Rakenteiden Mekaniikka (Journal of Structural Mechanics)
Vol. 57, No. 3, 2024, pp. 99–125
http://rakenteidenmekaniikka.journal.fi
https://doi.org/10.23998/rm.144320

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Estimating quantiles in acceptance criteria for structural components

Matti Pajari[1] and Lasse Makkonen

Summary The probabilistic foundations of methods to control the characteristic properties of structural materials, expressed as p-quantiles, are discussed. We argue that the acceptance criteria in the quality control should be based on quantile estimators complying with the definition of the quantile. We introduce the distribution-free concept of *definition-based quantile estimator*. For normal distribution, an application of different types of estimators, as well as some attribute methods and mixed methods presently used, are illustrated by operating characteristic curves. We recommend the prediction method by which the subjective limit values of the mixed methods are eliminated, information from the sample will be used more effectively than in the attribute methods, and the questions about the proper confidence level or a “known” variation coefficient need not be considered. However, the direct application of the prediction method results in stricter quality control than that presently used. Therefore, we recommend adopting $p' > p$ in such a way that if the predicted p'-quantile estimate $\hat{x}_{p'}$ is equal to or higher than the lower lilmit L, the required value of the p-quantile, the lot is accepted in the quality control. If the present quality level is appropriate, the value of p' can be chosen in such a way that the present level is maintained. If not, necessary modifications are possible simply by adjusting p’.

Key words: quantile, quantile estimator, quality control, acceptance criterion, small sample, confidence level, structural safety

Received: 17 March 2024. Accepted: 12 October 2024. Published online: 21 October 2024.

Introduction

Structural safety is characterized by the probability of failure which, when considering a single failure mechanism, is determined by two random variables: the resistance of the structure (R), and the load to be resisted (E). In the design philosophy of the European design codes called Eurocodes, both R and E are related to the characteristic values of the relevant variables. Typical characteristic values affecting the safety are the 0.05-quantile of the strength and the 0.98-quantile of the natural load. Both are estimated based on a limited number of observations. Since the estimated quantiles are used as input for

[1] Corresponding author: matti.pajari@berakon.fi

probabilistic safety considerations, it is important to use estimators which in the probabilistic sense are as correct as possible. For assorted reasons, discussed later, this is not presently done in industrial production.

While the probabilistic theory of structural safety has made great progress during the last eighty years, the knowledge of the strength of the structural materials is vague. In the factories, sample sizes of one to three with options of retesting are common. The uncertainty associated with small sample sizes and mild acceptance criteria enhances the risk of accepting products with defective quality. The main point of this article is to illuminate the statistical background of the acceptance criteria to provide standardization committees with easily understandable information. The methods to control the production process in such a way that the rate of rejection is limited to an economical level, see e.g. Schilling and Neubauer (2017), is not treated in this article.

For practical reasons, discrete strength classes, each with a specific lower limit L for the characteristic strength, have been used in structural design. In European design codes called Eurocodes, characteristic strength is defined as a p-quantile of the strength. p is a small probability, e.g. 0.05 or 0.10. A lot of products belongs to strength class C if the p-quantile of its strength distribution is at least as high as the lower limit L specified for C. For example, when a lot of rebars (reinforcing bars) belongs to class B500, the producer claims that the characteristic strength (0.05-quantile) of the rebars is greater than or equal to L = 500 MPa.

The producer's claim is verified by tests on a sample comprising n specimens randomly chosen from the lot. Based on the observed values $x_1,\ldots,x_\mathrm{n}$, different acceptance methods are available. Some of them estimate the p-quantile of the lot and compare the estimate $\hat{x}_p$ with the lower limit L, some others use more heuristic methods. This article aims at evaluating the probabilistic background of the former methods, revealing the weaknesses of the latter ones, and proposing what could be done to find a balance between the p'-quantile actually controlled and the p-quantile assumed in European design practice.

The acceptance methods may be classified in three categories: Attribute, variable, and mixed methods. In the attribute methods, a single on-off attribute is given to each inspected product. The product either works or not, a measured characteristic either exceeds the threshold or not, etc. In the variable methods, the property of the product is treated as a continuous random variable. For example, when the strength of three test specimens is measured and an estimate for the quantile considered is constructed using all measured values, we are using a variable method. If instead, we only check how many observed values exceed the threshold value, we are using an attribute method. When using an attribute method instead of a variable method, some information is lost. Setting a lower limit for the sample mean and another lower limit for the sample minimum is an example of mixed methods which include features from both attribute and variable methods.

We use the yield strength of rebars made of carbon steel as an example to demonstrate the acceptance methods. We also assume that the strength is normally distributed, and the samples are representative. It is a straight-forward process to extend the results to the lognormal distribution.

Even though there is no dispute about the meaning of a p-quantile, the acceptance (conformity) criteria in the national and international standards vary. In Europe, the national criterion for a lot of rebars is typically based on three tests and optional retests. A lot is accepted if none of the three measured strength values is below *L,* but under some provisions, one value lower than *L* is also acceptable.

In EN 10080 (2005) which is the European product standard for carbon steel reinforcement, the strength is understood as the long-term strength. This means that the products made during a long period, say six months, comprise the population and the sample consists of all test results over that period. The compliance with the strength class is determined based on this large sample. This involves a problem that if an unacceptable strength level is observed after a period of six months, there is no way to reject the products already delivered to the market. Such a method represents monitoring rather than control.

There is also a criterion for each lot, based on three tests. In the following, the acceptance criterion of EN 10080 refers to this lot-specific criterion because it controls the strength *in the structure* as understood in EN 1992-1-1 (2004), a part of the design standard collection for concrete structures called Eurocode 2. In EN 10080, the conformity criteria for the lots include nationally determined parameters. Since the probabilistic background of those parameters has been regarded as obscure, national specifications are still used. To facilitate unified European design, criteria for the strength control of the reinforcing steel have been included in Eurocode 2. Those criteria are relatively mild. In practice, they enable the design according to Eurocode 2 together with the national material standards, but they have little to do with the 0.05-quantile.

According to the standard ASTM 615-M20 (2020), the yield strength of a lot of reinforcing bars is acceptable when the yield strength measured from a *single specimen* is greater than or equal to the minimum yield strength *L* specified for the strength class, or when it is at most 7 MPa lower than *L*, and two additional test results are above it. Since a single test result gives no information of the probability distribution, and the same 7 MPa is used for all strength classes, proper statistical considerations based on quantiles are not possible. Therefore, neither ASTM 615-M20 nor any other standard applying acceptance criteria based on a single test and optional retests are considered here.

In the European concrete standard EN 206 (2005), the 0.05-quantile is estimated by a method which is not consistent, i.e. the estimator does not converge to the correct value even when the sample size increases without limit. Caspeele and Taerwe (2012) have proposed an improved method for cores drilled from existing structures, but this, as well as the method of EN 206, includes parameters based on engineering judgment rather than on statistics.

To sum up, one of the cornerstones of the European safety philosophy is the characteristic material strength expressed as a *p*-quantile of the strength, but the existing standards, such as ISO 12491 (1997), have not been followed in the acceptance criteria. This is not required in EN 1990 + A1 (2005), either, even though it can be understood as a guidance document for the safety issues.

This article has been provoked by the authors' experiences from the European standardization of construction products. It does not cover all existing acceptance methods and materials but underlines the principles of statistics that are not acknowledged in the present practice, clarifies some probabilistic concepts, recommends abandoning of misleading terminology, and encourages the profession to adopt simple and probabilistically unambiguous acceptance methods.

Operating characteristic curves

Based on a few observations, it is impossible to evaluate any quantile accurately. If the actual strength in the considered set of products, called a lot, is close to the required level, a small sample size in the quality control means a considerable risk, both of rejecting a conforming lot (type I error) and accepting a non-conforming lot (type II error). The former case is uneconomical for the producer, the latter unsatisfactory for the consumer.

A single product is said to be defective when its strength is lower than the limit value *L*. The *Acceptance Probability* (AP) is the probability that a lot is accepted when the chosen acceptance criterion is applied to a random sample taken from that lot. An operating characteristic curve (OC curve), see Fig. 1, presents AP as a function of the share of defective products in the lot. The OC curves depend both on the acceptance criterion and on the sample size. Under certain provisions, the OC curve does not depend on the parameters of the distribution.

The *limiting quality level* LQL means that a lot with the share of defective products greater than that at LQL should be rejected, and otherwise accepted. In Fig. 1, LQL corresponds to 5.0%. *Consumer's risk* is AP at LQL. At *acceptable* or *target quality level* (AQL) the *producer's risk* 1–AP is small enough to be economically acceptable to the producer. The *confidence level* α of an acceptance criterion can be defined as 1–AP at LQL. This is a more general definition than the confidence level defined later for the coverage estimator because 1–AP is also defined for criteria which do not include quantile estimator.

An ideal OC curve would be a step function that equals unity for a quality level better than or equal to LQL, and zero elsewhere. By increasing the sample size, see Fig. 2, the real OC curve becomes steeper and closer to the ideal one, but the risks of type I and type II errors never vanish. The curves in Figs 1 and 2 correspond to the prediction method described below.

