A NOTE ON THE EFFECT OF STATISTICAL SAMPLE SIZE ON FRACTURE TOUGHNESS CHARACTERIZATION IN THE DTB TRANSITION REGION

Sreten Mastilovic[1,*], Branislav Djordjevic[2], Aleksandar Sedmak[3]

[1] University of Belgrade, Institute for Multidisciplinary Research, Kneza Viseslava 1a, Belgrade
e-mail: misko.mastilovic@imsi.bg.ac.rs
[2] Innovation Center of Faculty of Mechanical Engineering, Kraljice Marije 16, Belgrade, Serbia
e-mail: brdjordjevic@mas.bg.ac.rs
[3] Faculty of Mechanical Engineering,University of Belgrade, Kraljice Marije 16, Belgrade, Serbia
e-mail: asedmak@mas.bg.ac.rs
[*] Corresponding author

Abstract

The ferritic steels, widely used as pressure-vessel materials in nuclear industry, are prone to embrittlement when exposed to neutron irradiation or temperature reduction within the DTB (ductile-to-brittle) transition region. This embrittlement may be accompanied by the increased size effect, which is a pronounced consequence of fracture mechanics not exhibited in the traditional plasticity theory. Therefore, the fracture toughness in the DTB transition temperature region is a stochastic *extrinsic* property well known for its aleatory variability. Consequently, the extremely-pronounced experimental data scatter necessitates the use of the statistical approach to material characterization. The recently proposed two-step-scaling approach to estimate the size effect of fracture toughness CDF (cumulative density function) in the DTB transition region relies heavily on regularity of arrangement of experimental data points for the two *input* sample sizes. This regularity of measurement values becomes an inherently iffy proposition in the case of statistically small data sets. Therefore, the ability of our novel approach to predict objectively the fracture toughness probability outside the experimental domain may be impaired in absence of the sufficient statistical size of the input data sets. Since the large-scale fracture toughness tests for nuclear pressure-vessel steels at low temperatures are very expensive, the present study is concerned with this issue of the statistically sufficient sample size. There are various statistical techniques to determine the sample size needed for a study, including power analysis and sample size calculation formulas. The appropriate method depends on the type of study, the research question, and the statistical analysis planned. These issues are addressed in this article.

Key words: ferritic steels, embrittlement, fracture toughness, size effect, statistical size

1. Introduction

The brittle fracture of ferritic steels is characterized by pronounced sample-to-sample variations of fracture toughness (especially for small-size specimens) and a statistical approach is a necessity. The field of Probabilistic Fracture Mechanics emerged from the fact that all fracture toughness measures are inherently distributed quantities (that is, they are best represented by a

range of values and not by a single value) [1]. The Weibull theory is one of the first size-effect theories of the strength of materials that is developed on purely statistical arguments [2]. The Weibull statistics is based on the weakest-link theory, which in this particular case implies lack of stress redistribution prior to cleavage fracture. When it comes to ferritic steels at DTB transition temperatures, addressed in the present study, plasticity mechanisms and stress redistribution are largely suppressed, which results in catastrophic failure of the whole specimen [3]. Consequently, the nature of the size effect appears inherently statistical – that is, of the kind traditionally described by the Weibull distribution. Landes and coworkers (e.g., [4]) based their statistical approach on the premise that the cleavage fracture toughness is controlled by the weakest link at the crack front. They used the two-parameter Weibull distribution, $\mathsf{W}(\beta, \eta)$

$$CDF\left(x \mid \beta,\eta\right)=1-\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right]; \qquad PDF\left(x \mid \beta,\eta\right)\equiv\frac{d}{dx}CDF\left(x\right)=\frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] \quad (1)$$

illustrated in Fig. 1, which is a warhorse of DTB fracture toughness assessment to this day (e.g., [3, 5]). The Weibull scale (η) and shape (β) parameters (Fig. 1b) are material constants sensitive to the specimen preparation, surface condition and temperature. Note the general size-effect trend (stemming from the weakest-link theory) that the increase in (sample) size ($W\uparrow$) should results not only in the reduction of the typical fracture toughness value ($\eta\downarrow$) but also in the *reduction* of the fracture toughness *scatter* ($\beta\uparrow$) and the number of necessary tests per CT size ($n_{\max}\downarrow$).

