Univerza
v Ljubljani
Fakulteta
*za gradbeništvo
in geodezijo*

Jamova 2
1000 Ljubljana, Slovenija
http://www3.fgg.uni-lj.si/

DRUGG – Digitalni repozitorij UL FGG
http://drugg.fgg.uni-lj.si/

University
of Ljubljana
Faculty of
*Civil and Geodetic
Engineering*

Jamova 2
SI – 1000 Ljubljana, Slovenia
http://www3.fgg.uni-lj.si/en/

DRUGG – The Digital Repository
http://drugg.fgg.uni-lj.si/

Characteristic value determination from small samples

Dejan Zupan,
University of Ljubljana, Faculty of Civil and Geodetic Engineering

Jelena Srpcic,
Slovenian National Building and Civil Engineering Institute
(ZAG, Ljubljana)

Goran Turk,
University of Ljubljana, Faculty of Civil and Geodetic Engineering

Abstract

When only small samples are available, the characteristic value is usually determined with the assumption that the distribution is known. If we review the European standards different distributions are usually prescribed for the determination of the resistance of different materials and for the determination of the resistance of structures: normal, lognormal, Gumbel, and Weibull distribution. For most cases 5 % characteristic values are prescribed. All the standards use the normal or lognormal formulation regardless the distribution of the parameters used in reliability of structures. By the use of analytical formulae as well as numerical simulations the tables, similar to those in European standards, are presented for the normal, lognormal, Gumbel, and Weibull distribution. At the end the use of proposed tables is presented on the data of experimentally obtained bending strengths of finger jointed elements.

Keywords: characteristic value, small sample, log-normal, Gumbel, Weibull

1 Introduction

The structural design is based on random variables which are represented by their characteristic values. This approach is suitable for further analysis and design because

we can use the fixed values without employing any probabilistic methods. However, when only relatively small sample is available, the characteristic value is only estimated from that sample. The estimate is based on the assumption that the distribution of the variable is known and that its parameters are approximated from a sample. If we review the European standards (e.g. [3], [4]) different distributions are usually prescribed for the determination of the resistance of different materials and for the determination of the resistance of structures: normal, lognormal, Gumbel, Weibull, etc. For most cases formulae for the 75% confidence interval for the estimates of 5% characteristic values based on normal and occasionally lognormal distribution are prescribed.

In Fig. 1 it is shown that the true characteristic value is deterministic but unknown. In the case the investigated variable is normally distributed, the estimate of the characteristic value is a random variable distributed according to the non-central t distribution. Its variance depends on the sample size. Smaller samples yield to larger variance of the characteristic value which is clearly illustrated.
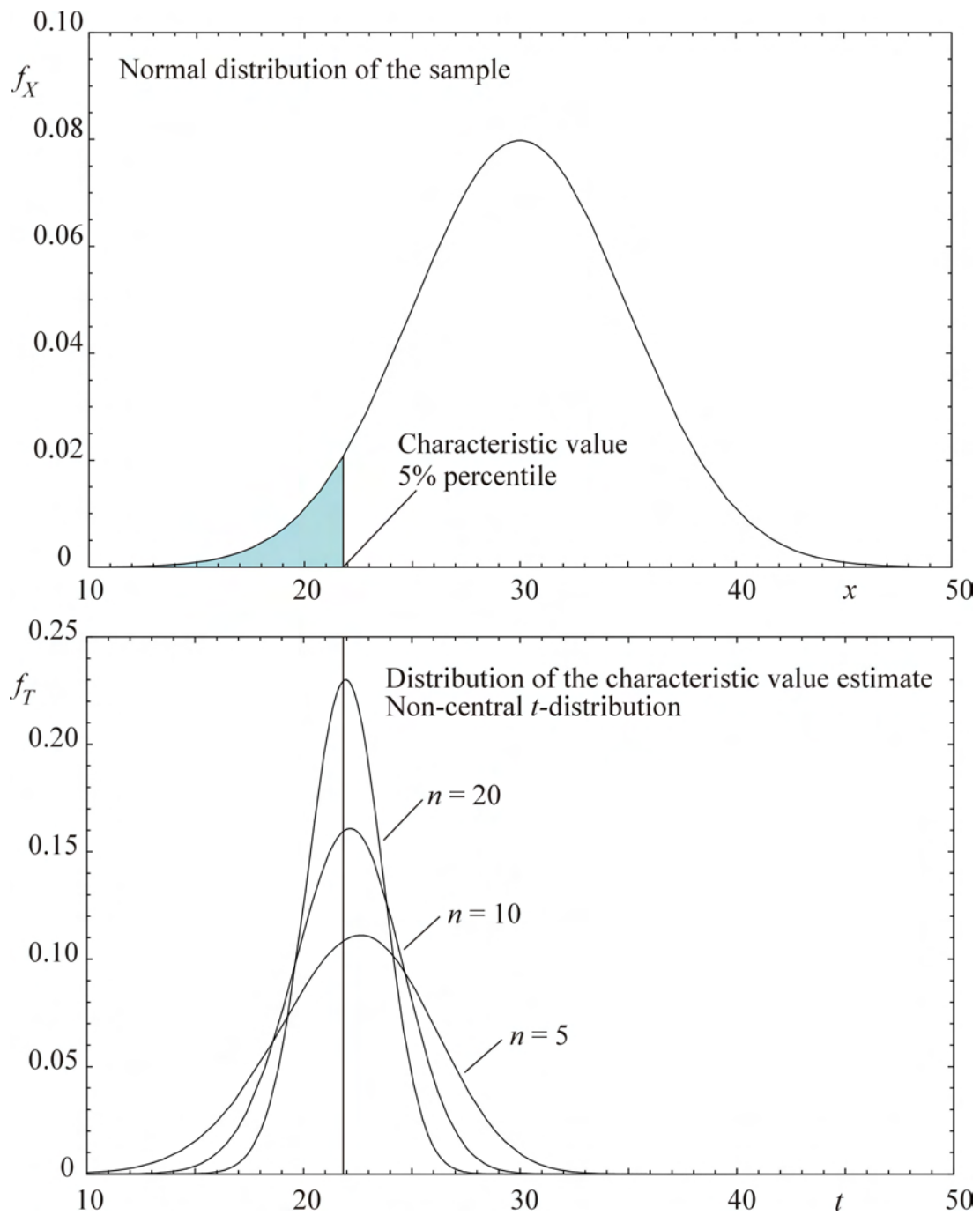
Figure 1: Probability density functions of the investigated variable and corresponding characteristic values