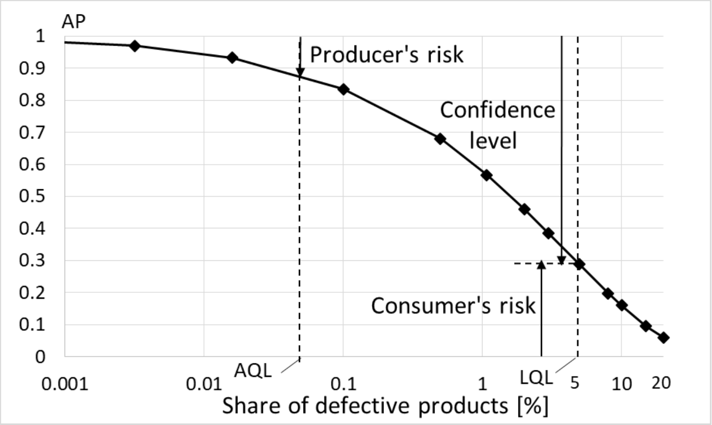

Figure 1. Typical OC curve.

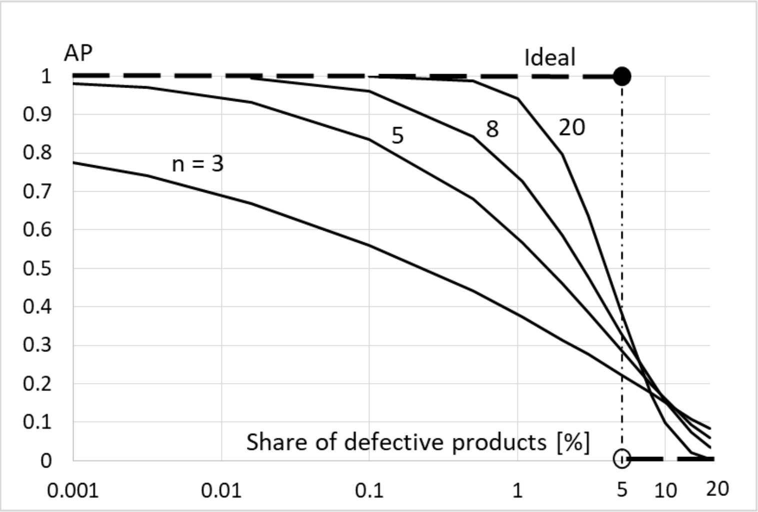

Figure 2. OC curves (prediction method for 0.05-quantile). Sample sizes 3, 5, 8 and 20. The dashed step curve is the ideal curve.

In theory the acceptance criterion could be defined by the producer and the consumer in such a way that both are satisfied. For structural bulk products, this is not realistic because other parties like the vendors, contractors, designers and building owners are involved. Furthermore, the strength control is a safety issue, and a subject to regulation by the authorities. Since it is impossible to agree on the terms of factory production control in each building project separately, standards and other specifications using different sample sizes and acceptance criteria have been developed.

Quantile estimators in quality control

General

By definition, x_p is the *p*-quantile of a random variable *X* if and only if

$$P\{x \leq x_p\} = p \qquad (1)$$

where *x* is a random outcome of *X*, see Fig. 3.

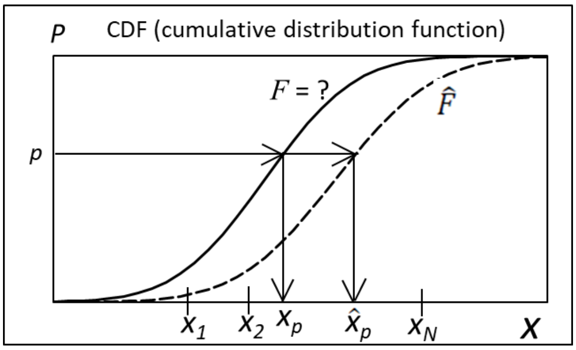

Figure 3. Determining *p*-quantile x_p and its estimate $\hat{x}_p$ from *F* and its estimate $\hat{F}$, respectively.

If F is the strictly monotonous cumulative distribution function (CDF) of a continuous X, $x_p = F^{-1}(p)$. For a normally distributed X with mean μ and standard deviation σ, x_p is obtained from

$$x_p = \mu + \phi^{-1}(p)\sigma \tag{2}$$

where ϕ is the CDF of the standardized normal distribution and ϕ^{-1} its inverse. In other words, for each μ and σ,

$$x_p = \mu + k(p)\sigma \tag{3}$$

where $k(p)$ depends on p but not on μ and σ. For the strength distribution of a lot to be evaluated, both μ and σ are unknown. Their estimates, sample mean $\bar{x}$ and sample standard deviation s, are calculated from strength values $x_1,\ldots,x_n$ measured in tests on n randomly chosen test specimens using Eqs (4) and (5)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{5}$$

Following Eq. (3), estimates of the type

$$\hat{x}_p = \bar{x} + k_n(p)s \tag{6}$$

have been used. Eq. (6) presents a rule T which maps the probability p and n observations x_i to a unique estimate $\hat{x}_p$ of x_p

$$T\colon (p, x_1, \ldots, x_n) \to \hat{x}_p = \bar{x} + k_n(p)s \tag{7}$$

Such rules are called *quantile estimators*. When L is the lower limit for x_p in quality control, an acceptance method or criterion can be expressed as:

If $\hat{x}_p \geq L$, the lot considered is accepted, otherwise rejected.

By varying $k_n(p)$, different estimators are obtained. Because each x_i contributing to $\bar{x}$ and s in Eq. (7) is an outcome of random variable X, $\bar{X}$ and S are random variables and so is also their function $\hat{X}_p$. It is often handy to regard $\hat{X}_p$ itself as a quantile estimator.

Definition-based quantile estimators

When U and V are two random variables and a is a constant, probabilities $P\{U \leq a\}$ and $P\{U \leq V\}$ are defined as follows. $P\{U \leq V\} = P\{u \leq v\}$ and $P\{U \leq a\} = P\{u \leq a\}$ where u and v are random outcomes of U and V, respectively. Using this notation, Eq. (1) or the definition of quantile x_p becomes

$$P\{X \leq x_p\} = p \tag{8}$$

A formally equivalent equation for $\hat{X}_p$ is

$$P\{X \leq \hat{X}_p\} = p \tag{9}$$

In Eq. (8) x_p is constant; in Eq. (9) $\hat{X}_p$ is a random variable which estimates x_p. This is the only difference. We call the quantile estimators complying with Eq. (9) *definition-based* (DBQE). When a DBQE is applied to n observations taken from an infinite population, one more observation x is taken and compared with $\hat{x}_p$, and this experiment is repeated *m* times, the share of steps fulfilling $x \leq \hat{x}_p$ approaches stochastically p with increasing *m*. In this sense $\hat{X}_p$ complying with Eq. (9) is a probabilistically ideal estimator.

In the following, Monte Carlo simulations with 10^5 cycles are performed using the free SageMath (2017) code to justify and illustrate some ideas and conclusions. Particularly

$$\frac{1}{M}\sum_{i=1}^{M} F(\hat{x}_{p,i}) \approx P\{X \leq \hat{X}_p\} = p' \tag{10}$$

with $M = 10^5$ is used to check which p'-quantile $\hat{X}_p$ actually estimates. See App. A for the justification of this distribution-free result. When $M \rightarrow \infty$, Eq. (10) implies that

$$P\{X \leq \hat{X}_p\} = E\left(F(\hat{X}_p)\right) = p' \tag{11}$$

where E denotes expectation. If p' is = p, $\hat{X}_p$ is a DBQE, and

$$E\left(F(\hat{X}_p)\right) = p \tag{12}$$

Furthermore, when Eq. (12) is true, $\hat{X}_p$ is a DBQE or $P\{X \leq \hat{X}_p\} = p$. This distribution-free result strongly underlines the importance of some early findings. Eq. (12) has been used e.g. to support the ideas that $\hat{X}_p$ fulfilling Eq. (12) is a good estimator for a p-quantile (Wilks (1941)), and $i/(n+1)$ is a good plotting position for order statistic $\hat{x}_{(i)}$ (Gumbel (1958)). However, this argument has left many later statisticians like Gringorten (1963), Cunnane (1978), Hyndman and Fan (1996) and Fuglem et al. (2013) unconvinced. Apparently unaware of Eq. (11), they have understood that Eq. (12) only represents one nice statistical characteristic without any major probabilistic role so that other criteria may also be appropriate. Madsen et al. (1986, 148–149) have shown that $P\{X \leq \hat{X}_{(i)}\} = i/(n+1)$, but this result has received little attention, and a belief on plotting positions better than $i/(n+1)$ is still reflected in many applications.

The accuracy of a DBQE increases with *n*. Wilks (1941) has introduced the concept of *tolerance limit*. When the estimator fulfills Eq. (12), $p_1 < p$ and $p_2 > p$, $0 < \beta < 1$ and the sample size *n* is high enough, $P\{p_1 < F(\hat{X}_p) < p_2\} = \beta$. $L_1 = F^{-1}(p_1)$ and $L_2 = F^{-1}(p_2)$ are called tolerance limits and β is the *two-sided confidence level*. A requirement for the accuracy of $\hat{X}_{0.05}$ might be $P\{0.045 < F(\hat{X}_{0.05}) < 0.055\} = 0.90$. In other words, when n is chosen high enough, 90% of the estimates $\hat{x}_{0.05}$ will be very close to $x_{0.05}$ on the probability scale.

Wilks' two-sided confidence level is a statistically sound measure for the accuracy of a quantile estimator. In the acceptance sampling of structural materials, the sample size is very small due to the costs of destructive testing. Therefore, strict tolerance limits cannot be set, and account of the sampling error needs to be taken by other means. Wilks' theory can be used for comparison of estimators by fixing first β, n, p_1 and p_2.