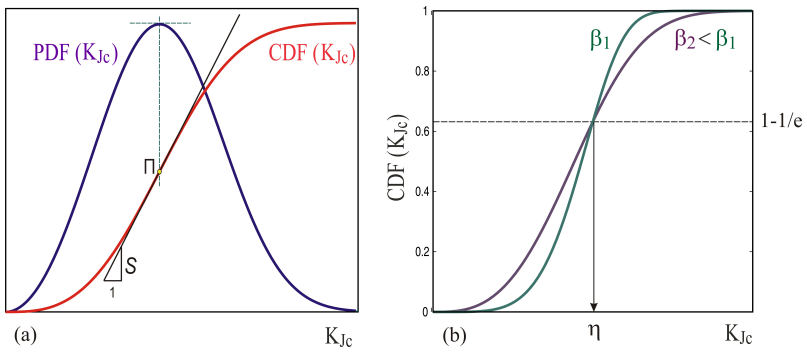

Fig. 1. The Weibull distribution, $\mathsf{W}(\beta, \eta)$. (a) Cumulative distribution function (CDF) and the probability density function (PDF). (b) Illustration of the Weibull parameters of scale (η) and shape (β). K_{Jc} [MPa$\sqrt{m}$] is the critical value of the stress intensity factor used in the master curve (one of the fracture toughness measures used in Linear Elastic Fracture Mechanics) [7, 8]. (Note that *S* marks the maximum CDF slope.)

A brief historical sketch of some of the, arguably, most influential statistical studies of cleavage fracture toughness of ferritic steels that make use of the Weibull statistics is, for example, recently presented in [6].

The two-step-scaling (2SS) approach has been proposed recently [3, 9] to predict the size effect on the fracture toughness CDF in the DTB transition region. It cannot be overemphasized that it relies crucially on the "regularity of arrangement" of experimental data points for the two *input* sample sizes. This regularity becomes an inherently iffy proposition in the case of *statistically* small data sets (i.e., the small number of realizations of the same statistics, *n*). Therefore, the ability of our novel approach to predict objectively the fracture toughness probability (especially in the extrapolation domain) may be impaired in absence of the sufficient statistical size of the input data sets.

The present study is concerned with this issue of the statistically "large enough" sample size. The very large data sets required to validate statistical methods are already identified as the main problem in the cleavage fracture toughness research [8]. Since the "large" for one analyst may be

"small" for another, what is sufficient in this context? As an example, a sample size of 30 is fairly common in Statistics since it often increases the confidence interval of data sets enough to warrant assertions against findings. Landes [11] suggested that the satisfactory handling of the Weibull slope appears to be achieved with sample sizes between 20 and 50. However, in experimental studies of fracture toughness of ferritic steels in the DTB transition region, the relatively small sample sizes of 10÷12 are often found (e.g., [5, 7]). Obviously, the sample sizes represent a compromise between the statistical analysis confidence and the economic aspects (reflected by judicious and justified use of time and resources). What is a sufficiently large sample size for the purpose? The appropriate method to get an answer, in general, depends on the type of study, the research question, and the statistical analysis planned. The sufficient size of a statistical sample depends on various factors, including the level of precision or accuracy desired, the variability in the population being studied, and the level of confidence required. In general, a larger sample size will provide more precise estimates of population parameters and a higher level of confidence in the results. Again, as a rough guideline, a sample size of at least 30 is often considered sufficient for many statistical analyses. However, this may not be the case in all situations and it is always recommended to investigate the issue on a case-by-case basis.

2. The sources of erratic patterns of experimental data distribution in terms of CDF

The stochasticity of the fracture toughness of ferritic steels in the DTB transition region is "the nature of the beast". In other words, the aleatory variability and epistemic uncertainty are *inherent* in the problem. Nonetheless, when it comes to the above-mentioned "regularity of arrangements" of experimental data points, the primary source of unreliably irregular (non-objective) behavior of fracture toughness CDFs, addressed in this note, appears to be the size of the statistical sample. There are some basics formulas in Statistics for sample size calculation, although sample size calculation differs from technique to technique [10]. (For example, when the means of two populations are compared, if the sample size is less than 30, the t-test is used; if the sample size is greater than 30, the z-test is recommended.) The only aspect a researcher needs in order to justify a sample size based on reliability is the desired width of the confidence interval with respect to their inferential goal, and their assumption about the sample standard deviation of the measure.

