right)_{n\in\mathbb{N}}$ *of quantile estimates satisfying* $\hat{q}_{n,\alpha}\left(\bar{X}_1,\ldots,\bar{X}_n\right)\to^{\mathbf{P}} q_{X,\alpha}^{[low]}$ *for*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

all real-valued random variables X and all random variables $\bar{X}_1, \ldots, \bar{X}_n$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_i| \to 0 \quad a.s. \tag{2.3}$$

for some independent $X_1, X_2, \ldots$ that have the same distribution as X.

Remark 3. Analogously, it is possible to show that there does not exist a sequence $(\hat{q}_{n,\alpha})_{n\in\mathbb{N}}$ of quantile estimates satisfying $\hat{q}_{n,\alpha}\left(\bar{X}_1, \ldots, \bar{X}_n\right) \to^{\mathbf{P}} q_{X,\alpha}^{[up]}$ under the same conditions.

2.2. Rate of convergence

Theorem 4. *Let $X, X_1, X_2 \ldots$ be independent and identically distributed real-valued random variables with cdf. F and let $\bar{X}_{1,n}, \ldots, \bar{X}_{n,n}$ be random variables that satisfy*

$$\eta_n := \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_{i,n}| \to 0 \quad a.s. \tag{2.4}$$

Let $\alpha \in (0,1)$ be arbitrary and assume that F is continuous at $q_{X,\alpha}^{[low]}$ and

$$\left|F(q_{X,\alpha}^{[low]}) - F(x)\right| \geq c_2 \cdot \left|q_{X,\alpha}^{[low]} - x\right|^{\gamma} \tag{2.5}$$

for all $\left|q_{X,\alpha}^{[low]} - x\right| \leq \zeta$, for some finite constants $c_2, \gamma, \zeta > 0$. Then

$$\left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}^{[low]}\right| = O_{\mathbf{P}}\left(\left(\frac{1}{\sqrt{n}}\right)^{1/\gamma} + (\sqrt{\eta_n})^{1/\gamma} + \sqrt{\eta_n}\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

One sees then that for $\gamma \leq 1$,

$$\left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}^{[low]}\right| = O_{\mathbf{P}}\left(\left(\frac{1}{\sqrt{n}}\right)^{1/\gamma} + \sqrt{\eta_n}\right)$$

and for $\gamma > 1$

$$\left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}^{[low]}\right| = O_{\mathbf{P}}\left(\left(\frac{1}{\sqrt{n}}\right)^{1/\gamma} + (\sqrt{\eta_n})^{1/\gamma}\right).$$

Under the assumption that F is differentiable at $q_{X,\alpha}^{[low]}$ with derivative greater than zero, (2.5) holds with $\gamma = 1$, yielding the $1/\sqrt{n}$ of the rate of convergence in Theorem 4. This is known from the rate of convergence of the order statistics with i.i.d. data without errors (see (1.5)). Because of (1.4) it is not possible to improve this part of the convergence rate by an asymptotically faster decreasing sequence. It is also known that an order statistic is asymptotically most concentrated about its distribution quantile in comparison with all other translation-equivariant and asymptotically uniformly median unbiased estimators (cf., Corollary 2 in Pfanzagl (1976)). The $\sqrt{\eta_n}$ in the convergence rate is due to the measurement errors of the data. We investigate whether the rate $\sqrt{\eta_n}$ is the best rate one can obtain for $\gamma = 1$.

Theorem 5. *Let $\alpha \in (0,1)$ be arbitrary. Under the assumptions of Theorem 4 with $\gamma = 1$, for every estimator $\hat{q}_{n,\alpha}$ there exists a random variable X and random variables $\bar{X}_{1,n}, ..., \bar{X}_{n,n}$ satisfying*

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}_{i,n}\right| \to 0 \quad a.s.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

for some independent $X_1, X_2, \ldots$ that have the same distribution as X such that

$$\left|\hat{q}_{n,\alpha} - q_{X,\alpha}^{[low]}\right| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}} + \tilde{\eta}_n\right)$$

does not hold, whenever $\tilde{\eta}_n$ is a sequence for which

$$\frac{\tilde{\eta}_n}{\sqrt{\eta_n}} \to^{\mathbf{P}} 0.$$

3. Application to simulated and actual data

In this section we consider estimates of 5%-, 50%-, 90%-, and 95%-quantiles. We first consider distributions with known quantiles in order to investigate our estimates, then we apply our estimator to experimental fatigue test data.