Weibull estimator

Acceptance methods defined by the criterion

> *If maximum m observations in sample $(x_1, \dots, x_n)$ are defective, the lot is accepted, otherwise rejected.*

are called attribute methods and denoted here Attr(*m*,*n*). The attribute methods need not estimate quantiles and can be applied to properties which are not continuous variables. For small sample sizes in destructive testing, only Attr(0,*n*), here called Weibull estimator $T_n^{wei}(1/(n+1))$, is relevant to the strength control of low quantiles. The corresponding acceptance method can also be called minimum value criterion: if the weakest observation is not defective, the lot is accepted.

If n observations $x_1, \dots, x_n$ sorted in ascending order are $x_1', \dots, x_n'$ and $p_i = i/(n+1)$, the Weibull estimator

$$T_n^{wei}: (\, p_i, x_1, \dots, x_n) \to \hat{x}_{p,i} = x_i' \qquad (13)$$

See Fig. 4, defines a distribution-free DBQE for n probabilities $p_i = i/(n+1)$, see Madsen et al. (1986, 148–149) and Makkonen et Pajari (2014). $x_1', \dots, x_n'$ are called order statistics. Superscript *wei* refers to the Weibull plotting positions p_i. Fig. 4 illustrates how T_n^{wei} can be extended to T_n^{wei+} by linear interpolation when $1/(n+1) \le p \le n/(n+1)$, but the interpolation results in an estimator which is only approximately definition-based. Neither T_n^{wei} nor T_n^{wei+} are defined for $p < 1/(n+1)$ and for $p > n/(n+1)$.

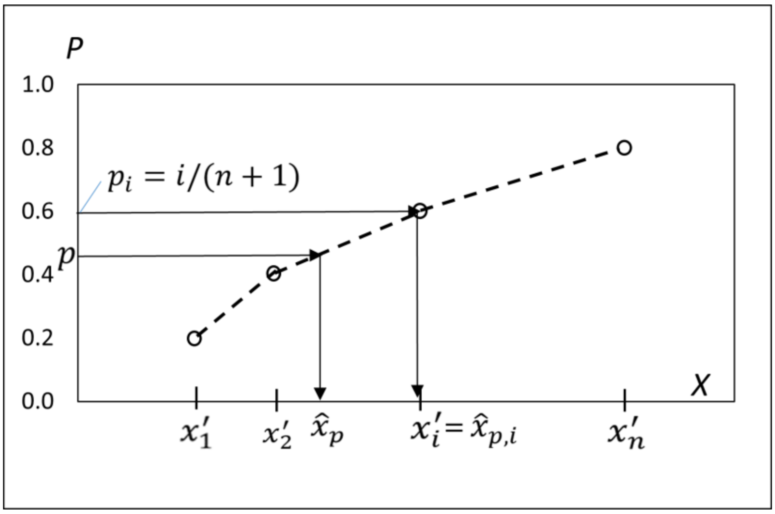

Figure 4. Weibull estimator $T_n^{wei}(i/(n+1))$ extended by interpolation to $T_n^{wei+}(p)$.

Prediction estimator

For a normally distributed X, application of the criterion (9) to Eq. (7) implies (see App. B) that

$$k_n(p) = k_n^{pre}(p) = t_{n-1}^{-1}(p)\sqrt{1+\frac{1}{n}} \tag{14}$$

where t_{n-1}^{-1} is the inverse of Student's t-function with n – 1 degrees of freedom. The estimator

$$T_n^{pre}: (p, x_1, \dots, x_n) \to \hat{x}_p = \bar{x} + k_n^{pre}(p)s \tag{15}$$

is definition-based and called *prediction estimator* in ISO 12491 (1997). Wilks (1941) has deducted expression (14) starting from Eq. (12). It is possible that Wilks regarded Eq. (11) as self-evident and compelling, but this has not been the case with all later statisticians. Even though the prediction method has been included in ISO standards, it has not been used widely.

ISO 12491 and Holicky (2013, 121–122) show that the prediction estimator is a special case of Bayesian estimators with no prior knowledge. Holicky also mentions that the prediction estimator fulfills Eq. (9). Caspeele and Taerwe (2012) call the prediction method "Bayesian method with vague prior information". As shown in Appendix B, the prediction estimator is a direct consequence of Eqs (7) and (9). Thus, no Bayesian justification is necessary.

Coverage estimator

Even though the prediction estimator is a DBQE, another estimator $T_n^{cov}(p, \alpha)$, which is not a DBQE, has been widely used for Gaussian variables. It is called *coverage estimator* because interval $[\hat{x}_p, \infty)$ covers the exact p-quantile x_p with probability or *one-sided confidence level* α. The coverage estimator is defined by

$$T_n^{cov}(p, \alpha): (\, p, x_1, \dots, x_n) \to \hat{x}_p = \bar{x} + k_n^{cov}(p, \alpha)s \tag{16}$$

where $k_n^{cov}(p, \alpha)$ is determined from

$$P\{\hat{X}_p \le x_p\} = \alpha \tag{17}$$

In the following, a shorter expression *confidence level* is used for α when there is no risk of mixing it with the two-sided confidence level β. The greater α, the lower is $\hat{x}_p$ and the larger is interval [$\hat{x}_p$,∞). Such an *interval estimate* is not optimal for the strength because the strength is input for the safety considerations in which a *point estimate* is needed. For this reason, $\hat{x}_p$ obtained from Eqs (16) and (17) is often used as a point estimate even though it is logically only the lower end of an interval. Criterion (17) provides no answer to the question: which α should be chosen? The intuitively appealing value 0.50 is not supported in the literature, and 0.75 is recommended e.g. by Holicky and ISO 12491, apparently without any other justification than the close relation of $k_n^{cov}(p, 0.75)$ to $k_n^{pre}(p)$.

Unlike Wilks' (two-sided) confidence level β, α defined by Eq. (17) does not reflect the average accuracy of the quantile estimator. For normal distribution, numeric values of $k_n^{cov}(p, \alpha)$ can be solved using noncentral *t*-distribution, see Madsen et al. (1986, 38–40) and a more detailed proof in App. B. A more general approach, also valid for non-Gaussian distributions, has been given by Zupan et al. (2007).

About known standard deviation or variation coefficient

In long-term production it may appear that, even though the sample mean varies with time due to the changes in the production parameters like raw material, machinery etc., the standard deviation of the sample is nearly constant, or remains below some upper limit. This suggests that the prediction estimator and the coverage estimator depend on the sample mean only. Applying criteria (9) and (17) would then result in k_p'-factors with considerably lower absolute values than those given in Table 1. We now compare the quantile estimators

$$\text{I} \qquad \hat{X}_p = \bar{X} + k_p'\sigma \qquad \sigma \text{ known} \tag{18}$$

$$\text{II} \qquad \hat{X}_p = \bar{X} + k_p S \qquad \sigma \text{ unknown} \tag{19}$$

It is unclear which of the cases I and II results in fewer rejected lots. If the sample size is small, the variation of S from one sample to another is high and the evaluation of σ is difficult, unless a value considerably higher than the mean of the observed s values is chosen. Increasing the sample size may justify a lower "known" σ, but it is both costly and reduces the difference between k_p and k_p'.

So far, there is no consensus on how to evaluate a known σ without conducting tests on large samples during an extended period. One possibility is to regard all observations from small samples, taken during an extended period, as one population, the standard deviation of which can be regarded as the known σ'. However, such a long-term σ' may be too conservative when applied to one lot, because the mean value varies from a lot to another due to the time-dependent changes in the production parameters. Consequently, it is possible that $\sigma' > s$ and $|k_p'\sigma'| \geq |k_p s|$ even though $k_p' < k_p$. This means that no advantage is gained by the assumption of a known σ.

Unbiased estimators

Quantiles have traditionally been estimated by first estimating the parameters of the assumed distribution function. The goodness of quantile estimators has been evaluated by criteria, such as unbiasedness or minimum root mean squared error of the distribution parameters. There is no probabilistic justification for such criteria, except when the estimated parameter is the mean. The same is true for using the unbiasedness of the quantiles themselves as a criterion, see Pajari & al. (2019). We argue that the quantile estimators shall be evaluated based on how well they conform to the definition of a quantile, i.e. Eq. (9).

Two common types of an unbiased quantile estimator are discussed in the literature, those unbiased relative to the distribution parameters and those unbiased relative to the quantile itself. In addition to the consistency and power, the unbiasedness has been regarded as one of the most relevant characteristics of a quantile estimator. The estimators used in quality control make an exception. For reasons unknown to the authors, the coverage estimator and the prediction estimator have been adopted in international standards despite their bias, obviously without major objections. The widely used Weibull estimator is also biased, but it has perhaps not been recognized as an estimator at all. However, during the last few years, the unbiasedness has been regarded as a goodness criterion for low-quantile estimators e.g. by Tur and Derechennik (2019).

The flaws in this kind of thinking have been discussed by Pajari et al. (2019). If a quantile estimator $\hat{X}_p$ (random variable) is unbiased and *Y* is its linear function, also *Y* is unbiased. In most practical cases the probability distribution *F(X)* is nonlinear and the unbiasedness of estimator $\hat{X}_p$ implies that its function $F(\hat{X}_p)$ is biased and vice versa. Eqs (11) and (12) show that if $\hat{X}_p$ is a DBQE, $F(\hat{X}_p)$ is unbiased when *F* is non-linear. Since $\hat{X}_p$ is non-linearly related to $F(\hat{X}_p)$ except when *X* is linearly distributed, $\hat{X}_p$ is biased, and if $\hat{X}_p$ is unbiased, $F(\hat{X}_p)$ is biased. Consequently, when the probability distribution is non-linear and $\hat{X}_p$ is unbiased, it is not a DBQE.