As an illustration for the statistical sample size effect of the fracture toughness CDF, the EURO data set is used [7, 8], which is obtained by using CT (compact tension) specimen of the quenched and tempered pressure-vessel steel 22NiMoCr37 frequently used in nuclear power plants. The experimental K_{Jc} data (considered "valid") are actually taken from Annex 2 of the report [7], extracted from the complete EURO data set. This particular data set is selected *exclusively* based on two conveniences: (i) the relatively large size (55 realizations), and (ii) the experimental results are obtained from the same laboratory (see Fig. 2 caption).

It can be observed from Fig. 2 that the progressive reduction of the size of the randomized-order sample eventually results in the break-up of the CDF objectiveness (representativeness). For example, the ½ reduction plot (28 solid green circles) follows reasonably closely the full-deck (55 empty purple squares) CDF plot. The agreement between Weibull slopes is within ±5% and between the Weibull scale factors even better. Then, after further size reductions, the CDFs diverge more and more. Eventually, a simple visual inspection is sufficient to observe that the Weibull parameters (η and β) corresponding to the ⅕ reduction (11 solid red triangles) differ significantly from the original. Interestingly, the ⅕ reduction data set in itself obeys the Weibull CDF sigmoid shape rather well but the parameter values are inconsistent.[1] These conclusions are further supported by the successfully reducing values of the statistical measure of the goodness-of-fit (the adjusted P-value) with the sample size reduction also shown in Fig. 2. (In some other examples, not

[1] For example, the horizontal line corresponding to CDF = 1-1/e ≈ 0.632 reveals the difference in the Weibull scale parameters of more than 10% between the full data set (n = 55) and the ⅕ reduction set (11).

presented herein, the distribution of 11-point set results does not even suggest the sigmoid shape of the Weibull CDF.) It goes without saying that using an experimental set of 11 data points, in this particular case, would bias the CDF (K_{Jc}) predictions obtained by the 2SS approach.

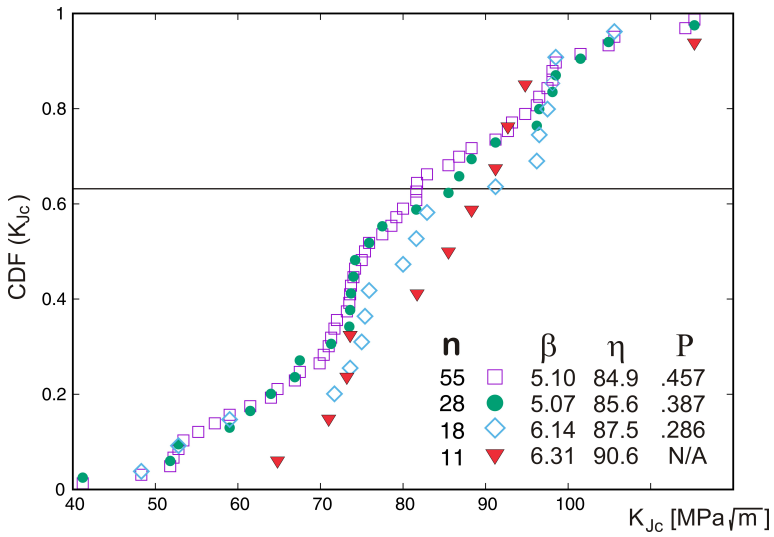

Fig. 2. Effect of random reductions of the data set (1/1, 1/2, 1/3, 1/5) on K_{Jc} CDF for 22NiMoCr37 steel at T=-110ºC and w=25 [7]. The Weibull parameters (β, η) correspond to the calculated P-value (χ^2 test, Appendix A). (Data obtained by the GKSS Research Centre – now: the Helmholtz-Zentrum Hereon, Germany.)

2.1 Random reductions of data set as a source of uncertainty

The random reductions of the original data set are a source of stochasticity in itself. This is an unavoidable consequence (an essential facet) of the process of random reduction. To illustrate this point, the ⅕ random reduction is performed independently four times on the original experimentally obtained fracture toughness set (n = 55 data points). The results are presented in Fig. 3.

