

DOI: 10.19416/j.cnki.1674-9804.2017.02.017

一种基于威布尔分布的金属疲劳试验 异常数据判别方法

An Outlier Detection Method of Metal Fatigue Tests Based on Weibull Distribution

唐兆田 / TANG Zhaotian

(上海飞机设计研究院, 上海 201210)

(Shanghai Aircraft Design and Research Institute, Shanghai 201210, China)

摘要:

金属疲劳试验的数据符合威布尔分布,是非对称的统计分布,而现有的异常数据判别方法均基于对称性分布(正态分布、 t 分布)。威布尔分布与对称性分布(正态分布、 t 分布)主要差异体现在低寿命区和高寿命区,这正是异常数据出现的区域。基于威布尔分布推导了一种异常数据判别方法,推导过程中进行了合理地工程处理。经比较,该方法相对于基于对称性分布的判别方法更可靠。

关键词:金属疲劳试验;异常数据;威布尔分布;统计方法

中图分类号:V215.5

文献标识码:A

[Abstract] Data of metal fatigue test is described with Weibull distribution which is asymmetrical. But general methods for outlier detection in metal fatigue tests are based on symmetrically statistical distribution. The difference between Weibull distribution and symmetrically statistical distribution is often found in the areas of small life and high life where outlier appears. A method of outlier detection is derived based on Weibull distribution in this paper. And some reasonable measures are used in the derivation. By comparison, the method in this paper is more reliable than those based on asymmetrical distribution.

[Keywords] metal fatigue test; outlier; weibull distribution; statistical method

0 引言

处理金属材料疲劳试验数据时,有时需要确定可疑数据是否为异常数据,通常采用的异常数据判别方法都是以对称性分布为前提的判别方法^[1](譬如 Paūta 法、Chauvenet 法、Grubbs 法、Dixon 法、罗马诺夫斯基准则法等),但是在金属材料疲劳试验中,上述方法不是总有效^[2],因其均以测量数据按对称性概率分布(正态分布、 t 分布)为前提,而金属疲劳试验的测量数据更符合非对称的威布尔分布,低寿命区分散性相对较小、高寿命区分散性相对较大。就金属疲劳寿命试验而言,正常情况下这两类分布模型都可以用于测量数据的统计分析,但是这两类

分布模型主要差异体现在低寿命区和高寿命区,下文以一组试验测量数据为例进行说明。

兹有一组试验中获得的金属材料疲劳寿命数据(单位为载荷循环次数):

164 441, 243 854, 151 520, 510 206, 170 841, 200 795, 150 511, 182 992,

206 555, 210 975, 120 273, 227 187, 748 838, 187 755。现分别采用威布尔分布、正态分布对其进行统计分析,获得概率密度曲线,横坐标为寿命、纵坐标为概率密度。为了清晰说明这两类分布模型主要差异体现在低寿命区和高寿命区的现象,本文分别对横坐标和纵坐标取对数。如图 1(a)、图 1(b)所示,威布尔分布、正态分布在低寿命区和高寿

