

Percentiles Confidence Intervals Building Using Bootstrap-Modeling: Application for High-Tech Production Quality Control

Irina Gadolina
Natalija Lisachenko

•
IMASH RAS
TECHNOLOGIJA
Moscow
gadolina@mail.ru

Abstract

We propose a method for building confidence intervals for percentiles with application to quality control of the random properties of composite polymer materials strength. The basises (i.e. lower confidence limits for percentiles) are analyzed. The new developed method employs the statistical bootstrap modeling method. For the explanation of bootstrap procedure, the problem of building confidence interval for the mean value for the random values of composite material strength is presented. The result is compared with classical one. Due to bootstrap, it is possible to overcome some problems of classical statistical constraints. There is no need in postulating some type of distribution (normal or Weibull). Some real problems are presented with the comparison of the results by standard procedures and by the new method. It is shown that the agreement is satisfactory.

Keywords: percentile, bootstrap-modeling, composite material, strength

I. Introduction

During the certification of high-tech products, which include composites made by autoclave molding of prepreg HexPly brand, it is necessary to conduct the experiment to determine some strength properties. The values being investigated are: 1) the ultimate strength (σ_B , MPa); 2) the modulus of elasticity in tension (E , GPa); 3) strength interlaminar shear at normal temperature (τ_{20} , MPa); 4) strength interlaminar shear at elevated temperature (τ_{120} , MPa). Since the experimental values are random, it is necessary to conduct the statistical analysis. For this purpose 1) the percentiles $\gamma\%$ are estimated. (These values are almost the same as quintiles q , $\gamma\%=q/100$); 2) for 1% and 10% percentiles the $\alpha=95\%$ confidence intervals are estimated. The lower boundaries of those confidence intervals are called basises: namely A- basis for 1% percentile and B- basis for 10% percentile. Although the percentiles are themselves at some extend the interval characteristics, the necessity to build the confidence intervals for them makes the researchers face new challenges.

Let us dwell on the characteristics of percentiles. The percentile $\gamma\%$ is the characteristics of the sample, which express the ranges of the elements in the array as the numbers from 1 to 100, and indicate what percentage of the values are below a certain level. More generally, define the quantile $q=\gamma\%/100$ is been used. Mathematically, the quantile is determined as follows. Suppose there are independent and identically distributed random variables, for which there exists a distribution

function F with density distribution $f=F'$. Define q -th quantile of the population, such that $F^{-1}(q) \equiv \inf\{x \in R: F(x) \geq q\}$. Quantile $q=0.1$ (or, equivalently, the $\gamma\%=10$ th percentile) indicates that 10% of all values are below this level. Quantile (percentile) is a random value, which is determined by the sample, so it requires the assessment of their variability.

To calculate confidence intervals for the values of the percentiles obtained for the random sample the bases are used. Two types of bases: A-basis and B-basis [1] are investigated. They are the lower limits with confidence 95% for the percentiles 1% and 10% respectively. Before now it was obligatory to choose the appropriate type of distribution for solving this problem. For each type of distribution the complex dependencies are developed. For example, for the calculation bases with the assumption of normal distribution of a random variable in [1] the formulae are proposed:

$$\begin{aligned} B &= \bar{x} - k_B s, \\ A &= \bar{x} - k_A s, \end{aligned} \quad (1)$$

where \bar{x} is the average; s – is the mean square deviation and k_B и k_A are the coefficients of tolerance appropriate to the sample size. The values of these coefficients are given in tables or can be calculated with an error of not more than 0.2% by the following formulas:

$$\begin{aligned} k_B &= 1,282 + \exp\left(0,958 - 0,520 \ln(n) + \frac{3,19}{n}\right), \\ k_A &= 2,326 + \exp\left(1,340 - 0,522 \ln(n) + \frac{3,87}{n}\right). \end{aligned} \quad (2)$$

where n is a sample size.

II. Methods

In the present work as an alternative for methods [1] a method of constructing bases with the use of statistical bootstrap is proposed.