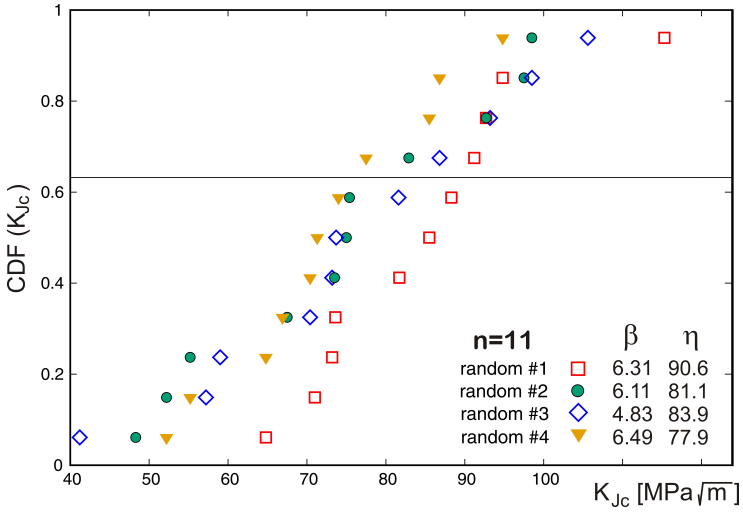

Fig. 3. Four different one-fifth random reductions of the original CDF (K_{Jc}) experimental data set.

Two observations readily come to mind. First, all four CDF patterns that reveal (at least to some extent) the desired sigmoid shape (Fig. 1) are themselves relatively irregular. (This is *apparently* an unavoidable consequence of the small size of the statistical sample; the rule rather than the

exception.) Second, the difference between the four data sets (and the perceived Weibull CDF data fits) is not negligible, to put it mildly, which is evident from the corresponding Weibull parameters (β, η) that are widely different.

Two observations illustrated in Figs. 2 and 3 question representativeness (objectivity) of the 11-point data sets. Fig. 2 suggests that the trend is improving with the increase of the sample size.

3. The use of the experimental data from different laboratories as a source of stochasticity

Assuming that all experimental specimens are cut from the same large segment of the material provided by the same manufacturer, it is a valid question whether testing in different laboratories is, in itself, a source of stochasticity of the fracture toughness measurements.

To investigate this eventuality, the data sets obtained from two laboratories (GKSS and Siemens) for the same CT specimen size ($W = 25$ mm) and two different temperatures (-154 ºC and -60 ºC) are compared. In this way, the uncertainty due to the specimen size is excluded from consideration. The corresponding CDF (K_{Jc}) plots are shown in Fig. 4.

The CDF comparisons yield the different results for two temperatures. Namely, at $T = -154$ ºC the Weibull CDFs from two data sets from different labs reveal significantly different Weibull parameters η and β. On the other hand, the Weibull CDFs at $T = -60$ ºC show a fair level of similarity. One can argue that the more brittle behavior at the lower cryogenic temperature is expected to be more stochastic. But also, it is rational to assume that the discrepancy between the two curves is due to the smaller statistical sample size for the GKSS data set (only 11 points).

Therefore, the analysis is inconclusive at present: it confirmed that the differences between the fracture toughness measurements from different laboratories *may* exist, but the determination of the root cause of this discrepancy with satisfactory certainty requires additional work on new data.