The aim of this paper is to give general instructions for characteristic value determination for an arbitrary distribution. The approximation based on analytical equations is developed. The results are verified by the use of simulations. For the normal and lognormal distribution we confirm the analytically developed values. Because of some additional approximations made for arbitrary distribution some

discrepancies are established for analytically obtained values; improved values are obtained by large number of simulations using bisection method.

2 Basic assumptions and definitions

Let X be a random variable with known cumulative distribution function (CDF) $F_X(x)$. The characteristic value of X is such value x_a, that the probability of X being less than x_a equals a :

$$P[X \le x_a] = F_X(x_a) = a \qquad \rightarrow \qquad x_a = F_X^{-1}(a).$$

(

It is obvious from (1) that the characteristic value x_a depends on the distribution of the random variable. The characteristic value can be uniquely determined if the CDF is known; e.g. if its parameters are prescribed. $F_X(x)$ is usually (directly or indirectly) described by the mean m_X and by the standard deviation s_X.

It is common to many practical problems that the correct values of m_X and s_X are unknown and can only be estimated from a random sample, by standard estimates: sample average $\overline{X}$ and sample standard deviation S_X^*. Thus, instead of the correct characteristic value, only its estimate could be obtained. The characteristic value estimate is itself a random variable, here denoted as $\hat{X}_a$. There are a number of possibilities for the determination of the characteristic value estimate. In the present paper we are interested in the estimates for which we can control the probability $P[\hat{X}_a \le x_a]$. Using the present approach for any prescribed confidence interval a_1 such characteristic value estimate, $\hat{X}_{a,1}$, can be determined that

$$P\left[\hat{X}_{a,l} \le x_a\right] = 1 - a_l .$$

(

The characteristic values and, consequently, their estimates are strongly dependent on distribution. Therefore, we need to discuss different distributions separately. Our approach is demonstrated in detail for the normal distribution and preformed also for the lognormal, Gumbel, and Weibull distribution.