3.1. Application to simulated data

In our simulated data, we used $n = 500, 1000$, and 2000 samples. To reduce the randomness contained in the quantile estimates due to the random number generation, we repeated the quantile estimation 100 times and we indicate the quantile estimate by an upper index i. We compared the quantile estimates by considering the mean value (MV) $\frac{1}{100}\sum_{i=1}^{100}\hat{q}^i$ and the mean squared error (MSE) $\frac{1}{100}\sum_{i=1}^{100}\left(\hat{q}^i - q_{X,\alpha}^{[low]}\right)^2$.

We first chose $X, X_1, X_2, \ldots$ to be independent $\mathcal{N}(0,1)$ and $\bar{X}_{i,n} = X_i + \frac{1}{n}E_i$, where $E_1, \ldots, E_n$ are samples from an exponential with expectation

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

$\lambda = 10$. Thus, we generated new samples with measurement error with change in n. We also considered $\bar{Y}_{i,n} = X_i + \frac{1}{i}E_i$, where samples with bigger measurement errors were retained. Here $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}_{i,n}\right| \to 0$ $a.s.$ and $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{Y}_{i,n}\right| \to 0$ $a.s.$ Since the cdf. of X is strictly increasing, the estimators $\hat{q}_{\bar{X},n,\alpha}$ and $\hat{q}_{\bar{Y},n,\alpha}$ are strongly consistent for $q_{X,\alpha}^{[low]}$. This can be seen in Table 1 for $\alpha = 0.9$ and $\alpha = 0.95$. The estimator $\hat{q}_{\bar{X},n,\alpha}$ shows, even for $n = 500$, estimates with a small mean squared error. The estimator $\hat{q}_{\bar{Y},n,\alpha}$ convergenced more slowly.

We next chose $X, X_1, X_2, \ldots$ as independent and identically distributed with $\mathbf{P}(X = 0) = \mathbf{P}(X = 1) = \frac{1}{2}$. Setting $\alpha = 0.5$ leads to the lower quantile $q_{X,\alpha}^{[low]} = 0$. The mean value and the mean squared error of $\hat{q}_{X,n,\alpha}$ are shown in Table 2. The estimator $\hat{q}_{X,n,\alpha}$ is obviously not strongly consistent for $q_{X,\alpha}^{[low]}$. However, by Theorem 2 we can modify our estimate to $\hat{q}_{X,n,\alpha_n}$ with $\alpha_n = \alpha - 2\sqrt{\frac{2\log(\log(n/2))}{n}}$. As seen in Table 1, this modification leads to a perfect estimation of $q_{X,\alpha}^{[low]}$. But, if we use the data $\bar{X}_{i,n} = X_i + \frac{B_i}{5n^{0.1}}$, where $B_1, ..., B_n$ are i.i.d. samples from a $b(1, \frac{1}{2})$, $\hat{q}_{\bar{X},n,\alpha_n}$ shows much larger errors, as seen in Table 2. Since we can bound $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}_{i,n}\right|$ by $\frac{1}{5n^{0.1}}$, Theorem 2 has it that we can get a consistent esimator if we choose the sequence $\gamma_n = \alpha - 2\sqrt{\frac{2\log(\log(n/2))}{n}} - \sqrt{\frac{1}{5n^{0.1}}}$ and consider the estimator $\hat{q}_{\bar{X},n,\gamma_n}$. The results in Table 2 show that this estimator approximates the quantile well.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

Table 1: Simulation results for $X, X_1, X_2, ...$ independent $\mathcal{N}(0,1)$, $\bar{X}_{i,n} = X_i + \frac{1}{n}E_i$ and $\bar{Y}_{i,n} = X_i + \frac{1}{i}E_i$, where $E_1, ..., E_n$ are samples from an exponential with expectation $\lambda = 10$.