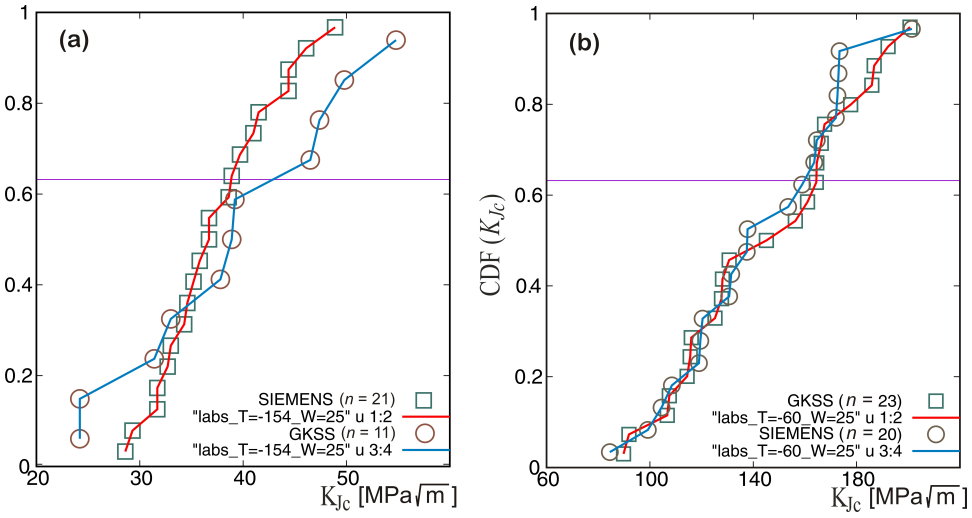

Fig. 4. CDF (K_{Jc}) for 22NiMoCr37 steel obtained from two laboratories: GKSS (squares and red lines) and Siemens AG (Power Generation Group, Erlangen, Germany) (circles and blue lines). The tests are performed at the same CT specimen size ($W = 25$ mm) at two different temperatures: (a) -154ºC and (b) -60ºC.

4. Summary

The present preliminary study is dedicated to investigation of the effect of the statistical sample size on modeling the Weibull CDF (K_{Jc}) in the DTB transition region by the novel 2SS approach. This approach is sensitive to the objectivity of the fitting of the fracture toughness measurement data sets at two input CT specimen sizes. The big question of every statistical analysis emerges: how many realizations (of the same statistics) are enough in this case?

There are various statistical techniques to determine the sample size needed for a study. At the risk of stating the obvious, in statistical practice "the more the merrier" but in engineering practice using a sample size that is too large could incur a significant waste of both resources and time (the penalty for excess conservatism is increased costs). Thus, based on the data sets examined, it seems reasonable to confirm that small sample sets consisting of less than approximately 15 data points (measurement realizations) are of questionable utility for the above-stated purpose. As a rule of thumb, the minimum data set size can be apparently set from 25 to 30. This observation is consistent with time-honored statistical practices in general and the previous DTB assessments of ferritic steels in particular.

Under these circumstances, bearing in mind that available data set sizes are often limited to 10-12 data points, combining data sets from different laboratories becomes necessary. This practice should be applied judiciously as it might be in itself an additional source of data uncertainty in the case of materials exhibiting the weakest-link type of fracture mechanism.

References

[1] U.S. NRC, *Important Aspects of Probabilistic Fracture Mechanics Analyses*. Technical Letter Report TLR-RES/DE/CIB-2018-01, 2018.

[2] Weibull, W., A Statistical Theory of the Strength of Materials. Generalstabens Litografiska Anstalts Förlag, Stockholm, 1939.

[3] Mastilovic S., Djordjevic B., Sedmak A. *A scaling approach to size effect modeling of J_c CDF for 20MnMoNi55 reactor steel in transition temperature region*. Engineering Failure Analysis 131: 105838, 2022.

[4] Landes, J., Zerbst, U., Heerens, J., Petrovski, B., Schwalbe, K., *Single-Specimen Test Analysis to Determine Lower-Bound Toughness in the Transition*, in Fracture Mechanics: Twenty-Fourth Volume, ed. J. Landes, D. McCabe, and J. Boulet (West Conshohocken, PA: ASTM International), 171-185, 1994.

[5] Djordjevic B., Sedmak A., Petrovski B., Dimic A., *Probability Distribution on Cleavage Fracture in Function of J_c for Reactor Ferritic Steel in Transition Temperature Region*, Engineering Failure Analysis, 125, 105392, 2021.

[6] Djordjevic B., Sedmak A., Mastilovic S., Popovic O., Kirin S. *History of ductile-to-brittle transition problem of ferritic steels.* Procedia Structural Integrity 42: 88–95, 2022.

[7] Lucon E., Scibetta M., *Application of Advanced Master Curve Approaches to the EURO Fracture Toughness Data Set*. Open Report of the Belgian Nuclear Reserach Centre SCK•CEN-BLG-1036. Mol, Belgium, 2007.

[8] Heerens J., Hellmann D., *Development of the Euro fracture toughness dataset*. Engineering Fracture Mechanics 69: 421–4492002, 2002.