2.1 Normally distributed variables

The basic idea on the *characteristic value determination for normally distributed variables* stems from the relationship between an arbitrary normal variable X and standardized normal random variable U

$$P[X \le x_a] = P\left[\frac{X - m_X}{s_X} \le \frac{x_a - m_X}{s_X}\right] = F_U\left(\frac{x_a - m_X}{s_X}\right) = a ,$$

(

where F_U is the CDF of standardized normal distribution, hence independent of parameters m_X and s_X. Therefore, the characteristic value can be expressed as:

$$x_a = m_X + s_X F_U^{-1}(a).$$

(

The simple and understandable form of expression (4) represents the basic estimate of the characteristic value of normally distributed random variable. If we replace the unknown parameters m_X and s_X by their sample estimates

$$\hat{X}_{a,l} = \overline{X} + S_X^* \, \boldsymbol{l} \, .$$

$($

Note that we replaced $F_U^{-1}(\boldsymbol{a})$ with parameter $\boldsymbol{l}$ which is the only free parameter in (5) and needs to be determined with respect to the previously prescribed confidence interval $\boldsymbol{a}_l$. Details concerning the determination of $\boldsymbol{l}$ can be found in [6] and [7].

As an example the results for $\boldsymbol{a} = 0.05$ and $\boldsymbol{a}_l = 0.25$ which corresponds to the determination of strength parameters for different sample sizes are shown in Tab. 1.

Table 1: Factor $\boldsymbol{l}$ for normal distribution for $\boldsymbol{a} = 0.05$ and $\boldsymbol{a}_l = 0.25$

n	3	4	5	7	10	20	50	100
$\boldsymbol{l}$	−3.125	−2.681	−2.463	−2.250	−2.104	−1.932	−1.811	−1.758

2.2 Lognormally distributed variables

As it has already been explained, the estimates of the characteristic values are dependent on distribution. Thus we must handle each particular distribution separately. In this section we develop the procedure for
characteristic value determination for lognormally distributed variable. The basic idea of the present approach is to employ the results of the normal distribution by using its relationship to the lognormal distribution.

Lognormal random variable Y is related to normal variable X through the exponential map:

$$Y = e^X \qquad \rightarrow \qquad X = \ln Y .$$

(

It can easily be shown that the value of l used for normal distribution can be utilized for lognormal distribution as well. However, the characteristic value is estimated by the following equation

$$\hat{Y}_{a,l} = \frac{\overline{Y}^2}{\sqrt{S_Y^{*2} + \overline{Y}^2}} \, e^{?\sqrt{\ln\left(S_Y^{*2}/\overline{Y}^2 + 1\right)}} ,$$

(

where $\overline{Y}$ and S_Y^* are the mean value and the standard deviation of the sample and l is the same parameter as in the previous section, i.e. the parameter from Tab. 1.

2.3 Variables of other known distributions

The idea for the lognormally distributed variable may in theory be extended to an arbitrary variable with known distribution. However, since the transformation from an arbitrary distribution to standardized normal always involves unknown parameters of the distribution, the results are not very accurate (see [6] for details). Therefore, it is impossible to use the same values l and modified formula for characteristic value determination.

Alternatively, we use the definition of characteristic value estimate (5) to determine the value of l for different distributions, different sample sizes and for some distributions different coefficients of variation via numerical simulations.

3 Simulations of the characteristic value determination

A huge number of repetitions of the sample selections can easily be simulated by computer using a random number generator. In computer simulations we can prescribe the values of mean and standard deviations in contrast to practical sampling where these parameters are usually unknown.

The simulations were performed by the following algorithm:

Reading the input values of m_X and s_X.

Calculation of the exact characteristic value x_a.

Determination of initial values for l.

Start of bisection iterations.

    Loop over simulations.

        Loop over elements of the sample.

            Random variate generation according to the chosen distribution (see [2] for details).

        End loop.

        Calculation of sample statistics $\overline{X}$ and S_X^* to estimate of m_X and s_X.

        Calculation of the estimate $\hat{X}_{a,l}$ from equation (3).

    End loop.

Estimation of probability $P\left[\hat{X}_{a,l} \leq x_a\right]$.

Update the value of l.

Continue bisection iterations until $\left|P\left[\hat{X}_{a,l} \leq x_a\right] - (1 - a_l)\right| \geq d$.

The estimation of probability $P\left[\hat{X}_{a,l} \leq x_a\right]$ is obtained by counting the number of estimates $\hat{X}_{a,l}$ that are less than x_a and by dividing this number by the number of simulations. In this procedure one million (1000000) simulations were employed. The maximum error d allowed in bisection procedure was set to 0.001.

The results of these simulations for Gumbel distribution are summarized in Tab. 2. In the case of Gumbel distributions the parameter l_G is independent of coefficient of variation which is an advantage compared to Weibull distribution.