	90%-quantile $q_{X,\alpha}^{[low]} = 1.2816$			95%-quantile $q_{X,\alpha}^{[low]} = 1.6449$		
size of n	500	1000	2000	500	1000	2000
MV for $q_{\bar{X},n,\alpha}$	1.2931	1.2901	1.2894	1.6730	1.6597	1.6473
MSE for $q_{\bar{X},n,\alpha}$	0.0066	0.0030	0.0013	0.0128	0.0047	0.0023
MV for $q_{\bar{Y},n,\alpha}$	1.4217	1.3691	1.3275	1.8289	1.7507	1.7029
MSE for $q_{\bar{Y},n,\alpha}$	0.0248	0.0105	0.0036	0.0436	0.0164	0.0057

We then chose $X, X_1, X_2, \ldots$ as independent uniforms on $(0,1)$. As our data with additional measurement error we took $\bar{X}_{i,n} = X_i + \frac{1}{n^{0.25}}$, so that $\eta_n = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}_{i,n}\right| \to 0$ $a.s.$ We computed the absolute error

$$d_n = \left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}\right|$$

for $\alpha = 0.9$ and sample sizes n in steps of 200. As illustrated in Figure 1, the absolute error shows approximately the same asymptotic behaviour as $\frac{1}{\sqrt{n}} + \eta_n$ in this case. Thus there exists data with measurement error such that a faster convergence rate than $\frac{1}{\sqrt{n}} + \sqrt{\eta_n}$ is obtained.

It is also possible to construct data with measurement errors such that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

Table 2: Simulation results for $X, X_1, X_2, \ldots$ independent $b\left(1, \frac{1}{2}\right)$ and $\bar{X}_{i,n} = X_i + \frac{B_i}{5n^{0.1}}$, where $B_1, ..., B_n$ are i.i.d. samples from a $b\left(1, \frac{1}{2}\right)$.

	50%-quantile $q_{X,\alpha}^{[low]} = 0$		
size of n	500	1000	2000
MV for $q_{X,n,\alpha}$	0.4000	0.4400	0.4800
MSE for $q_{X,n,\alpha}$	0.4000	0.4400	0.4800
MV for q_{X,n,α_n}	0	0	0
MSE for q_{X,n,α_n}	0	0	0
MV for $q_{\bar{X},n,\alpha_n}$	0.1074	0.1002	0.0935
MSE for $q_{\bar{X},n,\alpha_n}$	0.0115	0.0100	0.0087
MV for $q_{\bar{X},n,\gamma_n}$	0	0	0
MSE for $q_{\bar{X},n,\gamma_n}$	0	0	0

the absolute error of the estimator behaves asymptotically as the claimed rate $\frac{1}{\sqrt{n}} + \sqrt{\eta_n}$ from Theorem 4:

As a last example, we chose $\alpha = 0.9$ and $X, X_1, X_2, \ldots$ as in the previ-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