[9] Mastilovic S., Djordjevic B., Sedmak A., *Corrigendum to "A scaling approach to size effect modeling of J_c CDF for 20MnMoNi55 reactor steel in transition temperature region"* [Eng. Fail. Anal. 131 (2022) 105838] Engineering Failure Analysis 142: 106751, 2022.

[10] Hines W.W., Montgomery D.C., *Probability and Statistics in Engineering and Management Science*. John Wiley & Sons; 3rd Edition, 1990.

[11] Landes J.D., The effect of size, thickness and geometry on fracture toughness in the transition. GKSS 92/E/43. GKSS, Geesthacht, Germany, 1992.

Appendix A – The Hypothesis-Testing for Goodness of Fit

The original full (1/1) data set of K_{Jc} CDF for 22NiMoCr37 steel at T = -110ºC and W = 25 mm [7] and the three randomly reduced (1/2, 1/3, 1/5) data sets (Fig. 2) are subjected to the hypothesis testing procedure to establish whether the random variable in question follows the 2-parameter Weibull distribution, $\mathsf{W}(\beta, \eta)$, with the estimated parameters.

The formal goodness-of-fit test procedure is based on the chi-square (χ^2) distribution [10]. The χ^2 is a statistical test that examines whether a random sample data follows a theoretical probability distribution with estimated parameters. The random results of the fracture toughness measurements arranged in the four datasets (1/1, 1/2, 1/3, 1/5) are arrayed in frequency histograms (one for each data set), having k class intervals.[2] The observed frequency in the i-th class interval is marked O_i. From the hypothesized $\mathsf{W}(\beta, \eta)$ with two parameters estimated based on the Weibull plot and the maximum likelihood method (Fig. A1) the expected frequency E_i in the corresponding class interval can be readily computed. (A common practice in constructing the class intervals is to choose their boundaries so that the magnitudes of the expected frequencies are equal for all cells [10].) The test statistics is

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{A1}$$

It can be demonstrated that the test statistics (A1) follows approximately the χ^2 distribution with $dof=k-p-1$ degrees of freedom (where p is the number of parameters in the hypothesized distribution; in this case p=2). It cannot be overemphasized that this approximation improves as n increases [10]; the small samples are therefore inherently handicapped. The null hypothesis (H_0) that the random sample conforms to the hypothesized $\mathsf{W}(\beta, \eta)$ is rejected with the confidence level $(1-\alpha)\cdot 100\%$ if the test statistics exceeds the critical statistics, $\chi_0^2 > \chi_{\alpha,dof}^2$. The significance level (α) is the probability of rejecting the null hypothesis when it is true.

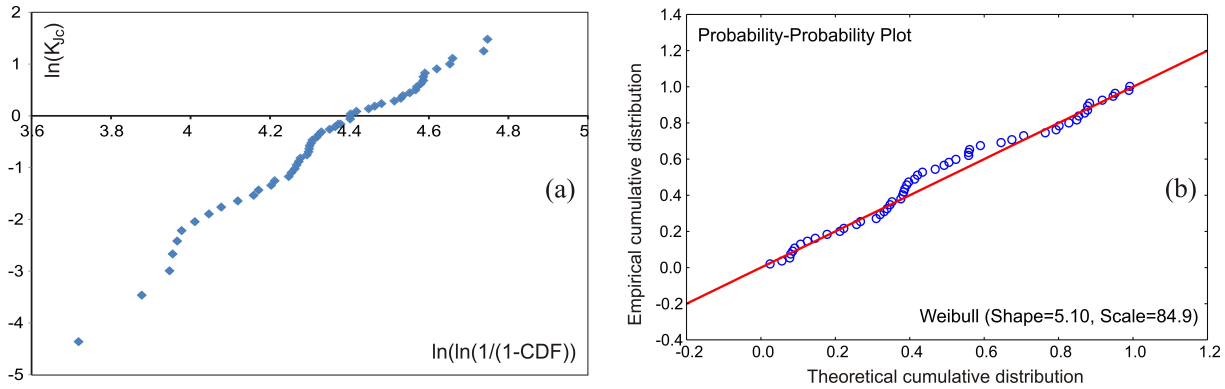

Fig. A1. An example of: (a) the Weibull plot and (b) the probability-probability plot for the full (1/1) data set of K_{Jc} CDF for 22NiMoCr37 steel at T = -110ºC and W = 25 mm.