Table 2: Factor l_G for Gumbel distribution for $a = 0.05$ and $a_1 = 0.25$

n	3	4	5	7	10	20	50	100
l_G	−4.429	−3.712	−3.351	−2.998	−2.730	−2.407	−2.174	−2.072

Similarly, the values for Weibull distribution are summarized in Tab. 3. In this case the values depend on the coefficient of variation V_X.

Table 3: Factor l_W for Weibull distribution for $a = 0.05$ and $a_1 = 0.25$

V_X	n							
	3	4	5	7	10	20	50	100
0.05	−4.204	−3.518	−3.192	−2.859	−2.618	−2.332	−2.125	−2.037
0.10	−3.977	−3.337	−3.037	−2.732	−2.512	−2.256	−2.074	−1.995
0.25	−3.353	−2.838	−2.602	−2.369	−2.208	−2.026	−1.898	−1.842
0.50	−2.538	−2.183	−2.027	−1.878	−1.778	−1.659	−1.574	−1.535

The values of parameter l for different distributions and different sample sizes are illustrated in Fig. 2.

We may observe several things:

(i) The highest values of l are needed in the case of Gumbel distribution.

(ii) The value of l depends on coefficient of variation in the case of Weibull distribution. Interestingly, higher values of l correspond to lower variance.

(iii) The values of l changes quite rapidly for smaller samples between $n = 3$ and $n = 20$. For larger sample sizes this value doesn't change much.
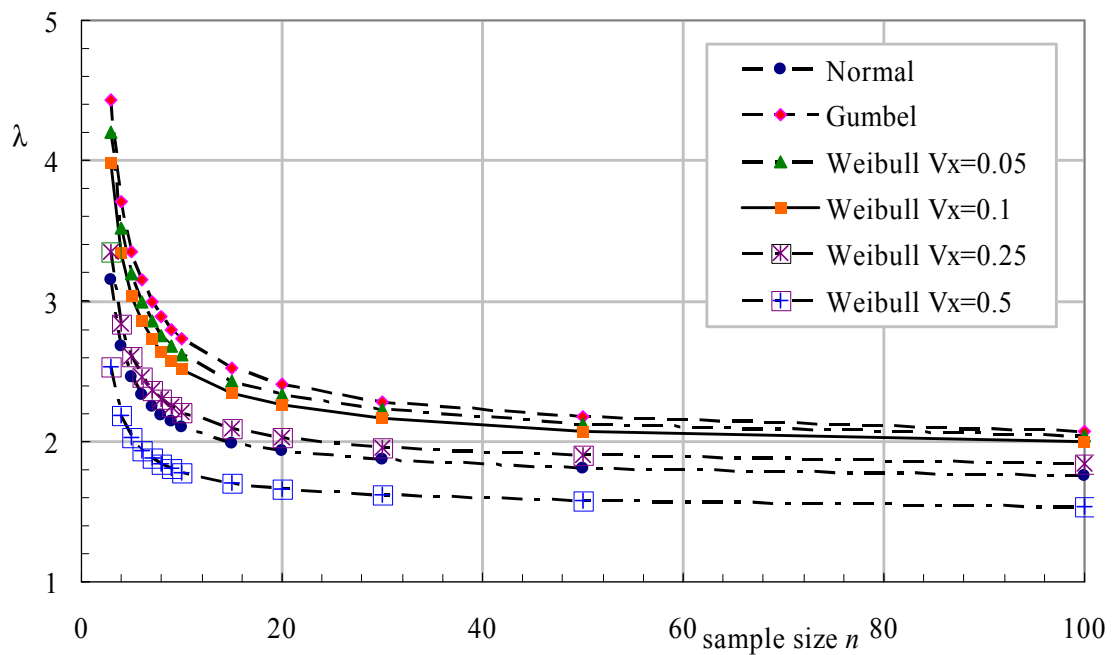
Figure 2: Value of *l* for different distributions and various sample sizes

4 An example

In order to define characteristic bending strength of finger jointed elements, destructive tests of small beams were performed in the laboratory of Slovenian National Building and Civil Engineering Institute. For the tests 20 specimens with dimensions 40 x 140 x 2600 mm, made from two pieces fully jointed in the middle of the length, were manufactured. The wood used was European spruce (Picea abies) and the joint was glued with one component polyurethane adhesive 1-K-PUR (PURBOND).