Even though the controversy between the DBQEs and unbiased quantile estimators is clear, it is interesting to get an impression of the magnitude of the difference. Some examples are presented below.

One might expect that

$$\hat{F}(x) = \frac{1}{\sqrt{2\pi s^2}} \int_{-\infty}^{x} e^{-\frac{(t-\bar{x})^2}{2s^2}} dt \tag{20}$$

is an appropriate estimate for the cumulative distribution function F of a normally distributed X because $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2, respectively. Eq. (20) implies that $\hat{x}_p$ is obtained from

$$\hat{x}_p = \hat{F}^{-1}(p) = \bar{x} + \phi^{-1}(p)\sqrt{s^2} \tag{21}$$

For p = 0.05, $\phi^{-1}(0.05) \approx$ 1.645 and

$$\hat{x}_{0.05} = \bar{x} + \phi^{-1}(0.05)s \tag{22}$$

represents an estimator $T_n^{un,par}$, unbiased with respect to the parameters μ and σ^2. Another estimator $T_n^{un,q}$, unbiased with respect to the quantile itself, is defined by

$$\hat{x}_{0.05,n} = \bar{x} + \xi_n \phi^{-1}(0.05)s \tag{23}$$

where the factor ξ_n > 1 is properly chosen, see Table 1. For small sample sizes, both unbiased estimators differ considerably from the prediction estimator.

Table 1. Factor ξ_n, see Eq. (22), for two unbiased estimators and the prediction estimator.

n	3	5	30	∞
$T_n^{un,par}(0.05)$	1.000	1.000	1.000	1.000
$T_n^{un,q}(0.05)$	1.128	1.064	1.009	1.000
$T_n^{pre}(0.05)$	2.050	1.420	1.050	1.000

Based on Table 1, unbiasedness with respect to quantile is slightly better than unbiasedness with respect to distribution parameters, but they both result in an estimator which is far from the DBQE.

Evaluating common estimators

The Weibull estimator T_n^{wei} with some modifications and the coverage estimator $T_n^{cov}(p)$ are often used in acceptance criteria. The prediction estimator $T_n^{pre}(p)$ is not so popular but serves as a benchmark due to its good properties. The other estimators discussed in this article behave so badly that they are excluded from this comparison.

Even though both $T_n^{wei}\left(\frac{1}{n+1}\right)$ and $T_n^{pre}\left(\frac{1}{n+1}\right)$ are DBQEs for the same p, their OC-curves differ, see Fig. 5. For a given *n*, the risk of both type I and type II error is smaller for T_n^{pre}, which exploits all measured data, while T_n^{wei} only uses the lowest order statistic x_1'. The difference is minimal for *n* = 3, but increases with *n*. The main disadvantage of T_n^{wei} as a quantile estimator is the fact that it is not applicable to $p < \frac{1}{n+1}$ and $p > \frac{n}{n+1}$. Interpolation for $\frac{1}{n+1} < p < \frac{n}{n+1}$ is possible but provides no advantage over the prediction estimator.

T_n^{wei} has been a starting point for numerous acceptance methods provided with additional rules which allow the acceptance of some lots which would be rejected by T_n^{wei} alone, see e.g. EN 1991–1–1 (2004, Annex C). Such acceptance methods are probabilistically difficult to interpret because their OC curves in general depend on the unknown standard deviation of the population.

Before evaluating the coverage estimator, the confidence level α must be fixed. Some values of k_n^{pre} and k_n^{cov} for *p* = 0.05 are given in Table 2. For each $k_n^{pre}(p)$ it is possible to find such an α that $k_n^{cov}(p,\alpha) = k_n^{pre}(p)$, and for each $k_n^{cov}(p,\alpha)$ it is possible to choose p' in such a way that $k_n^{pre}(p') = k_n^{cov}(p,\alpha)$.

Table 1 and the OC curves in Fig. 6 show that the prediction estimator and the coverage estimator are not far from each other when *p* = 0.05 and α = 0.75. From the producer’s point of view, there is no reason to prefer the coverage estimator at this confidence level when *n* > 3 because, when the quality is satisfactory, it results in a lower AP than the prediction estimator.

Fig. 7 compares the 0.05-quantiles estimated by the prediction estimator and the coverage estimator with four different α. As an example, when n = 3, the coverage estimators for $x_{0.05}$ with α = 0.50 and 0.90 actually estimate $x_{0.112}$ and $x_{0.022}$, respectively. When α = 0.75, the resulting error is relatively small for sample sizes 3, 4 and 5, but there is no point of using the coverage estimator and worrying about α when the prediction estimator is available for all sample sizes.

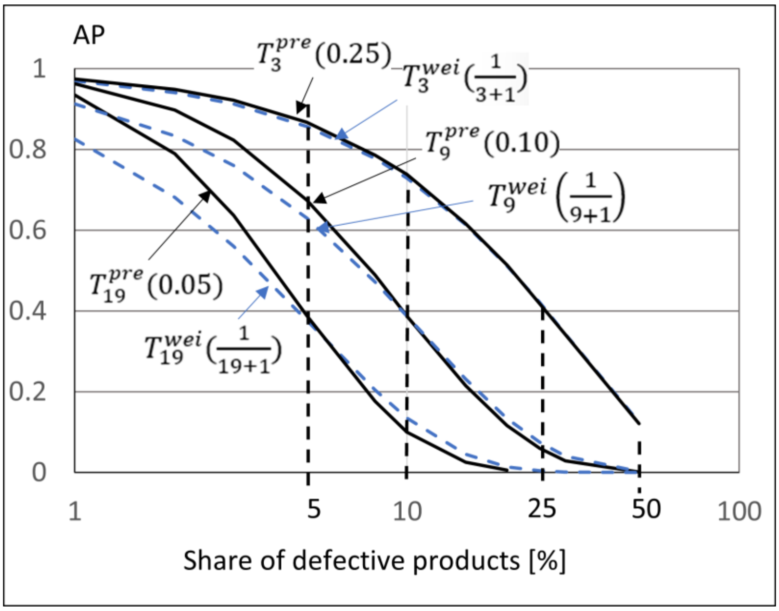

Figure 5. OC curves for the Weibull and prediction methods; p = 0.05, 0.10 and 0.25, n = 19, 9 or 3, respectively.

Table 2. Factors k_n for estimators $T_n^{pre}(p)$ and $T_n^{cov}(p,\alpha)$ when p = 0.05 and α = 0.50, 0.75 or 0.90.

Sample size n	3	4	5	6	8	10	20	30
$-k_n^{pre}(p)$	3.372	2.631	2.335	2.177	2.010	1.923	1.772	1.727
$-k_n^{cov}(p,0.50)$	1.938	1.830	1.779	1.750	1.719	1.702	1.671	1.662
$-k_n^{cov}(p,0.75)$	3.152	2.681	2.463	2.336	2.188	2.104	1.932	1.869
$-k_n^{cov}(p,0.90)$	5.312	3.957	3.400	3.092	2.755	2.569	2.208	2.080

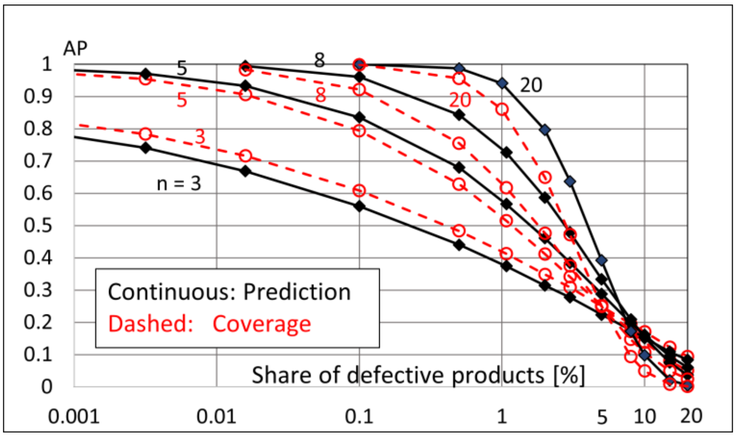

Figure 6. OC curves for prediction method and coverage method with confidence level α = 0.75 and p = 0.05; n is the sample size.

When n > 3 and α > 0.75, see Fig. 8, the average underestimation of the coverage estimator increases with increasing α, and $\hat{x}_{0,05} \rightarrow -\infty$ when $\alpha \rightarrow 1$. For very large samples α plays no role. For better safety, using α = 0.90 or even 0.95 instead of 0.75 have been recommended e.g. by ISO 12491. The safety is indeed improved by strong underestimation of the quantile, but the effect depends on the sample size, and the consequences on the structural safety are not easy to evaluate quantitatively. It would be more transparent to use the prediction method to control a lower quantile, for example $x_{0.02}$, and increase the sample size or use higher safety factors in design when extra safety is needed.