We will briefly review the description of the method of the statistical bootstrap. It was introduced in 1977, by the mathematician Bradley Efron [2]. The statistical bootstrap is a way of obtaining robust estimates of standard errors and confidence intervals, but not only this. It is being used to evaluate the variability of different characteristics. The method is based on the repeated simulation of the so-called bootstrap samples which constructed on the basis of the original sample with the replacement and is based on intensive use of computers. The number of bootstrap samples (denoted by R) should be large: in the present study were used $R=100$ and $R= 1000$. At the nowadays computers speed those number is not a problem. It will be only the fractions of a second calculations. The size of each bootstrap sample corresponds to the size of source sample, namely n , and the elements of the bootstrap samples are formed from the elements of the original sample, this is a random choice with replacement. For statistics, for which the exact mathematical expressions of variability exist, a number of studies have shown satisfactory agreement of the estimates based on bootstrap with the classical estimates (see also Appendix). To date, already has significant experience of applying statistical bootstrap to engineering problems, see for example [3]. On the other hand, mathematicians warn of excessive enthusiasm to this method: where the statistic theory is well developed and where the methods of data analysis in some sense close to optimal were found, the bootstrap has nothing to do.

In our case, for such statistics as, for example, $\gamma\%$ percentile, the mathematical expressions for the variance are complex and their optimality is not strictly proven. In this regard, it is interesting to compare the interval estimates of the bootstrap $\gamma\%$ percentile with ones, constructed by the used nowadays methods and to consider the possibility of new method introduction in the practice of engineering design. The evaluation of confidence intervals for quantiles using the bootstrap simulation was considered also in [4,5]. In [4] smoothing method confidence intervals for quantiles was proposed, in particular, nuclear assessment. In [5] the accuracy of bootstrap estimates of confidence intervals of quantiles, depending on the distribution of random variables was investigated.

Let's take for example h-th delivery of the s-th random value. In the investigated pool $h=7,8\dots 11$ (total 5 deliveries) and $s=1,2,3,4$ (total 4 characteristics). We accept $R=1000$ bootstrap samples which seems to be sufficient. Separately for each h and s we perform $R=1000$ random choices with replacement. There is a good function for it in [6] – "sample" function. It be metioned that it is necessary to include parameter "replace=true" among the others, or else the choice will be without replacement, which contradict the main bootstrap rule. [6] provide several algorithms for estimation of quantiles. We employed the algorithm $i=3$, which is taken by default. Sample quantiles of the algorithm type i are defined by:

$$Q[i](p) = (1 - \gamma) x[j] + \gamma x[j+1] \tag{3}$$

where $1 \leq i \leq 9$, $(j-m)/n \leq p < (j-m+1)/n$, $x[j]$ is the jth order statistic, n is the sample size, the value of γ is a function of $j = \text{floor}(np + m)$ and $g = np + m - j$, and m is a constant determined by the sample quantile type. For $i=3$ $\gamma = 0$ if $g = 0$ and j is even, and 1 otherwise. Type 3 is discontinuous sample quantile, as well as the types $i=1,2$.

After the algorithm of calculating quantiles has been chosen, the bootstrap modeling is performed. For the building confidence limits the random values of bootstrap samples are arranged into the variational series. The members of variational series with indexes LOW and UP form the $\alpha\%$ confidence interval for the statistic of interest. Here are the expressions:

$$\text{LOW} = \text{integer part} \left[\frac{1-\alpha \cdot 100}{2} R \right] \tag{4}$$

$$\text{UP} = \text{integer part} \left[\frac{1+\alpha \cdot 100}{2} R \right] \tag{5}$$

For building bases only the low limits are necessary. If number of bootstrap samples is $R=1000$ and confidence level as it is defined for bases $\alpha=95\%$, the lower index is: $\text{LOW}=25$.

III. Results

For the delivery #8 of carbon fiber specimens for the random value of ultimate strength σ_B , Mpa in Fig.1 the histogram of $q=0.1$ quantile (or that is the same $\gamma=10\%$ percentile) is shown.

