World Conference on Timber Engineering Oslo 2023

NON-PARAMETRIC LOWER CONFIDENCE BOUNDS FOR THE FIFTH PERCENTILE – EN 14358 IN COMPARISON TO A FULLY NON-PARAMETRIC APPROACH

Andreas Weidenhiller[1], Dan Ridley-Ellis[2], Andreas Neumüller[3]

ABSTRACT: In Europe, fifth percentile values are required for the calculation of characteristic values of strength and density. The European standard EN 14358:2016 defined three ways to calculate a 75% lower confidence bound (LCB) for such fifth percentile values, based either on a lognormal parametric approach, on a normal parametric approach or on a non-parametric approach. Using simulated data with different sample sizes and with different underlying distributions, this paper studied the effects of using each of the three approaches of EN 14358. As the third approach in EN 14358 did not seem to be fully non-parametric, the simulation study included, as a fourth approach, a fully non-parametric calculation of the LCB for the fifth percentile. The simulation study confirmed that both non-parametric approaches led to acceptable results for some important distributions, although the non-parametric approach defined in EN 14358 seemed to be more conservative especially for data with a non-normal distribution. The study also confirmed that the use of an incorrect parametric assumption can lead to systematically misleading LCB values for the fifth percentile. The authors recommend replacing the non-parametric approach currently defined in EN 14358 by a fully non-parametric approach. This approach can easily be implemented in a standard.

KEYWORDS: timber strength; fifth percentile; lower confidence bound; simulation study; distribution assumptions

1 INTRODUCTION

Fifth percentile values are frequently required when assessing strength. Such values are also used for assessing the strength of timber, timber products and timber structures. In recent years, the European standard EN 14358 [1] has evolved towards encompassing a number of cases for calculating fifth percentile values.

In its original form from 2006, the standard EN 14358 [2] contained instructions for calculating 75% lower confidence bounds (LCB) for the fifth percentile assuming a lognormal distribution; the purpose was to assess test results for connections and wood-based products. Construction timber was explicitly excluded and was dealt with in a separate standard (EN 384 [3]). The calculation in EN 384 differed from the one in EN 14358 in two important aspects: it was a nonparametric fifth percentile (in contrast to the parametric lognormal distribution assumption in EN 14358), and the fifth percentile itself was reported as the result (in contrast, EN 14358 required to report the 75% LCB). With the revision of both standards in the year 2016 [1, 4], the calculation of the fifth percentiles was completely moved from EN 384 to EN 14358. In the course of this revision, EN 14358 was extended to encompass, in addition to lognormal fifth percentiles, also fifth percentiles for normally distributed quantiles, nonparametric fifth percentile calculation, and mean values of normally distributed quantities, all with a 75% LCB [1].

For the parametric quantities, the standard gives three options for calculation: using an exact formula, using a rational approximation, and using tabled coefficients. For the 75% LCB m_k of the nonparametric fifth percentile $y_{0,5}$, only an approximation formula is given [1]:

$$m_k = y_{0,5}\left(1 - \frac{k_{0,5,0,75}V}{\sqrt{n}}\right) \quad (1)$$

where n is the sample size, V is the coefficient of variation (CoV), and $k_{0,5,0,75}$ is a factor of which the standard claims that it leads to the 75% LCB [1]:

$$k_{0,5,0,75} = \frac{0.49n + 17}{0.28n + 7.1} \quad (2)$$

The authors could not trace the origins of these formulas and the underlying assumptions remained unclear. But the use of the coefficient of variation V in formula (1) makes it clear that some parametric assumptions have to be involved.

On the other hand, there is a well-known method to calculate a fully nonparametric LCB (see, for example [5]). This method is discussed in more detail in section 2.1 below.

[1] Andreas Weidenhiller, Holzforschung Austria, Austria, a.weidenhiller@holzforschung.at
[2] Dan Ridley-Ellis, Edinburgh Napier University, UK, dan.ridley-ellis@napier.ac.uk
[3] Andreas Neumüller, Holzforschung Austria, Austria, a.neumueller@holzforschung.at

https://doi.org/10.52202/069179-0101

How does this method perform in comparison to the approaches defined in EN 14358 [1]?
In the long run (after, say, 1000 repetitions), we would expect that, in 75% of the cases, the true fifth percentile value is above the 75% LCB. This fact was used in the present paper to compare the 75% LCB calculation approaches from EN 14358 [1] and from the fully nonparametric approach with the true fifth percentile values for simulated data with a known distribution.