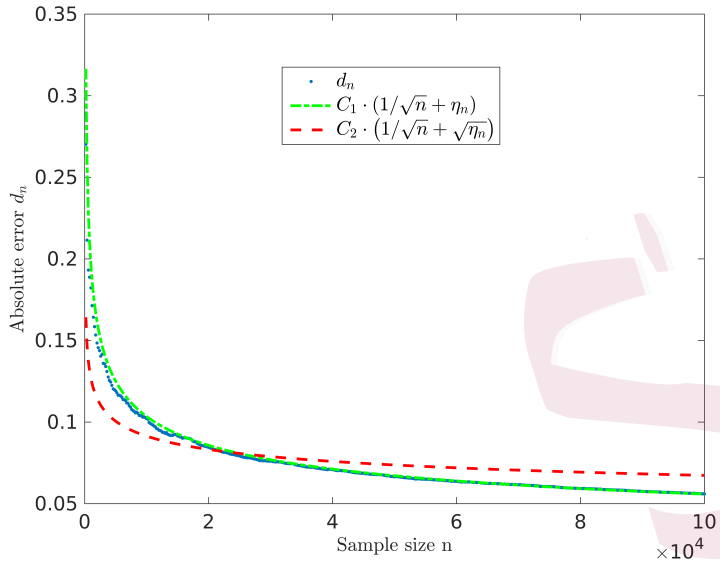

Figure 1: Typical asymptotic behaviour of $d_n = \left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}\right|$ in the setting of our third example.

ous example and

$$\bar{X}_{i,n} = \begin{cases} X_i + \frac{1}{n^{0.25}} & \text{if } X_i \in \left[\alpha - \frac{1}{n^{0.25}}, \alpha\right] \text{ and } X_i \text{ is one of the } \lfloor \frac{1}{n^{0.25}} \cdot n \rfloor \\ & \text{largest samples of } (X_j)_{j=1,\ldots,n} \text{ in } \left[\alpha - \frac{1}{n^{0.25}}, \alpha\right] \\ X_i, & \text{otherwise.} \end{cases}$$

Here

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n} \left|X_i - \bar{X}_{i,n}\right| \leq \frac{1}{n} \cdot \lfloor \frac{1}{n^{0.25}} \cdot n \rfloor \cdot \frac{1}{n^{0.25}} \to 0 \ \ a.s.$$

This leads to an absolute error d_n that has approximately the same asymptotic behaviour as $\frac{1}{\sqrt{n}} + \sqrt{\eta_n}$, as illustrated in Figure 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

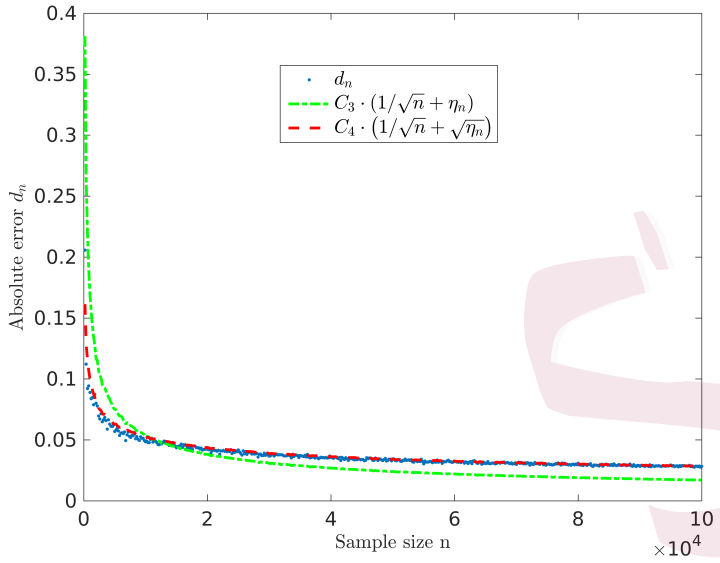

Figure 2: Typical asymptotic behaviour of $d_n = \left|\hat{q}_{\bar{X},n,\alpha} - q_{X,\alpha}\right|$ in the setting of the last example.

3.2. Application to data

We applied our methods in the context of fatigue behaviour of steel under cyclic loading. This was motivated by experiments of the Collaborative Research Center 666 at the Technische Universität Darmstadt, which studies integral sheet metal design with higher order bifurcations. Here the main idea is to produce structures out of one part by linear flow and bend splitting, which has several advantages concerning the material properties. Our main goal was to study, whether this modified, splitted material shows better fatigue behavior under cyclic loading than the base material.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