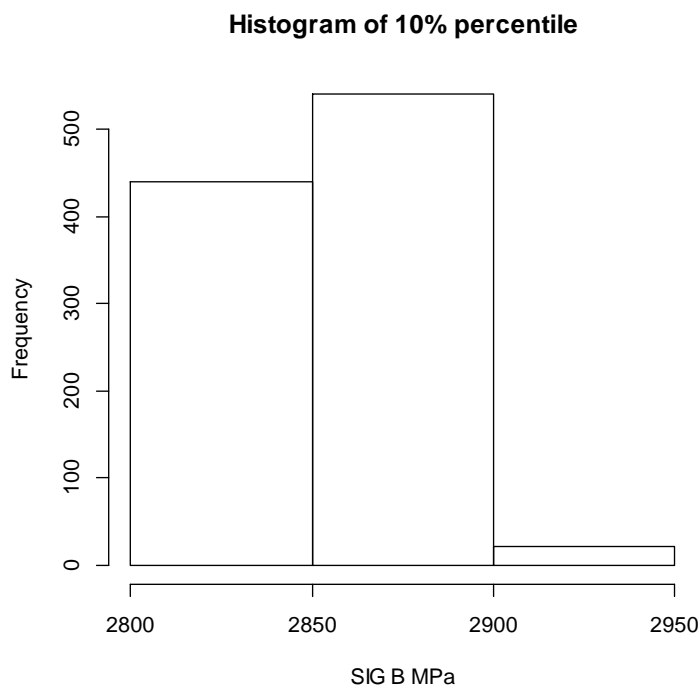


Figure 1: Histogram of percentiles

It can be seen, that the distribution is far from being normal, so we have all the reasons to employ the bootstrap, which is free from requirements of the normal distribution. In the Figure 2 the cumulative distribution function of the bootstrap estimations of percentile $Y=10\%$ is shown (“ecdf” function in [6]). The random value of ultimate strength σ_B is shown in MPa.

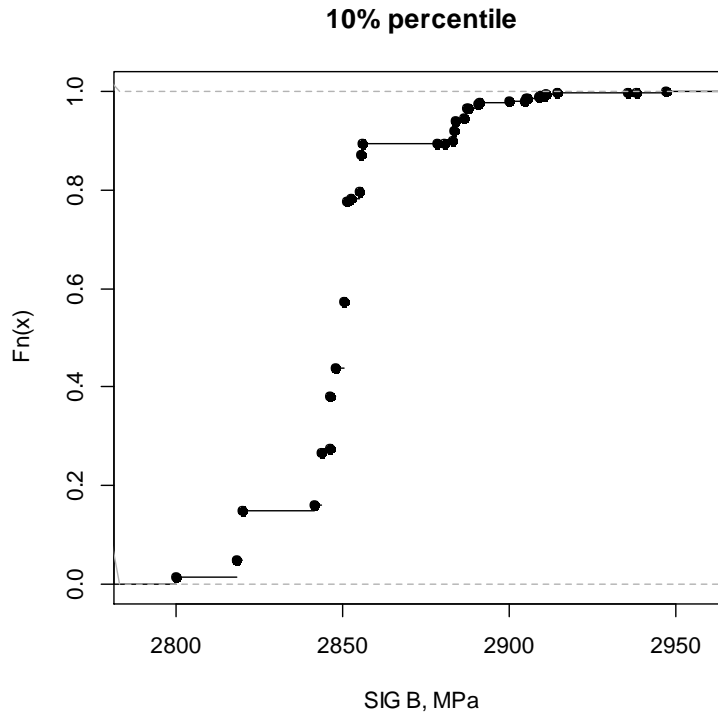


Figure 2: Percentiles cumulative distribution

Following the rules for building the bootstrap confidence interval (4,5) we estimate the indexes of variational series: $LOW=25$; $UP=975$. For the bases we need, actually, only the lower index. For deliveries the data of bases are presented in Table 3.

Table 3: Ultimate strength σ_B , $MIIa$ for some deliveries of HexPlay

Delivery index	Mean	Standard deviation	10% percentile	B-basis	
				bootstrap	[2]
7	3013.3	142.6	2817.0	2789	2693
8	2962.5	84.9	2850.5	2818	2780
9	2719.3	154.6	2512.2	2580	2361
10	2848.9	172.3	2630.2	2580	2463
11	2729.7	134.7	2579.3	2501	2403

IV. Discussion

Because of standard requirements in industry it is important to estimate bases of some statistical characteristics. The proposed method was applied to the quality control of the strength characteristics of carbon fiber composite specimens. The comparison has been made with the results, obtained by nowadays applied methods. The agreement is shown to be satisfactory. On the other hand the new developed method possesses some advantages, for example, it is free from distribution. It all refers to B-bases (the lower 95% limit of 10% percentile). As for A-bases (the

lower 95% limit of 1% percentile) there are some problems. They are due to very small samples size (up to $n=30$). For building A-basises for small samples it might be necessary to develop new methods, based on numerical imitation.

Appendix

Example of applying bootstrap for average value confidence intervals building.

To explain the method of construction confidence intervals using bootstrap-modeling and to make reader more familiar with bootstrap procedure, we employ the problem of building confidence intervals for the average value of the normal set. This problem was chosen because in has a good classical decision. As the object of investigation the shear strength under room temperature τ_{20} , MPa of the composite samples has been taken. The delivery number was 7, because for this set a good agreement with the normal distribution was obtained [7]. Here the number of bootstrap trials was chosen as $R=100$.

The initial set is presented in Table A1.