2 MATERIAL AND METHODS

2.1 FULLY NONPARAMETRIC 75% LOWER CONFIDENCE BOUND FOR THE FIFTH PERCENTILE

To get a non-parametric LCB for the fifth percentile of a sample with n measurement values, we order the measurement values by increasing value, so that x_1 is the smallest value, x_2 the second smallest value, and x_n the largest value. We say that these values have rank 1, 2 and n, respectively.
The probability with which the measurement value x_i with rank i underestimates the fifth percentile can be calculated using the binomial distribution [6].
This probability is independent of the actual values x_i – it only depends on the ranks i. Therefore, one can calculate a table which tells, for each sample size n, the maximum rank i_{max} which fulfils the following condition: The probability that $x_{i_{max}}$ underestimates the fifth percentile is 75% or more. Underestimating the fifth percentile means that we get a conservative estimate of the fifth percentile.
Table 1 lists such values i_{max} (ranks) for selected values of n. Such a table could easily be incorporated in a standard. For intermediate values of n, one would have to pick the smaller rank value (e.g., for $n = 50$, still the value x_1 with rank 1 would have to be picked as the 75% LCB).

Table 1: *Lowest sample size n for which the element with a given rank (counted from the smallest to the largest value) is a nonparametric 75% lower confidence bound (LCB) for the fifth percentile.*

rank	n	rank	n	rank	n
1	28	8	193	15	347
2	53	9	215	16	368
3	78	10	237	17	390
4	102	11	259	18	412
5	125	12	281	19	433
6	148	13	303	20	455
7	170	14	325	21	476

Due to the discrete nature of the binomial distribution, for intermediate values of n, the probability that $x_{i_{max}}$ underestimates the fifth percentile can be quite a bit higher than the 75% which we aimed for – for example, the probability for $n = 40$ is 87%. However, it is also possible to determine interpolated ranks directly at the desired probability of 75%, for example, by using the R package cbinom which provides a continuous analog of the binomial distribution [7].
In the present study, we used the sample sizes $n = 40$, 80, 500, 1000 and 100000 (see 2.3) – for these, the interpolated ranks are 1.488, 3.107, 22.13, 45.77 and 4954, respectively. These ranks were used for the calculation of the fully non-parametric LCB values in the study.

2.2 INCLUDED PROBABILITY DISTRIBUTIONS

We included normal and lognormal distributions, because they are included in EN 14358. For normal distributions, the location parameter does not influence the results, so we chose 420 (required mean density in kg/m³ for C24 in EN 338 [8]) and used CoV values of 5%, 10% and 20%. For the lognormal distribution, the location parameter also changes the shape of the distribution, so we calculated mean strength values corresponding to the requirements for C16, C24 and C35 in EN 338 [8] together with CoV values of 5%, 10% and 20%. A further important class of distributions are truncated normal distributions, because these represent an extreme case of what efficient strength grading can do. Here we took the normal distributions defined above but cut off everything below 310, 350, respectively 390 (required fifth percentile of density in kg/m³ for C16, C24 respectively C35). The calculations on the truncated normal distributions were handled using the R package truncnorm [9].
Regarding normal distributions, it is important to note that those always include negative values (if maybe only with a small probability). Negative values are not realistic in the application and cause the calculation of parametric lognormal LCB values to fail. Therefore, the normal distributions are actually modelled as truncated normal distributions truncated below 1.
The parameterization of the different distributions is listed in Table 2.

2.3 SIMULATION STUDY

We simulated 1000 samples of each of the distributions given in Table 2, for each of the following sample sizes: 40, 80, 500, 1000 and 100000. 40 was chosen as the smallest sample size, because below that, the nonparametric calculation according to EN 14358 may not be used [1]. 100000 was chosen as an "almost infinite" sample size, where we expected the LCB to be almost equal to the sample fifth percentile, and we expected the sample fifth percentile to almost perfectly match the true fifth percentile according to the distribution.