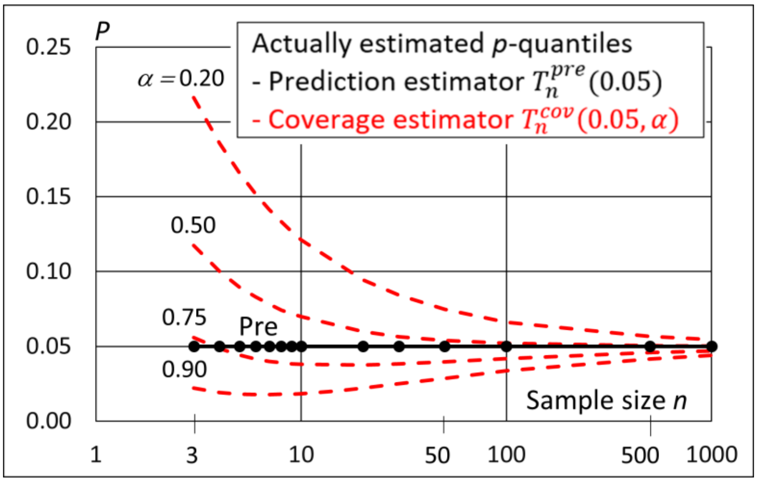

Figure 7. p-quantiles actually estimated by coverage estimator $T_n^{cov}(0.05; \alpha)$ and prediction estimator $T_n^{pre}(0.05)$.

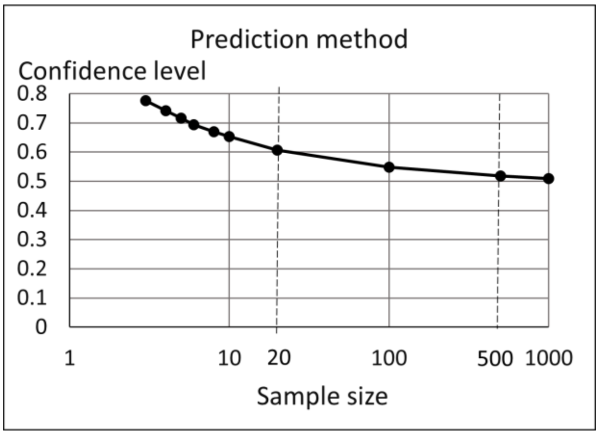

Figure 8. Confidence level of the prediction estimator until n = 1000. p = 0.05.

In the prediction estimator, the criterion $P\{X \leq \hat{X}_p\}$ fixes both k_p^{pre} and the confidence level $\alpha = P\{\hat{X}_p \leq x_p\}$. Fig. 8 depicts the α-n relationship of the prediction estimator for the 0.05-quantile. With increasing n, α approaches 0.50 from above.

The confidence level of the prediction method decreases with increasing n and accuracy of estimation. This underlines the fact that the confidence level is an inappropriate measure for the goodness of a quantile estimator. More importantly,

$P\{\hat{X}_p \le x_p\} = \alpha$ is a vague starting point for structural safety considerations, because such an $\hat{X}_p$ estimates *p'*-quantile which varies with *n*.

Fig. 9 illustrates Wilks' two-sided confidence level β for estimators $T_n^{pre}(0.05)$ and $T_n^{cov}(0.05, 0.75)$ for two tolerance limits defined by (p_1, p_2) = $(0.03{,}0.07)$ or $(0.04{,}0.06)$. As expected, the prediction estimator is the better one. When *n* = 3 or 10, the risk of falling outside the tolerance limits is higher than 90% or 70%, respectively, even when the better prediction estimator is used.

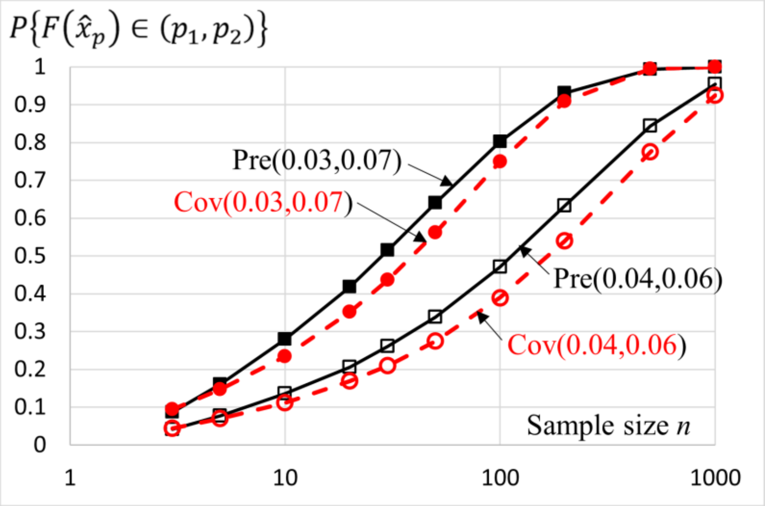

Figure 9. Probability of falling inside Wilks' tolerance limits when estimating 0.05-quantile with prediction estimator and coverage estimator with one-sided confidence level = 0.75.

Conclusion from the evaluation is clear: instead of the coverage estimator, the prediction estimator shall be used. It is also possible to choose the confidence level α in such a way that these two estimators become identical, but such an α depends on the sample size. A self-evident way to calibrate the coverage estimator is to abandon it entirely, and to use the prediction estimator instead.

Mixed methods

It is not uncommon to set two separate lower limits for acceptance: $\bar{x}_{lim}$ for the sample mean and $x_{lim,min}$ for the minimum value x_{min} of the sample. In this method both criteria $\bar{x} \ge \bar{x}_{lim}$ and $x_{min} \ge x_{lim,min}$ must be met simultaneously, but no *p*-quantile is estimated. Both $\bar{x}_{lim}$ and $x_{lim,min}$ depend on *p* and *n*. Such an acceptance method is called mixed because it has features both of a variable method ($\bar{x}_{lim}$) and an attribute method ($x_{lim,min}$).

One way to calibrate $\bar{x}_{lim}$ and $x_{lim,min}$ for a given *p* is to fix the value of acceptance probability AP at point *x* = *L*. Assume next that the acceptance probabilities $A_{P,mean} = P\{\bar{x} \ge \bar{x}_{lim}\}$ and $A_{P,min} = P\{x_{min} \ge x_{lim,min}\}$ are given, and the resulting overall acceptance probability is = AP. When $X \sim N(\mu, \sigma)$ and L is the limit for the *p*-quantile of *X*, see App. C,

$$x_{lim,min} = L + \left(-k_p + z_1\right)\sigma \qquad z_1 = \phi^{-1}\left(1 - A_{P,min}{}^{\frac{1}{n}}\right) \; (< 0) \tag{24}$$

$$\bar{x}_{lim} = L + \left(-k_p + z_2\right)\sigma \qquad z_2 = \frac{1}{\sqrt{n}}\phi^{-1}\left(1 - \; A_{P,mean}\right) \; (< 0) \tag{25}$$

Whichever A_{P}-values are chosen, the limits depend both on n and on the unknown standard deviation σ of the lot to be evaluated.

As an example, consider the yield strength of reinforcing steel bars and set p = 0.05, n = 5 and L = 500 MPa (in the following, the numeric strength values are expressed in dimensionless form). Even if the exact value of σ is unknown, it is possible to fix an interval which is likely to cover the exact value. Assume that $15 \leq \sigma \leq 25$, which suggests that $\sigma = \sigma_0$ = 20 would be a reasonable compromise. We try to determine $\bar{x}_{lim}$ and $x_{lim,min}$ with σ_0 = 20 in such a way that, when x_p of the population corresponds to the limiting quality level L, the mixed method yields the same AP = 0.286 as the prediction method. Using MC simulation, it comes out that $A_{P,mean} = \; A_{P,min}$ = 0.408 results in AP = 0.284 which is close enough to 0.286. In the following, Mix$_{\mathrm{Y}}$(Z) denotes a mixed method OC curve with limit values calculated using n = 5, p = 0.05, and σ_0 = Y, applied to a lot with true σ = Z.

Fig. 10 depicts the limiting OC curves Mix$_{20}$(15) and Mix$_{20}$(25) as well as Mix$_{15}$(15) = Mix$_{20}$(20). The real OC curve depends on the actual σ, but if the assumption $15 \leq \sigma \leq 25$ is true, it is somewhere between Mix$_{20}$(15) and Mix$_{20}$(25). The big difference between these curves shows that the mixed method, unlike the prediction method, coverage method and Weibull method, is not objective: AP is different for two populations with the same share of defective products but different σ. The risk of rejecting good and accepting bad lots is then pronounced, unless the limits of the true σ can be determined more precisely. The prediction method ($T_5^{pre}(0.05)$) gives a curve which intersects both Mix$_{20}$(15) and Mix$_{20}$(25) curves. Instead of constant σ_0, the sample standard deviation s may be used for $\bar{x}_{lim}$ and $x_{lim,min}$ which then become lot-specific. It is much simpler to use the prediction method than to find appropriate $A_{P,mean}$ and $A_{P,min}$ for each lot separately, and then apply Eqs (24) and (25).

Assume next that $\sigma = \sigma_0$ = 20 is known exactly. This knowledge also affects the prediction method by reducing $k_5^{pre}(0.05)$ = 2.335 to 1.802. Setting $A_{P,mean} = \; A_{P,min}$ = 0.497, the mixed method and the prediction method yield the same AP = 0.366 at the limiting quality level, see Fig. 11. Elsewhere, the mixed method rejects more good lots and accepts more bad ones than the prediction method. It Even if it were possible to maintain a constant value for σ to determine it and to verify it, the mixed method would provide no benefits.