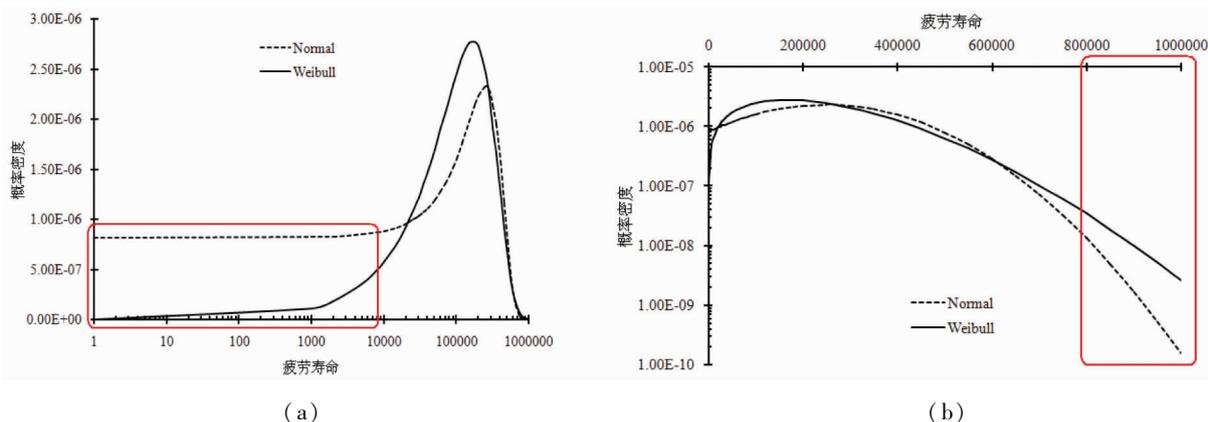


图 1 分别采用威布尔分布和正态分布模型对数据进行统计分析获得的概率密度曲线

命区可能有一个数量级以上的差异。

这两个区域正是统计意义上异常数据的发生区域,因此如果采用以对称性分布为基础的异常数据判别方法,会出现判别不出低寿命区的异常数据、高寿命区又容易犯“弃真”的错误。有文献认为如果测量数据不符合正态分布,应采用非参数检验来判别异常数据,但是一般认为非参数检验没有参数检验可靠,而且通常要求样本量大于 100 甚至更多^[3],然而通常金属疲劳试验中都不会有这么大的样本量。

到目前为止,几乎没有文献针对金属疲劳试验研究基于威布尔分布的异常数据判别方法。本文认为主要原因是,如果从统计的角度对可疑数据进行判别,需要精确地获得总体分布的分布参数,而通常情况下金属疲劳试验测量的数据量是很有限的,依据有限的样本量,现有的统计方法很难比较精确地计算出威布尔分布的分布参数,从而无法较精确地计算出可疑数据的发生概率,即不能判别可疑数据是否为异常数据。

本文将根据对两种常见的计算威布尔分布参数的方法的研究,给出一种基于威布尔分布的异常数据判别方法,更适用于金属材料疲劳试验的异常数据判别。

1 分布参数的两种计算方法

威布尔分布的概率密度函数为:

$$f(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^\alpha} \quad (1)$$

累积概率密度函数为:

$$F(x) = 1 - e^{-1(x/\beta)^\alpha} \quad (2)$$

对于分布参数 α (形状参数) 的确定,通常有两种方法;至于分布参数 β (位置参数) 的确定,则依赖分布参数 α :

1) 头两序法^[4],是一种非参数统计方法。

将一组试验测量数据按照从小到大的顺序排列,即 $x_{1,n} < x_{2,n} < \dots < x_{n,n}$,那么:

$$\begin{cases} \hat{\alpha} = \frac{n \ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{x_{2,n}}{x_{1,n}}\right)} \\ \hat{\beta} = \left(\frac{\sum_{i=1}^n x_i^{\hat{\alpha}}}{n}\right)^{\frac{1}{\hat{\alpha}}} \end{cases} \quad (3)$$

2) 最大似然法^[5-6]。

公式如下:

$$\begin{cases} \frac{\sum_{i=1}^n x_i^{\hat{\alpha}} \ln x_i}{\sum_{i=1}^n x_i^{\hat{\alpha}}} - \frac{1}{n} \sum_{i=1}^n \ln x_i - \frac{1}{\hat{\alpha}} = 0 \\ \hat{\beta} = \left(\frac{\sum_{i=1}^n x_i^{\hat{\alpha}}}{n}\right)^{\frac{1}{\hat{\alpha}}} \end{cases} \quad (4)$$

显然,这两种方法中头两序法更简单,便于快速计算,但是头两序法完全依赖于最小的两个测量数据,如果最小的测量数据为异常数据,那么头两序法将失效。为了避免上述缺点,根据对金属疲劳寿命试验数据的一般认识,本文以距离中位数最近的两个测量数据 ($x_{i,n}, x_{j,n}, x_{i,n} < x_{j,n}$) 为基础推导分布参数 α 的计算公式,推导过程同头两序法,见文献[4],可获得式(5)。

$$\begin{cases} \alpha_{i,j} = \hat{\alpha} = \frac{n \ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{x_{j,n}}{x_{i,n}}\right)} \\ \beta = \hat{\beta} = \left(\frac{\sum_{i=1}^n x_i^{\hat{\alpha}}}{n}\right)^{\frac{1}{\hat{\alpha}}} \end{cases} \quad (5)$$

2 基于威布尔分布的判别方法的介绍

2.1 筛选可疑数据

假设符合威布尔分布的一组测量数据中存在一个可疑数据,分布参数 α 和分布参数 β 对可疑数据的敏感性如图 2 所示,图中可疑数据的对数变化率为可疑数据与分布参数 β 的比值取对数。显然,分布参数 α 对测量数据的变化特别敏感,而分布参数 β 则对测量数据的变化不敏感,即使存在异常数据也不会引起 β 的显著变化。