2.4 STATISTICAL ANALYSIS

For each simulated sample, we calculated parametric 75% LCB values for the fifth percentile according to EN 14358 [1] assuming a normal or lognormal distribution. Further, nonparametric LCB values according to EN 14358 [1] and according to the procedure outlined in section 2.1 were calculated.

https://doi.org/10.52202/069179-0101

Table 2: *Parameters for the simulated distributions. mu and sigma describe location and spread. "lower bound" is a parameter for the truncated normal distribution that ensures that no values below the lower bound are included in the distribution. For simulation, this is implemented using accept-reject sampling (see help on rtruncnorm in [6]). Additionally, the fifth percentile of the distribution is given.*

distribution	mu	sigma	lower bound	fifth percentile
normal	420	21	1	385
	420	42	1	351
	420	84	1	282
lognormal	2,85	0,05		15,9
	2,94	0,0998		16,1
	3,1	0,198		16,0
	3,26	0,05		24,0
	3,34	0,0998		23,9
	3,5	0,198		23,9
	3,64	0,05		35,1
	3,72	0,0998		35,0
	3,88	0,198		35,0
truncated normal	420	21	310	385
	420	42	310	353
	420	84	310	329
	420	21	350	386
	420	42	350	365
	420	84	350	361
	420	21	390	396
	420	42	390	395
	420	84	390	397

For each simulated sample and each calculated LCB value, we determined the "true percentile" this LCB value corresponded to according to the distribution by which the sample was generated.
For example, the "true percentile" of 350 in a normal distribution with mean 420 and a CoV of 10% is 4.78. The "true percentile" value 4.78 in this example is lower than five. This means that 350 is a conservative estimate of the fifth percentile of the underlying distribution – in fact, the actual fifth percentile of the underlying distribution is 351.
All calculations were performed using R 4.2.1 [10].

3 RESULTS AND DISCUSSION

In Figure 1, the results of the simulation study were summarised using boxplots. On the vertical axis, for each 75% LCB value, the "true percentile" value in the respective underlying distribution was plotted. The upper end of each box corresponded to the threshold below which we found 75% of the observed values. For the 75% LCB, we would expect 75% of the observed values to be below the target value five, which is indicated by the black horizontal line.
For an optimal 75% LCB, the upper end of the box would be at the target value five, which means that 75% of the calculated values would be conservative estimates of the fifth percentile of the underlying distribution.
If the upper end of the box is above five, the calculated LCB values are too optimistic. If the upper end of the box is far below five, the calculated LCB values are very conservative. If the height of the box is large, the LCB values have a large spread. The ideal situation would be a box with low height with its upper end at or closely below five.
In Figure 1a, the true percentiles for the parametric normal calculation of 75% LCB values according to EN 14358 are shown. When the underlying distribution was normal (leftmost five boxes in Figure 1a), the height of the boxes as well as the distance of the true percentiles to the target value five decreased with increasing sample size n. For $n = 100000$, there were almost no deviations of the true percentile from the target value five. However, for the other underlying distributions in Figure 1a, the heights of the boxes and the distances from the target value five did not decrease. In particular, even for $n = 100000$, the true percentile values could be far away from the target value five. Even if the discrepancies are on the conservative side, it is clear that the parametric normal LCB calculation was misleading when the underlying distribution was not normal.
A similar pattern could be observed for the parametric lognormal calculation of 75% LCB values (Figure 1b). When the underlying distribution was lognormal (central five boxes in Figure 1b), the parametric lognormal calculation was close to optimal. For other underlying distributions, the parametric lognormal LCB calculation was misleading. When the underlying distribution was normal, the upper ends of the boxes were above the target value five, meaning that more than 25% of the LCB values were too optimistic, and this share increased with increasing n. For $n = 1000$, 75% of the LCB values were