The χ^2 test is extremely sensitive to the sample size. Since the full data set has a relatively large number of data points (n = 55) an attempt is made to use k = 6 data cells (class intervals). Consequently, the limits of the class intervals (a_i, i = 1, k) in the first column of Table A1 are

[2] The number of class intervals should be reasonably large. More importantly, although there is no general agreement regarding the minimum value of expected frequencies E_i, the values of 3, 4, and 5 are *commonly* regarded as minimal [10]. This requirement imposes a stringent limitation to the small data sets.

determined by using Eq. (A2) under the constraint that all expected frequencies have the same magnitude $E_i = n \cdot p_i = n/k = 55/6 \approx 9.167$:

$$p_i = \int_{a_{i-1}}^{a_i} PDF(x)\,dx = \frac{1}{k}; \quad (i = 1, k),\ a_0 \equiv -\infty \tag{A2}$$

Based on Eq. (A1) and the data in Table A1, the computed value of the chi-square statistics is

$$\chi_0^2 = \frac{(9-9.167)^2}{9.167} + \frac{(8-9.167)^2}{9.167} + \frac{(14-9.167)^2}{9.167} + \frac{(7-9.167)^2}{9.167} + \frac{(6-9.167)^2}{9.167} + \frac{(11-9.167)^2}{9.167} = 4.672 \tag{A3}$$

Since two parameters in the Weibull distribution have been estimated ($\beta = 5.10$, and $\eta = 84.9$; Fig. A1b), the calculated value (A3) should be compared to a chi-square distribution with 3 degrees of freedom. Moreover, the χ^2 P-value for the observed and expected frequencies shown in Table A1 is calculated in Excel to be 0.457. Thus, the null hypothesis that the random sample conforms to W (5.10, 84.9), cannot be rejected since: (i) $\chi_0^2 = 4.672 < \chi^2_{0.05,3} = 7.81$ (Table III,Ref. [10]), and (ii) the corresponding P-value is greater than the significance level (0.457 > 0.05). Consequently, it can be concluded that there is *no reason to believe* that the K_{Jc} measurement data **is not** distributed in accordance with W (5.10, 84.9).

	Class Interval	Observed Frequency, O_i	Expected Frequency, E_i
1	$x \le 60.81$	9	9.167
2	$60.81 \le x \le 71.13$	8	9.167
3	$71.13 \le x \le 79.02$	14	9.167
4	$79.02 \le x \le 86.49$	7	9.167
5	$86.49 \le x \le 95.20$	6	9.167
6	$95.20 \le x$	11	9.167
	sum	**55**	**55.00**

Table A1. Class intervals, observed and expected frequencies for $n = 55$ and $k = 6$.

The χ^2 test applied to the reduced data sets (½) with $n = 28$, yielded the same conclusions. Namely, for $n = 28$ and the estimated Weibull parameters $\beta = 5.07$ and $\eta = 85.6$:

$$\chi_0^2 = 4.143 \quad < \quad \chi^2_{.05,\,5-2-1} = \chi^2_{.05,\,2} = 5.99 \tag{A4}$$

while P-value = 0.387 > $\alpha = 0.05$. Similarly, for $n = 18$ (⅓) and the estimated Weibull parameters $\beta = 6.14$ and $\eta = 87.5$:

$$\chi_0^2 = 3.778 \quad < \quad \chi^2_{.05,\,4-2-1} = \chi^2_{.05,\,1} = 3.84 \tag{A5}$$

while P-value = 0.286 > $\alpha = 0.05$. Consequently, the two null hypotheses cannot be rejected with the confidence level of 95%. Notably, for the smaller sample size (A5), the hypothesis came very close to be rejected since the two χ^2 values are within 2% from each other and the P-value is smaller than for other samples. Thus, the *n*-reduction trend is toward rejection of the null hypothesis.

Finally, the smallest randomly reduced data set (⅕), with $n = 11$, is evidently too small for the χ^2 test since the expected frequencies ($11 / 4 = 2.5 < 3$) would be below even the most liberally fixed minimal value (= 3; see footnote 2 on the preceding page).