Table A1: Shear strength under room temperature τ_{20} , MPa, delivery number 7, initial sample

Order index	1	2	3	4	5	6	7	8	9	10
τ_{20} , MPa	86.1	89.7	97.1	95.9	93.7	94.4	90.6	93.6	96.4	88.8
Order index	11	12	13	14	15	16	17	18	19	20
τ_{20} , MPa	101	89.3	85.9	92.8	94.3	91.5	90.6	91.5	92.8	90.1
Order index	21	22	23	24	25	26	27	28	29	30
τ_{20} , MPa	97.2	94.9	91.5	93.8	96	95.9	89.3	99.8	95.9	107.8

Initial sample parameters (Table A1)

$$\text{Mean value}(\tau_{20}) = 93.61 \text{ MPa} \quad (6)$$

Standard deviation (τ_{20}) = 4.502 MPa

According to the bootstrap rules R bootstrap samples are simulated. In Table A2 an example of simulated k -th bootstrap sample is shown, constructed based on the source sample shown in Table A1:

Table A2: k -th bootstrap sample (example) modeled on the base of initial sample (Table A1)

	1	2	3	4	5	6	7	8	9	10
τ_{20} , MPa	107.8	93.6	90.1	99.8	92.8	97.1	89.3	90.5	91.5	89.3
Order index	11	12	13	14	15	16	17	18	19	20
τ_{20} , MPa	107.8	96.4	93.6	94.4	93.8	90.1	91.5	89.3	101.0	92.8
Order index	21	22	23	24	25	26	27	28	29	30
τ_{20} , MPa	89.3	88.8	89.7	101.0	97.2	93.7	93.8	89.3	89.7	93.8

It can be seen that some random values are repeated in k -th bootstrap sample (in Table A2, for example, the elements 1 and 11 are the same: $\tau_{20}=107.8$ MPa). Some elements of the Table A1 was not included even once in the k -th sample (for example, item number 1: $\tau_{20}=86.1$ MPa), but it might be included in the bootstrap sample number $k+1$.

k -th sample parameters (Table A2):

$$\text{Mean value}(\tau_{20}) = 93.99 \text{ MPa} \quad (7)$$

Standard deviation (τ_{20}) = 5.137 MPa

It is seen that the characteristics at the initial sample (6) and k -th sample (7) differ, albeit only slightly. For each $k=1,2,\dots,R$ bootstrap sample it is necessary determine the characteristics of interest,

namely average. The set of estimations characterize the variability of the point estimate.

The histogram shape looks like the shape of the normal distribution. The average value for the bootstrap samples is the bootstrap mean(τ_{20})=93.603 MPa. This value is close to the average for the original sample (6). For building the confidence intervals with probability $\alpha=90\%$ for the average the formulae (4,5) are applied. For the given parameters $\alpha=90\%$ and $R=100$ the values in indexes are: $LOW=3$ and $UP=97$. The standard statistical estimation [8] for the same purpose provides very close values. The confidence limits built by two methods are shown in Table A3.

Table A3: $\alpha=95\%$ confidence limits for the mean value of shear strength value [MPa]

LOW 90%, Student's	LOW 90%, bootstrap	UP 90%, Student's	UP 90%, bootstrap
92.22	92.38	95.0	95.10

References

- [1] Composite Materials Handbook – 17 (CMH-17). SAE International on behalf of CMH-17, Wichita State University, March 2012 – Chapter 8 Statistical Method. P. 552-712.
- [2] P.Diaconis, B.Efron. Computer-intensive method in Statistic// Scientific American. 1983. v. 248. N5 . P. 116-131.
- [3] Adler U.P., Gadolina I.V., Ljandres M.N. Bootstrap-modeling for confidence-interval building by censored sets] // Zavodskaja laboratorija. 1987. №10. P.90-94.
- [4] By Yvonne H.S., S.Lee. Iterated smoothed bootstrap confidence intervals for population quantiles// The Annals of Statistics. 2005. v.33. No.1. P.437-462.
- [5] B.V.Veshn'akov, A.I.Kibzun. Application of bootstrap method for quantile function estimation] // Avtomatika I telemekhanika. 2007. v.1, P.46-60
- [6] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [7] Lisachenko N.G., Popov A.G., Gadolina I.V. The analysis of the stability of the strength properties of modern carbon fiber reinforced plastics / Conference proceedings: Deformirovanie i razrushenie kompozitsionnykh materialov DFCMS -2016, IMASH RAN, Moscow, Russia, 18-20 October, 2016, P.74-76.
- [8] Stepnov M.N., Zinin A.V. Prediction of materials and machines parts characteristics. M. Innovasionnoe mashinostroenie. 2016. 391 c.