https://doi.org/10.52202/069179-0101

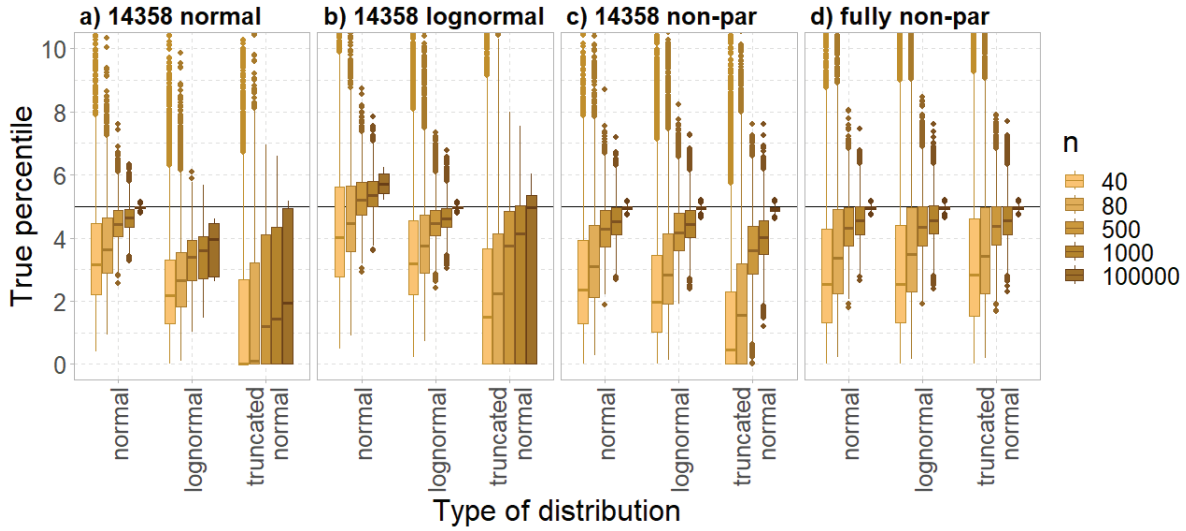

Figure 1: *Boxplots of "true percentiles" (the percentile which a calculated 75% LCB corresponds to in the true underlying distribution) separated by underlying distribution (horizontal axis), sample size n (colour scale) and mode of calculation for the 75% LCB of the sample fifth percentile. a) parametric normal LCB acc. to EN 14358, b) parametric lognormal LCB acc. to EN 14358, c) non-parametric LCB acc. to EN 14358, d) fully non-parametric LCB as outlined in section 2.1. The range of the vertical axis is optimised for the boxes – not all extreme values are included. Each box summarises 1000 runs times the number of different parameterisations for each type of distribution – see also Table 2.*

too optimistic, and for $n = 100000$, all parametric lognormal LCB values were above the target value 5.

Both nonparametric approaches to calculating 75% LCB values (Figure 1c and d) worked for all underlying distributions examined in this study. In all the cases, the height of the boxes as well as the distance of the true percentiles to the target value five decreased with increasing sample size n. For $n = 100000$, there were almost no deviations of the true percentile from the target value five.

In Figure 2, the boxes from Figure 1c and d were rearranged to facilitate comparisons between the two nonparametric approaches to calculating 75% LCB values. When the underlying distribution was normal, the two approaches behaved similarly, with the fully non-parametric LCB values being slightly closer to the target. For an underlying lognormal distribution, and even more for an underlying truncated normal distribution, the fully non-parametric LCB values were distinctly closer to the target fifth percentile.

In Figure 1, the dangers of using a parametric calculation of the 75% LCB can be observed. If the distribution assumptions are not met, there is a risk of obtaining misleading results.

The non-parametric approaches to calculating the 75% LCB seemed to work well for very different types of distributions, at the cost of a higher spread of the values (indicated by greater heights of the boxes in the plots).

To facilitate the inclusion of the fully non-parametric approach in a standard, it would be helpful to have an approximation function for the interpolated ranks which are needed to calculate the fully non-parametric LCB values (section 2.1). Using linear regression, such an approximation was calculated for all sample sizes from 40 to 10000.

The rank r can thus be calculated as

$$r = 0.422 + 0.05\,n - 0.147\,\sqrt{n} \qquad (3)$$

where n is the sample size. The relative errors of this approximation in the range from $n = 40$ to $n = 10000$ are in the range between -0.01% and +0.26%.

4 CONCLUSIONS

In this paper, different approaches to calculating 75% lower confidence bound (LCB) values for the fifth percentile of a sample were compared by means of a simulation study.

Both parametric and non-parametric approaches were included, focusing on approaches defined in the European standard EN 14358 [1]. As the non-parametric calculation of 75% LCB values defined in the standard EN 14358 [1] is not fully non-parametric, a fully non-parametric approach was also included in the study.

Samples of sizes 40, 80, 500, 1000 and 100000 were simulated for the following underlying distributions: normal, lognormal, and truncated normal. For all distributions, different coefficients of variation were included, and for the lognormal and truncated normal distributions, the mean respectively the lower bound parameter were also varied.