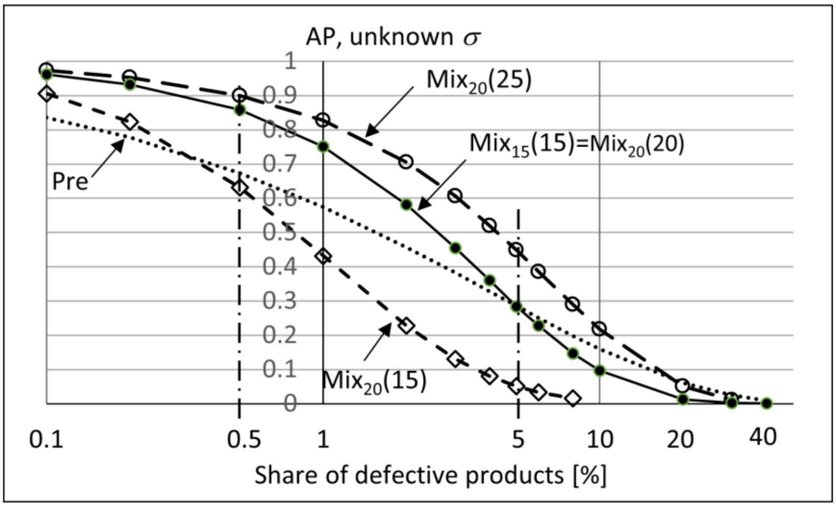

Figure 10. OC curves for 0.05-quantile, n = 5. Prediction method and three σ-dependent curves for a mixed method.

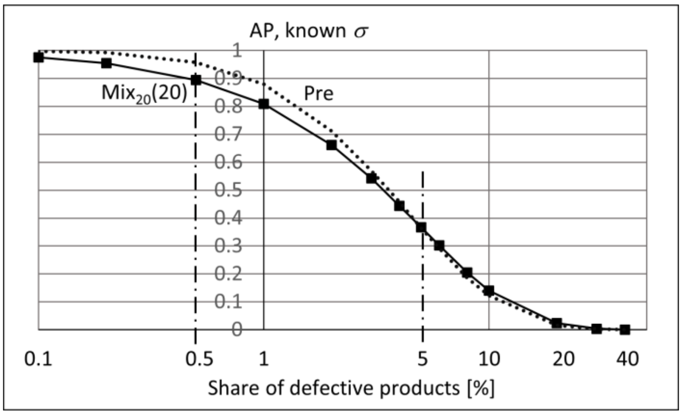

Figure 11. OC curves for 0.05 quantile, n = 5, σ is known. Prediction and mixed method.

Separate limits for $\bar{x}$ and x_{min} bring both complexity and probabilistic obscurity to the quality control. Despite this, mixed methods have been adopted e.g. in European standards EN 206–1 (2005), EN 10080 (2005) and EN 1992–1–1 (2004, Annex C). In EN 10080 the limits are a national choice, but neither default values nor any guidance for choosing them are given.

Retesting

The number of erroneously rejected lots can be reduced by increasing the sample size. A naïve trick to reduce the costs of testing would be to apply the acceptance method with sample size n_1 in Phase 1. If the lot is rejected in Phase 1, n_2 more observations are taken in Phase 2 and the acceptance method is applied with sample size $n = n_1+n_2$. This trick results in an OC curve denoted by n_1&n_2 in Fig. 12. It is above the OC curve for T_n^{pre} denoted by n.

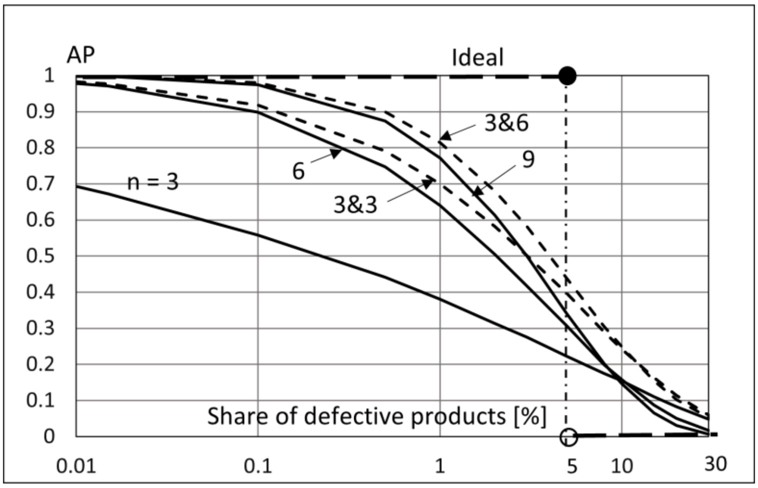

Figure 12. OC curves for five acceptance criteria. 3, 6 and 9: Prediction method for sample sizes 3, 6 and 9, respectively. 3&3 and 3&6: Prediction method for sample size 3, and if rejected, followed by prediction method for sample size 6 and 9, respectively.

Both the number of erroneous rejections and the costs are lower than when conducting n_1+n_2 tests on each lot, but this happens at the cost of considerable increase in the consumer's risk because the number of erroneously accepted lots in Phase 2 adds to that of erroneously accepted lots in Phase 1. When applying such criteria, the consumer's risk tends to be forgotten.

European strength control of rebars

In Europe, a lot of reinforcing steel bars has traditionally been accepted when the minimum of three observed strength values exceeds the limit strength L for the 0.05 quantile. As pointed out above, the minimum of three outcomes is a definition-based Weibull estimator for the 0.25-quantile. This criterion can be reformulated as follows:

> *When the estimated 0.25-quantile is greater than or equal to the requirement for the 0.05-quantile, the lot is accepted, otherwise rejected.*

In several specifications the criterion is even milder. In the method of EN 1990-1-1 (2004) (EC2) the lot is accepted if either none of three strength observations is below L, or if the minimum is ≥ kL and the mean is ≥ $L+a$. The recommended values are k = 0.97 and a = 10.

The EC2 method is a combination of $T_3^{wei}(0.25)$ and additional rules which make the OC curve dependent on σ. Two EC2 curves for σ = 10 and 20 as well as two prediction-based curves are shown in Fig. 13. The EC2(3,10) curve would be close to the $T_4^{pre}(0.25)$ curve but the EC2(3,20) curve far above it. Replacing the presently used σ-dependent acceptance methods with an objective method based on a DBQE is not possible without affecting the acceptance probability.

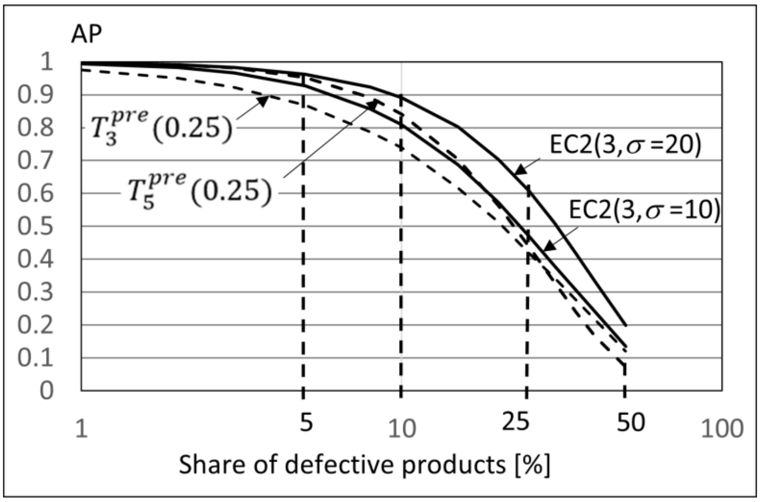

Figure 13. OC curves. EC2 method for standard deviations σ = 10 and 20. Prediction method for p = 0.25, n = 3 and 5.

Let us consider four cases:

(a) *The criteria are understood as controlling the 0.25-quantile.*
From Fig. 13 it is obvious that when $\sigma = 10$, the curves $T_4^{pre}(0.25)$ and EC2(3,10) are approximately equivalent. On the other hand, there is no such n that the curve EC2(3,20) could be approximated by curve $T_n^{pre}(0.25)$.

(b) *The criteria are understood as controlling the 0.05-quantile.*
All criteria are non-conservative. However, the rejection of 4–14 % of the lots is uneconomical, and the producer needs to set the target strength higher in such a way that the average share of defective products is lower than 5%. The demand for such a *shifting* effect decreases with the effectiveness of the acceptance method. The greater n, the lower the target quality.

(c) *The criteria are tuned to result in the present safety level*.
As far as EC2(3, σ) is considered to represent the present safety level, it depends on σ. $T_4^{pre}(0.25) \geq L$, where L is the lower limit for 0.05-quantile, might be a reasonable σ-independent choice to replace the criterion of Eurocode 2. Unlike EC2(3, σ), it gives easily understandable statistical background to the calibration of the material safety factors which also depend on things not related to the formal statistics. Since the sampling error decreases with increasing n, the need for the shifting mentioned above also decreases, and choosing n > 4 might mean a lower average outgoing strength level.

(d) *Acceptance criterion $T_n^{pre}(0.05) \geq L$ is applied.*
For small sample sizes used in destructive testing, this statistically sound alternative would be so much stricter and more expensive than the present practice that it will hardly be adopted.

Discussion

Structural design needs knowledge of the strength distribution of materials. In Eurocodes it is assumed that the actual outgoing strength level in the factories conforms to the p-quantile of the strength used when calibrating the material safety factors. We have shown that in Europe, the acceptance rules for rebars, applied to control the 0.05-quantile, actually control 0.25- or higher quantiles. On the other hand, the sample size of the order of three is so small that the risk of erroneous rejection is high. To minimize the amount of rejected products, the target strength level in production, and obviously also the average outgoing strength level, are higher than the lowest acceptable level. In addition, some measures like monitoring the long-term strength of the reinforcing steel (average strength level over some months), even though it cannot reject anything, works in the same direction. Whether these reasons are enough to compensate for the statistical mismatch in the acceptance rules, is unknown.