Therefore, for each material m, data

$$\left\{\left(\epsilon_1^{(m)}, \left(N_1^{(m)}, \tau_1^{(m)}\right)\right), ..., \left(\epsilon_{l_m}^{(m)}, \left(N_{l_m}^{(m)}, \tau_{l_m}^{(m)}\right)\right)\right\}$$

was obtained by a series of experiments, in which for a strain amplitude $\epsilon_i^{(m)}$ the number of cycles $N_i^{(m)}$ until failure and the corresponding stress amplitude $\tau_i^{(m)}$ was determined. We had available a database of 132 materials, and 1222 data points in total. The data were used to compare the estimated 5%−quantiles of the number of cycles until failure from the modified and the base material of ZStE500 for different strain amplitudes ϵ. Thus, we were interested in estimating the number of cycles such that no failure occurs, with a probability of approximately 95%. Since the experiments are very time consuming, we only had available 4 to 35 data points per material, not enough for a nonparametric estimation. To nevertheless estimate the quantile of the number of cycles until failure, we assumed

$$N^{(m)}(\epsilon) = \mu^{(m)}(\epsilon) + \sigma^{(m)}(\epsilon) \cdot \delta, \tag{3.1}$$

to hold, where $\mu^{(m)}(\epsilon)$ is the expected number of cycles until failure and $\sigma^{(m)}(\epsilon)$ is the standard deviation for each material m and strain amplitude ϵ; δ is an error term with expectation zero. We estimated the α−quantile of δ as well as $\mu^{(m)}(\epsilon)$ and $\sigma^{(m)}(\epsilon)$, so that we could estimate the α−quantile of $N^{(m)}(\epsilon)$ by a simple linear transformation. Thus, we used a similar

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

approach as in Bott and Kohler (2015):

To estimate the expected number of cycles $\mu^{(m)}(\epsilon)$, we applied a standard-method from the literature (cf., Williams, Lee, and Rilly (2002)), that uses the measured data to estimate the coefficients $p = \left(\sigma_f', \epsilon_f', b, c\right)$ of the strain life curve according to Coffin-Morrow-Manson (cf., Manson (1965)) by linear regression, and to estimate $\mu^{(m)}(\epsilon)$ from the corresponding strain life curve.

For the estimation of the standard deviation $\sigma^{(m)}(\epsilon)$, we augmented our data points for every material m by 100 artifical ones, as in Furer and Kohler (2015), and weighted the Nadaraya-Watson kernel regression estimates applied to the real and the artificial data.

Thus, we determined the data samples

$$\hat{\delta}_i^{(m)} = \frac{N_i^{(m)} - \hat{\mu}_i^{(m)}}{\hat{\sigma}_i^{(m)}} \quad \text{for } i = 1, ..., l_m \text{ and all materials } m$$

of the random variable δ. These samples contained measurement errors because we only estimated $\mu^{(m)}(\epsilon)$ and $\sigma^{(m)}(\epsilon)$. Since we assumed in (3.1) that δ does not depend on the material m, we used all data samples to estimate the $\alpha-$quantile $\hat{q}_{\delta,\alpha}$ of δ and get an estimation of the $\alpha-$quantile of $N^{(m)}(\epsilon)$ by the transformation $\hat{q}_{N^{(m)},\alpha}(\epsilon) = \hat{\sigma}^{(m)}(\epsilon) \cdot \hat{q}_{\delta,\alpha} + \hat{\mu}^{(m)}(\epsilon)$.

The estimated quantiles of $N^{(m)}(\epsilon)$ for $\epsilon \in [0\%, 0.25\%]$ for the modified and the base material are illustrated in Figure 3. One can see that the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