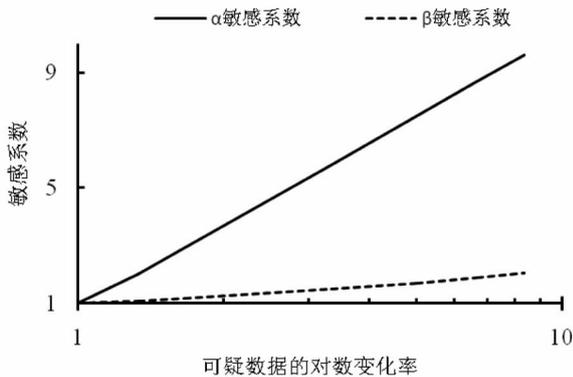


图 2 分布参数 α 和分布参数 β 对可疑数据的敏感性

本文将根据上述特点,研究基于威布尔分布的异常数据判别方法。

首先判断最小测量数据和最大测量数据是否为可疑数据。如本文第 1 节所述,同理可以利用式(5)逐一计算出 $x_{1,n}$ 和 \bar{X} 、 $x_{n,n}$ 和 \bar{X} 、 $x_{1,n}$ 和 $x_{n,n}$ 对应的 α_1 、 α_n 、 $\alpha_{(1,n)}$ 、 \bar{X} 为不含 $x_{1,n}$ 、 $x_{n,n}$ 的均值。

$$\alpha_1 = \frac{n \ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{\bar{X}}{x_{1,n}}\right)}, \alpha_n = \frac{n \ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{x_{n,n}}{\bar{X}}\right)}, \alpha_{1,n} = \frac{n \ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{x_{n,n}}{x_{1,n}}\right)}$$

根据 α_1 、 α_n 、 $\alpha_{(1,n)}$ 的相对关系,找到可疑数据。显然,正常情况下 $\alpha_{(1,n)}$ 与 α_1 、 α_n 存在显著差异,如果 α_1 或 α_n 与 $\alpha_{(1,n)}$ 相近(根据对大量试验数据的观察,本文建议取:若 $\alpha_{(1,n)} < \alpha_1$ 则 $0.6\alpha_1 < \alpha_{(1,n)}$,若

$\alpha_1 < \alpha_{(1,n)}$ 则 $\alpha_{(1,n)} < 1.6\alpha_1$;若 $\alpha_{(1,n)} < \alpha_n$ 则 $0.6\alpha_n < \alpha_{(1,n)}$,若 $\alpha_n < \alpha_{(1,n)}$ 则 $\alpha_{(1,n)} < 1.6\alpha_n$),则认为 α_1 或 α_n 对应的测量数据为可疑数据。

譬如引用参考文献[4]中金属疲劳试验的一组测量数据(单位为载荷循环次数):42 000、45 000、48 000、52 000、55 000、60 000、500 000。该文献表明,其中 500 000 被判定为异常数据。按本方法算得: \bar{X} 、 $\alpha_1 = 5.05$ 、 $\alpha_n = 0.48$ 、 $\alpha_{(1,n)} = 0.44$ 、 $0.6\alpha_n < \alpha_{(1,n)}$,故 α_n 对应的 500 000 为可疑数据。

2.2 判别异常数据

威布尔分布中均值 \bar{X} 与位置参数 β 存在如下关系:

$$\bar{X} = \beta \Gamma\left(1 + \frac{1}{\alpha}\right) \text{ 或 } \beta = \frac{\bar{X}}{\Gamma\left(1 + \frac{1}{\alpha}\right)} \quad (6)$$

因为 β 对测量数据的变化不敏感,本文认为正常情况下由偶然误差引起的测量数据的波动对 β 的影响可以忽略,那么可以作以下近似推导。根据对大量金属疲劳试验数据的分析^[4],测量数据符合 $\alpha > 1$ 者属威布尔分布,当 $\alpha \leq 1$ 时,威布尔分布函数转变为其它分布函数,本文以 $\alpha \leq 1$ 作为异常数据的判据。鉴于总体上铝合金、钛合金、中强钢、高强钢的 α 值分别为 4.0、3.0、3.0、2.2^[7],即 $\Gamma\left(1 + \frac{1}{\alpha}\right)$ 介于 0.89 和 0.91 之间,故近似取为 0.9,则 $\beta = \bar{X}/0.9$,这在工程上是完全可以接受的。

令 $z = \frac{x}{\beta}$,则由式(2)得 $F(z) = 1 - e^{-z^\alpha}$,解得:

$$\alpha = \frac{\ln(-\ln)(1-F)}{\ln(z)} \quad (8)$$

根据对大量金属疲劳试验数据的观察,本文设定:累积发生概率小于 0.000 1 的测量数据为异常数据。对于左侧可疑数据,则 $\alpha = -\frac{9.210290370}{\ln(z)}$;

对于右侧可疑数据,则 $\alpha = \frac{2.220326806}{\ln(z)}$ 。对可疑数据 500 000 不难算得: $z = 9.9338$,继而由式(8)算得 $\alpha = 0.97 < 1$,故可疑数据 500 000 为异常数据。