So far, the present safety level has been regarded as satisfactory. Adopting stricter rules like the prediction method for the 0.05-quantile is not realistic as far as the involved parties are satisfied with the status quo. However, we strongly recommend using the probabilistically founded, simple and transparent prediction method, and abandoning the coverage method as well as all variants of the mixed and attribute methods. The solution is simple. Choose p' > p. Accept the lot if $\hat{x}_{p'} \geq L$, otherwise reject it, and specify the reasons, why it is justified to claim that the acceptance criterion $\hat{x}_{p'} \geq L$ implies either that the outgoing lots meet the requirement $x_p \geq L$ or that such a requirement is not necessary.

Conclusions

In the quality control of structural parameters like the strength, it is preferable to use an acceptance method in which the p-quantile is first estimated, and the considered lot is accepted or rejected depending on whether the estimate is higher or lower than the limit value L. The definition-based prediction estimator is preferable for normally and lognormally distributed variables. The definition of a p-quantile is enough to justify it, and no Bayesian approach is needed.

The classical coverage estimator should be abandoned, because the p-quantile which it estimates varies with the sample size and the arbitrarily chosen confidence level. For small sample sizes, the confidence level 0.75 gives a reasonable accordance with the prediction estimator. However, there is no reason to use an approximation when a better alternative is available and is equally easy to use. Choosing a confidence level as high as 0.90, which is often recommended for better safety, results in strong underestimation of the p-quantile, the effect of which on the structural safety is difficult to quantify. When extra safety is needed, it is preferable to increase the sample size or to modify the safety factors.

Mixed acceptance methods, comprising one lower limit for the sample mean and another one for the lowest test result, should not be used. Since the spread is characterized by the lowest value only, information is lost in comparison with the prediction method.

The lower limits depend on the standard deviation which must be known or estimated. In both cases the acceptance probability is sensitive to the standard deviation.

Acceptance sampling by attributes exploits information from the sample less effectively than acceptance sampling by variables. It should not be used when the sample size is small and the characteristic strength is defined as a low *p*-quantile of the probability distribution. The conventional method of accepting a lot of products when three test results are above the specified lower limit, actually controls the 0.25-quantile. The even milder methods presented in some European and national standards control *p*-quantiles in which *p* is > 0.25. This is far from the 0.05-quantile which is claimed to be controlled.

The high risk of rejection associated with a small sample size makes the producer set the target quality higher than the minimum acceptable level. Consequently, the average outgoing quality tends to be higher than that formally controlled. Based on this argument we propose that in the quality control of the structural materials, the prediction method for a *p*'-quantile (*p*' > *p*) be adopted to control the *p*-quantile. By adjusting *p*', a rough calibration to one of the presently used acceptance methods is possible. In this way the acceptance process will not change essentially but the *p*'-quantile becomes transparent and allows appropriate calibration of the safety factors.

Appendix A: Expectation of $F(\hat{x}_p)$

Assume that X has a continuous density function f and cumulative distribution function F. Consider Monte Carlo simulation in which an estimator's conformity to the criterion

$$P\{X \le \hat{X}_p\} = p' \tag{A1}$$

is investigated or the p'-quantile which $\hat{X}_p$ actually estimates is determined. M estimates $\hat{x}_{p,m}$ are generated and for each estimate, K random numbers $x_{m,k}$ for each $\hat{x}_{p,m}$ are drawn from X. The following MxK matrix illustrates the situation. $I\{A\}$ denotes indicator function which is = 1 if A is true, otherwise = 0.

Table A1. Monte-Carlo simulation justifying Eqs. (A2) and (A3).

$$\begin{array}{ccccccc}
 & & & & & K \to \infty & \\
 & \dfrac{I\{x_{1,1} \le \hat{x}_{p,1}\}}{KM} & \dfrac{I\{x_{1,2} \le \hat{x}_{p,1}\}}{KM} & \cdots & \dfrac{I\{x_{1,K} \le \hat{x}_{p,1}\}}{KM} & \Sigma \to & \dfrac{1}{M} P\{X \le \hat{x}_{p,1}\} \\
 & \dfrac{I\{x_{2,1} \le \hat{x}_{p,2}\}}{KM} & \dfrac{I\{x_{2,2} \le \hat{x}_{p,2}\}}{KM} & \cdots & \dfrac{I\{x_{2,K} \le \hat{x}_{p,2}\}}{KM} & \Sigma \to & \dfrac{1}{M} P\{X \le \hat{x}_{p,2}\} \\
 & \vdots & \vdots & & \vdots & & \\
 & \dfrac{I\{x_{M,1} \le \hat{x}_{p,M}\}}{KM} & \dfrac{I\{x_{M,2} \le \hat{x}_{p,M}\}}{KM} & \cdots & \dfrac{I\{x_{M,K} \le \hat{x}_{p,M}\}}{KM} & \Sigma \to & \dfrac{1}{M} P\{X \le \hat{x}_{p,M}\} \\
M \to \infty & \Sigma \downarrow & \Sigma \downarrow & & \Sigma \downarrow & & \\
 & \dfrac{P\{X \le \hat{X}_p\}}{K} & \dfrac{P\{X \le \hat{X}_p\}}{K} & & \dfrac{P\{X \le \hat{X}_p\}}{K} & &
\end{array}$$

The sum of elements on each horizontal line m approaches stochastically $P\{X \le \hat{x}_{p,m}\}/M$ = $F(\hat{x}_{p,m})/M$ when $K \to \infty$. The sum of elements on each vertical line approaches stochastically $P\{X \le \hat{X}_p\}/K$ when $M \to \infty$. The sum of all elements approaches stochastically both $\sum_{i=1}^{M} F(\hat{x}_{p,i})\,/M$ and $P\{X \le \hat{X}_p\}$ when both K and $M \to \infty$. It follows that

$$P\{X \le \hat{X}_p\} \approx p' = \tfrac{1}{M} \textstyle\sum_{i=1}^{M} F(\hat{x}_{p,i}) \tag{A2}$$

when M is very large. Eq. (A2) greatly simplifies the simulation because there is no need to generate any random $x_{m,k}$. Furthermore,

$$P\{X \le \hat{X}_p\} = E\left(F(\hat{X}_p)\right) \tag{A3}$$

Appendix B: Expressions for *k*-factors

General background data

Assume that $X \sim N(\mu, \sigma)$. Random variables sample mean and sample deviation are denoted by $\bar{X}$ and S, their outcomes by $\bar{x}$ and s. It is well-known that

1. If $X \sim N(\mu, \sigma)$, p-quantile of X is $x_p = \mu + k_p \sigma$ where $k_p = \phi^{-1}(p)$
2. Sample mean $\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$
3. Random variable $\frac{Q}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X - \bar{X})^2$ is χ^2-distributed with *n*–1 degrees of freedom
4. If $Y \sim N(0,1)$, $\frac{Y}{\sqrt{\frac{Q}{\sigma^2 (n-1)}}} \sim t_{n-1}$ where t_{n-1} is the Student's *t*-distribution with *n*–1 degrees of freedom
5. If $Y \sim N(0,1)$ and c is constant, $\frac{Y+c}{\sqrt{\frac{Q}{\sigma^2 (n-1)}}} \sim t_{n-1,c}$ where $t_{n-1,c}$ is the noncentral *t*-distribution with *n*–1 degrees of freedom and noncentrality parameter *c*.

Prediction estimator, $\boldsymbol{k_n^{pre}(p)}$

Find such a $k_n^{pre}(p)$ that $P\{X \leq \hat{X}_p\} = p$ where $\hat{x}_p = \bar{x} + k_n^{pre}(p) s$.

$$X - \bar{X} \sim N\left(0, \sqrt{\sigma^2 + \frac{\sigma^2}{n}}\right) = N\left(0, \sigma\sqrt{1 + 1/n}\right) \tag{B1}$$

Since

$$\frac{S}{\sqrt{n}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{Q}{n(n-1)}} \tag{B2}$$

where *s* is the sample standard deviation, we can write

$$U = \frac{X - \bar{X}}{S\sqrt{1+1/n}} = \frac{\frac{X - \bar{X}}{\sigma\sqrt{1+1/n}}}{\sqrt{\frac{Q}{\sigma^2 (n-1)}}} \tag{B3}$$

The numerator encompasses a normally distributed variable $X - \bar{X}$ divided by its standard deviation, i.e. the numerator is $\sim N(0,1)$. The denominator is χ^2-distributed with *n*–1 degrees of freedom. It follows that $U \sim t_{n-1}$ and

$$t_{n-1}\left(\frac{X - \bar{X}}{s\sqrt{1+1/n}}\right) = p \ \Leftrightarrow \frac{X - \bar{X}}{s\sqrt{1+1/n}} = t_{n-1}^{-1}(p)$$
(B4)

Setting

$$k_n^{pre}(p) = \sqrt{1 + 1/n}\, t_{n-1}^{-1}(p) \text{ and } \hat{x}_p = \bar{x} + k_n^{pre}(p) s \tag{B5}$$

yields

$$P\{X \leq \hat{X}_p\} = p \tag{B6}$$

Hence, the prediction estimator (B5) results from purely frequentistic considerations, and no Bayesian approach is needed.