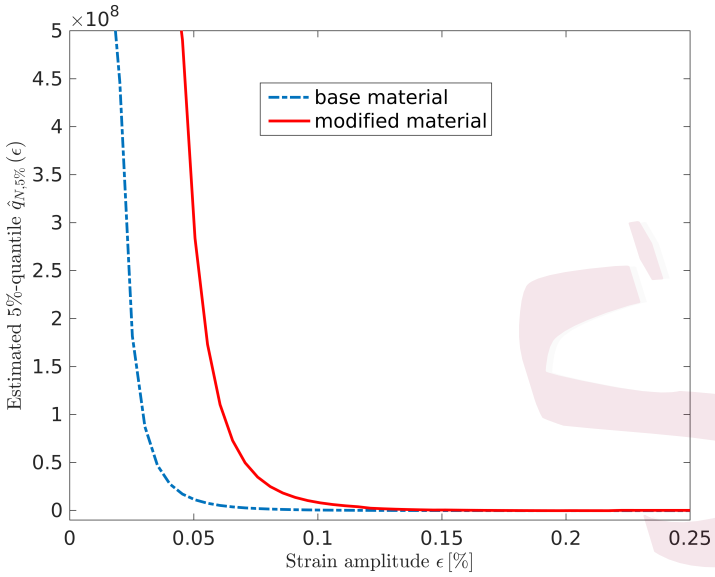

Figure 3: Comparison of the estimated 5%−quantiles of the number of cycles until the failure occurs $\hat{q}_{N,5\%}$ from the base and the modified material of ZSTE500. Here the strain amplitude is divided by the length of the material sample used in the experiments.

material shows much better fatigue behaviour after the flow splitting, which confirms the conjecture that the strain hardening occuring during the flow splitting improves the fatigue behaviour of materials.

Supplementary Materials

Proofs of Theorems 1 to 5 can be found in the supplementary materials.

Acknowledgements

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Center 666.

The authors would also like to thank an associate editor and the referee for their helpful comments. In particular, the referee's idea to assume the Hölder condition on F in Theorem 4 instead of the differentiability led to a valuable generalization of the result.

References

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1992). *A First Course in Order Statistics*. John Wiley & Sons.

Bahadur, R. R. (1966). A Note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577-580.

Bott, A., Devroye, L., and Kohler, M. (2013). Estimation of a distribution from data with small measurement errors. *Electronic Journal of Statistics* **7**, 2457-2476.

Bott, A. and Kohler, M. (2015). Nonparametric estimation of a conditional density. To appear in *Annals of the Institute of Statistical Mathematics*.

Feldman, D. and Tucker, H. G. (1966). Estimation of non-unique quantiles. *Ann. Math. Statist.* **37**, 451-457.

Furer, D. and Kohler, M. (2015). Smoothing spline regression estimation based on real and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

artificial data. *Metrika* **78**, 711-746.

Furer, D., Kohler, M., and Krzyżak, A. (2013). Fixed design regression estimation based on real and artificial data. *Journal of Nonparametric Statistics* **25**, 223-241.

Kohler, M. (2006). Nonparametric regression with additional measurement errors in the dependent variable. *Journal of Statistical Planning and Inference* **136**, 3339-3361.

Manson, S. S. (1965). Fatigue: A complex subject - some simple approximation. *Experimental Mechanics* **5**, 193-226.

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics, **193**. Springer, Berlin.

Mosteller, F. (1946). On some useful "inefficient" statistics. *Ann. Math. Statist.* **17**, 377-408.

Pfanzagl, J. (1976). Investigating the quantile of an unknown distribution. In *Contributions to Applied Statistics (Ziegler, W. J., ed.)*, 111-126. Birkhäuser, Basel.

Puri, M. L. and Ralescu, S. S. (1986). Limit theorems for random central order statistics. In *Adaptive Statistical Procedures and Related Topics* **8**, 447-475, Institute of Mathematical Statistics, Hayward, CA.

Reiss, R.-D. (1974). On the accuracy of the normal approximation for quantiles. *Ann. Probab.* **2**, 741-744.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Smirnov, N. V. (1952). Limit distributions for the terms of a variational series. *American Math-*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0061

ematical Society Translations **11**, 82-143.

Williams, C. R., Lee, Y.-L., and Rilly, J. T. (2002). A practical method for statistical analysis of strain-life fatigue data. *International Journal of Fatigue* **25**, 427-436.

Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7,

64289 Darmstadt, Germany.

E-mail: hansmann@mathematik.tu-darmstadt.de

Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7,

64289 Darmstadt, Germany.

E-mail: kohler@mathematik.tu-darmstadt.de