Specimens were tested according to EN 408 (Section13 Determination of bending strength). Test pieces were symmetrically loaded at two points at the distance of l/3 (820 mm) and were laterally restrained as shown in the Fig.3. Load was applied with the prescribed speed (maximum load was reached within ca 300 s).
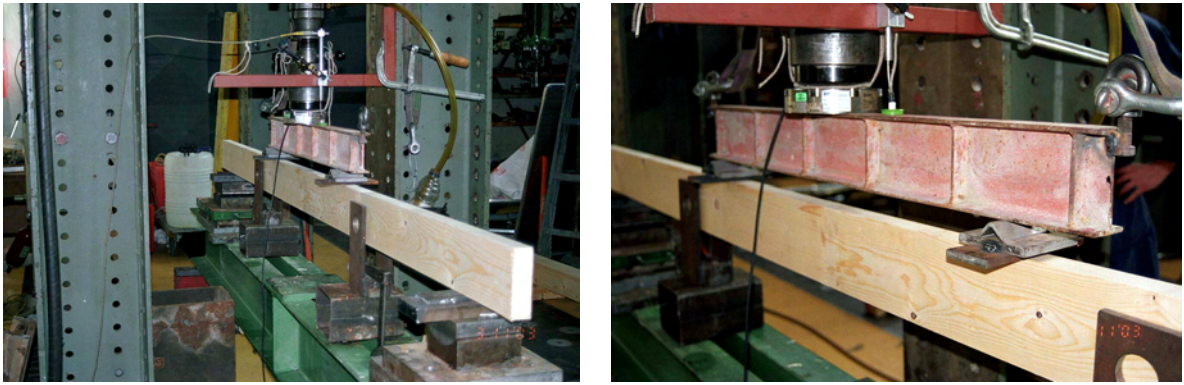
Figure 3: Test setup

After the failure beams were examined to determine locations and types of failure. All the beams failed in the cross sections where constant bending moment was applied (between the forces) – most of them on the joints or adjacent to them. Typical failed beams and details of failure are presented in Fig. 4.

The experimental results are summarized in Tab. 4.

Table 4: Strength g of beams (in MPa)

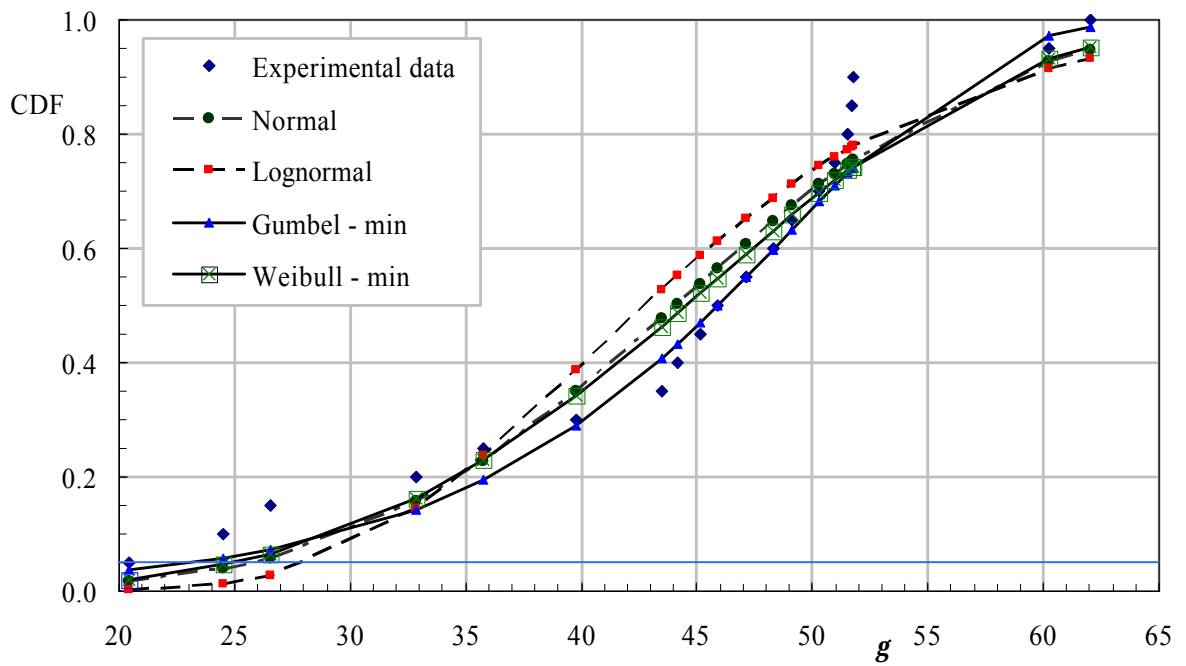| No.	1	2	3	4	5	6	7	8	9	10
γ	43.470	60.244	47.129	49.120	48.286	50.954	45.140	51.750	24.513	39.804
No.	11	12	13	14	15	16	17	18	19	20
γ	44.152	62.063	26.543	51.715	20.461	50.317	51.530	45.887	32.857	35.788