Coverage estimator, $\boldsymbol{k_n^{cov}(p,\alpha)}$

Find such a $k_n^{con}(p,\alpha)$ that $P\{\hat{X}_p \leq x_p\} = \alpha$ where $\hat{x}_p = \bar{x} + k_n^{con}(p,\alpha)s$, α is the confidence level and $x_p = \mu + \phi^{-1}(p)\sigma$ is the true *p*-quantile of *X*. From this criterion

$$\hat{x}_p = \bar{x} + k_n^{cov}(p,\alpha)s \leq x_p \tag{B7}$$

$$\Leftrightarrow \frac{\bar{x}-x_p}{s} \leq -k_n^{cov}(p,\alpha) \tag{B8}$$

$$\Leftrightarrow \frac{\bar{x}-x_p}{s}\sqrt{n} \leq -k_n^{cov}(p,\alpha)\sqrt{n} \tag{B9}$$

Observing that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{\mu-x_p}{\sigma/\sqrt{n}}$ is constant and writing

$$V = \frac{\bar{X}-x_p}{s}\sqrt{n} = \frac{\frac{\bar{X}-x_p}{\sigma/\sqrt{n}}}{\sqrt{\frac{Q}{\sigma^2(n-1)}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}+\frac{\mu-x_p}{\sigma/\sqrt{n}}}{\sqrt{\frac{Q}{\sigma^2(n-1)}}} \tag{B10}$$

we see that $\frac{\bar{X}-x_p}{S}\sqrt{n} \sim t_{n-1,c}$ where $c = \frac{\mu-x_p}{\sigma/\sqrt{n}} = -\phi^{-1}(p)\sqrt{n}$. So

$$P\{\hat{X}_p \leq x_p\} = P\left\{\frac{\bar{X}-x_p}{s}\sqrt{n} \leq -k_n^{cov}(p,\alpha)\sqrt{n}\right\} = t_{n-1,c}\left(-k_n^{cov}(p,\alpha)\sqrt{n}\right) = \alpha \tag{B11}$$

$$k_n^{cov}(p,\alpha) = \frac{-t_{n-1,c}^{-1}(\alpha)}{\sqrt{n}} \tag{B12}$$

Appendix C: Determining limits $\bar{x}_{lim}$ and $x_{lim,min}$

Assume that the acceptance probabilities $A_{P,mean} = P\{\bar{X} \geq \bar{x}_{lim}\}$ and $A_{P,min} = P\{X_{min} \geq x_{lim,min}\}$ are given, and when they are applied, the resulting overall acceptance probability is = AP. When $X \sim N(\mu, \sigma)$ and L is the limit for the p-quantile of X

$$L = x_p = \mu + k_p \sigma \tag{C1}$$

Note that $k_p < 0$ for p < 0.5. At the lower limit

$$\mu = L - k_p \sigma \tag{C2}$$

This is achieved when $A_{P,mean} A_{P,min} = A'_P$ where $A'_P < A_P$ is properly chosen. Consider first $x_{lim,min} = x_0$. For each x_0, there is a probability $p_0 = F(x_0)$ which satisfies

$$P\{X \leq x_0\} = p_0 = 1 - P\{X > x_0\} \quad \text{or} \quad P\{X > x_0\} = 1 - p_0 \tag{C3}$$

$$P\{X_{min} > x_0\} = P\{all(X_i) > x_0\} = (1 - p_0)^n \tag{C4}$$

where $x_1, \ldots, x_n$ is a sample taken from X. To determine p_0 and x_0 we require that they are in accordance with the given acceptance probability $A_{P,min}$

$$A_{P,min} = P\{X_{min} > x_0\} = (1 - p_0)^n \Rightarrow p_0 = 1 - A_{P,min}{}^{\frac{1}{n}} \tag{C5}$$

From Eq. (C5) we get

$$F(x_0) = \phi\left(\frac{x_0 - \mu}{\sigma}\right) = p_0 = 1 - A_{P,min}{}^{\frac{1}{n}} \tag{C6}$$

$$\frac{x_0 - \mu}{\sigma} = \phi^{-1}\left(1 - A_{P,min}{}^{\frac{1}{n}}\right) \tag{C7}$$

$$x_0 = x_{lim,min} = \mu + \sigma \phi^{-1}\left(1 - A_{P,min}{}^{\frac{1}{n}}\right) = \mu + z_1 \sigma = L + \left(-k_p + z_1\right)\sigma \tag{C8}$$

The sample mean $\bar{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. So

$$A_{P,mean} = P\{\bar{X} \geq \bar{x}_{lim}\} \tag{C9}$$

$$1 - A_{P,mean} = P\{\bar{X} < \bar{x}_{lim}\} = F\{\bar{x}_{lim}\} = \phi\left(\frac{\bar{x}_{lim} - \mu}{\sigma/\sqrt{n}}\right) \tag{C10}$$

$$\bar{x}_{lim} = \mu + \frac{\sigma}{\sqrt{n}} \phi^{-1}\left(1 - A_{P,mean}\right) = \mu + z_2 \sigma = L + \left(-k_p + z_2\right)\sigma \tag{C11}$$

To summarize:

$$x_{lim,min} = L + \left(-k_p + z_1\right)\sigma \qquad z_1 = \phi^{-1}\left(1 - A_{P,min}{}^{\frac{1}{n}}\right) \; (< 0) \tag{C12}$$

$$\bar{x}_{lim} = L + \left(-k_p + z_2\right)\sigma \qquad z_2 = \frac{1}{\sqrt{n}} \phi^{-1}\left(1 - A_{P,mean}\right) \; (< 0) \tag{C13}$$

References

Ang, A.H., W.H. Tang. 2007. *Probability concepts in engineering. Emphasis on applications to civil engineering*. New York, Wiley. ISBN-10 0-471-72064-X.

ASTM Standard A615/A615M–20. 2020. *Specification for Deformed and Plain Carbon-Steel Bars for Concrete Reinforcement.* https://doi.org:10.1520/A0615_A0615M-20

Caspeele, R., L. Taerwe. 2012. Quantitative Comparison of Estimation Methods for Determining the in Situ Characteristic Concrete Compressive Strength. *Structural Engineering International*, 22: 215–22. https://doi.org/10.2749/101686612X13291382990840

Cunnane, C. 1978. Unbiased plotting positions–a review. *Journal of Hydrology*, 37:205–222.

EN 206–1. 2005. Concrete–Part 1: *Specification, performance, production and conformity*.

EN 1990 + A1. 2005. *Eurocode–Basis of structural design*.

EN 1992–1–1. 2004. *Eurocode 2: Design of concrete structures. Part 1–1: General rules and rules for buildings*.

EN 10080. 2005. *Steel for the reinforcement of concrete – Weldable reinforcing steel – General*.

Fuglem, M., G. Parr, I. J. Jordaan. 2013. Plotting positions for fitting distributions and extreme value analysis. *Canadian Journal of Civil Engineering*, 40: 130–139. http://dx.doi.org/10.1139/cjce-2012-0427

Gringorten, H. 1963. A plotting rule for extreme probability paper. *Journal of Geophysical Research*, 68: 813–814.

Holický, M. 2013. *Introduction to Probability and Statistics for Engineers*. Heidelberg, Springer. ISBN 978-3-642-38299-4.

ISO 12491.1997. *Statistical methods for quality control of building materials and components*.

Madsen, H. O., S. Krenk, N. C. Lind. 1986. *Methods of structural safety*. Englewood Cliffs, New Jersey, Prentice-Hall. ISBN 0-13-579475-7.

Makkonen, L. 2008. Problems in the extreme value analysis. *Structural Safety*, 30: 405–419. https://doi:10.1016/j.strusafe.2006.12.001

Makkonen, L., M. Pajari. 2014. Defining sample quantiles by the true rank probability. *Journal of Probability and Statistics*, 326579. http://dx.doi.org/10.1155/2014/326579

Makkonen, L., M. Tikanmäki. 2019. An improved method of extreme value analysis. *Journal of Hydrology* X, 2: 100012.
https://doi.org/10.1016/j.hydroa.2018.100012

Pajari, M., M. Tikanmäki, L. Makkonen. 2019. Probabilistic evaluation of quantile estimators. *Communications in Statistics – Theory and Methods*, 50: 3319–3337.
https://doi.org/10.1080/03610926.2019.1696975

Sagemath (2017). *The Sage Mathematics Software System* (Version 8.1). The Sage Developers.
https://www.sagemath.org

Schilling, E. G., D. V. Neubauer. 2017. *Acceptance Sampling in Quality Control*. 3rd edition. Chapman & Hall, ISBN 9781498733571.

Tur, V., S. Derechennik. 2020. Non-parametric evaluation of the characteristic in-situ concrete compressive strength. *Journals of Building Engineering*, 27: 100938.
https://doi.org/10.1016/j.jobe.2019.100938

Wilks, S. S. 1941. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12: 91–96.
http://www.jstor.org/stable/2235627

Zupan, D., J. Srpčič, G. Turk. 2006. Characteristic value determination from small samples. *Structural Safety*, 29: 268–278.
https://doi:10.1016/j.strusafe.2006.07.006

Matti Pajari
Berakon
Peräsin 2 B, FI-02320 Espoo, Finland
matti.pajari@berakon.fi

Lasse Makkonen
VTT Technical Research Centre of Finland
Yläkalliontie 27, FI-02760 Espoo, Finland
lasse.makkonen3@gmail.com