再以本文提到的另一组试验数据为例:164 441, 243 854, 151 520, 510 206, 170 841, 200 795, 150 511, 182 992, 206 555, 210 975, 120 273, 227 187, 748 838, 187 755。将测量数据按照从小到大的顺序排列,可计算得: $\alpha_1 = 0.57$ 、 $\alpha_n = 1.19$ 、 $\alpha_{(1,n)} = 1.76$ 、 $\alpha_{(1,n)} <$

1.6,故 α_n 对应的 748 838 为可疑数据。对之可算得 $z = 3.2118$,继而算得 $\alpha = 1.90 > 1$,由此判定 748 838 不是异常数据。

3 与其它判别方法的比较

所谓其它判别方法,是指用基于对称性分布的判别方法来判别本文列举的两组数据中是否存在异常数据^[1]。这两组数据均来自试验测量数据,已确认第一组数据中的可疑数据 500 000 为异常数据,第二组数据中的可疑数据 748 838 为非异常数据。

对此两组数据用其它方法判定结果分别见表 1 和表 2。

第一组:42 000、45 000、48 000、52 000、55 000、60 000、500 000。判别过程略,判别结论见表 1。

表 1 算例 1

可疑数据	拉依达法	肖维纳	格拉布斯	狄克逊	罗马诺夫斯基准则
500 000	非异常数据	异常数据	异常数据	非异常数据	非异常数据

第二组:164 441,243 854,151 520,510 206,170 841,200 795,150 511,182 992,206 555,210 975,120 273,227 187,748 838,187 755。判别过程略,判别结论见表 2。

表 2 算例 2

可疑数据	拉依达法	肖维纳	格拉布斯	狄克逊	罗马诺夫斯基准则
748 838	非异常数据	异常数据	异常数据	非异常数据	非异常数据

4 结论

本文引用的两组试验数据中,第一组数据(42 000、45 000、48 000、52 000、55 000、60 000、500 000)是来自波音公司的试验数据^[4],文献[4]已判定 500 000 为异常数据;第二组数据(164 441,243 854,151 520,510 206,170 841,200 795,150 511,182 992,206 555,210 975,120 273,227 187,748 838,187 755)取自某试验结果,经试验人员在试验过程中和试验后对试验件及试验过程记录仔细检查,确认没有出现制造及试验偏离,不

存在异常数据。第 2 节、第 3 节判别结果的比较表明,本文方法对此两组数据的判别结论与它们的原始判定结论完全一致;然而若用基于对称分布的诸方法来判别,分别得到不同的结论,亦即具备不确定性,容易造成误判。可见按本文方法得到的判定结论更加可靠,更适合处理金属疲劳试验的测量数据。

值得指出,本文提供的判别异常数据的方法是基于威布尔非对称分布的非参数统计法,应用了与“头两序法”相同的原理,但作了合理的工程处理,使之有效地避免了“头两序法”高寿命区容易失真和低寿命区容易误判的弊病。

参考文献:

- [1] 唐兆田. 判别金属材料疲劳试验中异常试验数据的一般方法[J]. 民用飞机设计与研究,2017,1:77-82.
- [2] 何平. 剔除测量数据中异常值的若干方法[J]. 航空计测技术,1995,18(1):19.
- [3] National Aeronautics and Space Administration. NASA-HDBK-8739. 19-3. Measurement Uncertainty Analysis Principles and Methods [S]. Washington DC:NASA,2010.
- [4] Whittaker Ian C., Philip M. Besuner. A Reliability Analysis Approach to Fatigue Life Variability of Aircraft Structures [R]. AFML-TR-69-65, Wright-Patterson AFB: Air Force Materials Lab. 1972.
- [5] R. B. Abernethy, J. E. Breneman, C. H. Medlin, G. L. Reinman. Weibull Analysis Handbook [R]. AFWAL-TR-83-2070, West Palm Beach, Florida: Pratt & Whitney Aircraft Government Products Division United Technologies Corporation, 1983.
- [6] Richard L., Hoffert B. S. A Monte Carlo Study of Composite Sequential Likelihood Ratio Tests for the Weibull Scale Parameter [D]. Wright-Patterson Air Force Base, Ohio: Engineering of the Air Force Institute of Technology Air University, 1976.
- [7] 郑晓玲,等. 民用飞机结构耐久性与损伤容限设计手册(疲劳与损伤容限)[M]. 第 1 版. 北京:航空工业出版社,2003:335.

作者简介

唐兆田 男,硕士,高工。主要研究方向:疲劳与损伤容限技术;E-mail: tangzhaotian@comac.cc