The study highlighted the risk of misleading results if a parametric approach to calculating 75% LCB values is used when the assumptions about the type of underlying distribution are not met.

https://doi.org/10.52202/069179-0101

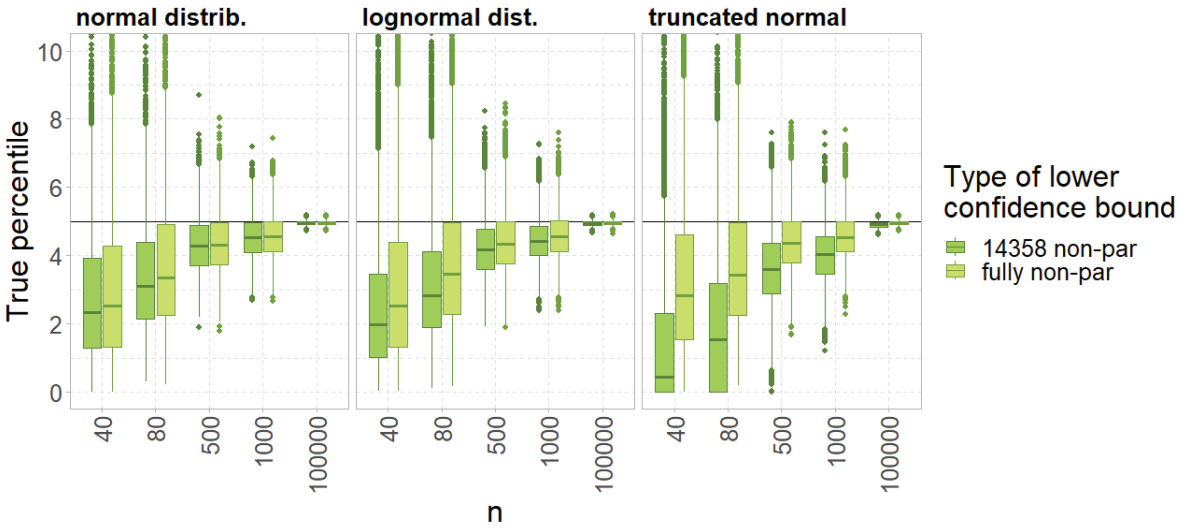

Figure 2: *Rearranged boxplots of "true percentiles" from Figure 1 for the two non-parametric approaches to calculating the 75% LCB ("14358 non-par" and "fully non-par"). For each combination of sample size n and type of distribution, the boxes for the two approaches "14358 non-par" and "fully non-par" are plotted side by side to facilitate pairwise comparison.*

Both non-parametric approaches to calculating 75% LCB values led to acceptable results. However, the non-parametric approach defined in EN 14358 [1] led to 75% LCB values which were more conservative than necessary, especially if the underlying distribution was not normal.

Therefore, the authors recommend revising the standard EN 14358 [1], replacing the current non-parametric approach to calculating 75% LCB values by a fully non-parametric approach.

The fully non-parametric approach is suitable for implementation in the standard. Implementation could be done using tabled values, an approximation formula and/or an exact formula, in a similar manner to the way in which the parametric approaches to calculating 75% LCB values are currently implemented in EN 14358 [1].

ACKNOWLEDGEMENT

Part of the research for this paper has been conducted within the InnoGrading project funded by the Austrian Research Promotion Agency FFG (grant nr. 869170).

REFERENCES

[1] CEN European Committee for Standardization. EN 14358:2016. Timber structures – calculation and verification of characteristic values. Brussels; 2016.

[2] CEN European Committee for Standardization. EN 14358:2006. Timber structures - Calculation of characteristic 5-percentile values and acceptance criteria for a sample. Brussels; 2006.

[3] CEN European Committee for Standardization. EN 384:2004. Structural timber – Determination of characteristic values of mechanical properties and density. Brussels; 2004.

[4] CEN European Committee for Standardization. EN 384:2016. Structural timber – Determination of characteristic values of mechanical properties and density. Brussels; 2016.

[5] David HA, Nagaraja HN. Order Statistics. 3rd ed. New York, NY: John Wiley & Sons 2004.

[6] Nagaraja CH, Nagaraja HN. Distribution-free Approximate Methods for Constructing Confidence Intervals for Quantiles. International Statistical Review 2020; 88(1): 75–100.

[7] Dalthorp D. cbinom: Continuous Analog of a Binomial Distribution; 2021. Available from: URL: https://CRAN.R-project.org/package=cbinom.

[8] CEN European Committee for Standardization. EN 338:2016. Structural timber – Strength classes. Brussels; 2016.

[9] Mersmann O, Trautmann H, Steuer D, Bornkamp B. truncnorm: Truncated Normal Distribution; 2018. Available from: URL: https://CRAN.R-project.org/package=truncnorm.

[10] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2022. Available from: URL: https://www.R-project.org/.

https://doi.org/10.52202/069179-0101