Figure 4: Typical failures

Figure 5: Experimental data and fitted cumulative distribution functions (CDF)

The experimental data and several fitted models according to different statistical distributions are shown in Fig. 5. The line for 5% probability is also shown, so that one can get an approximate estimate of 5% percentile, i.e. characteristic value, directly from the figure. The parameters of all distributions were obtained by the method of moments (see [1] or [5]). All parameters are summarized in the Tab. 5, where the parameters are defined as in Benjamin and Cornell reference book on statistics for civil engineers [1]. We may see that in the lower tail which is of our interest, the two distributions that corresponds the data best are Gumbel and Weibull. Lognormal distribution is the least suitable for our data.

Table 5: Distribution parameters obtained by the method of moments and the characteristic value estimates

Normal			Lognormal			Gumbel – min			Weibull – min		
m_X [MPa]	s_X [MPa]	g_{char} [MPa]	$\widetilde{m}_Y$ [MPa]	$s_{\ln Y}$	g_{char} [MPa]	a	u [MPa]	g_{char} [MPa]	k	u [MPa]	g_{char} [MPa]
44.086	11.150	22.54	42.740	0.2490	26.42	0.1150	49.104	17.25	4.485	48.319	21.50

Firstly, the characteristic value was determined according to EN 14358 [4] for 5% percentile. By using $n = 20$, $k_s = 1.93$ the characteristic value is $g_{char} = 24.01 \, \text{MPa}$.

Alternatively, characteristic value was determined from equation (5) for normal, Gumbel and Weibull and equation (7) for lognormal distribution. The results vary considerably as can be observed from Tab. 5. The result obtained based on the assumption of lognormal distribution is considerably higher than other values (about 15% higher than value according to the code[4]). On the other hand the value obtained for Gumbel (min) distribution gives a characteristic value that is 25% lower than the value according to the standard [4].

5 Conclusions

Determination of the characteristic values from small samples was analyzed for several different distributions. The main points of the present approach are as follows:

(i)   For normal distribution exact analytical formulation of the problem can be found. Analytical derivation results in a one-dimensional non-linear formula for determination of parameters.

(ii)   These parameters are used directly with the estimates of mean and standard deviation from the sample to evaluate the estimate of the characteristic value with previously prescribed confidence interval. Analytical results are confirmed by simulations.

(iii)   Lognormal distribution is directly connected to normal distribution through the exponential map. This relationship allows us to extend the formal algorithm from the normal to lognormal distribution.

(iv)   For other distributions relation to normal distribution is more complicated. The parameters l used in equation (5) have to be determined via numerical simulations. The result is a number of useful tables which can be very easily used for the estimation of characteristic value if relatively small samples are available.

(v)   An example for characteristic value determination in the case of strength of timber beams is used to illustrate possible differences when different assumptions about the statistical distributions are taken.

(vi) Lognormal distribution has a general shape which corresponds to the distributions of maximum values since its coefficient of asymmetry is always positive (it depends on coefficient of variation), whereas the coefficient of asymmetry is negative for all distributions of minimum values. As a result it is expected that the lognormal distribution would be a bad approximation for the strength of material where the distribution of minimum value (the weakest point) is sought. This is clearly true for the data analysed in this paper.

References

[1] Benjamin JR, Cornell CA, Probability, statistics and decision for civil engineers, McGraw-Hill: New York, 1970.

[2] Devroy L, Non-Uniform Random Variate Generation, Springer-Verlag: New York, 1986.

[3] EN 1990:2002 – Eurocode: Basis of Structural Design, 2002.

[4] EN 14358:2004 – Timber structures – Fasteners and wood-based products – Calculation of characteristic 5-percentile value and acceptance criteria for a sample.

[5] Siegrist K, Virtual Laboratories in Probability and Statistics, Univ. Of Alabama, Huntsville,

http://www.math.uah.edu/stat/point/index.xhtml.

[6] Zupan D, Turk G, On unbiased estimates of characteristic values, Risk analysis III (ed. Brebbia CA), pp. 385–394, WIT, Southampton (UK), 2002.

[7] Zupan D, Turk G, Characteristic value determination for arbitrary distribution, Risk analysis IV (ed. Brebbia CA), pp. 503–512, WIT, Southampton (UK), 2004.