解 説

金属学のための極値統計学

井上 毅
Tsuyoshi Inoue
舞鶴工業高等専門学校
自然科学科(数学) 教授

Statistics of Extremes for Metallurgical Application

1 極値統計学とは

一般の統計学は、問題としている母集団の平均的な姿を把握することを主な目的とする。例えば、ある商品が若年層に売れるかという統計調査では、それが平均的な若者に受けるかどうかが最大の関心事であり、ごく一部の偏屈者にはあまり関心がない。一方、極値統計学では平均値はあまり重要でなく、ごく少数の飛びぬけて大きいもの又は小さいもの、言い換えれば少数派偏屈者(これを統計学では極値と呼ぶ)に関心がむけられる。

極値統計学はもともと、洪水の予測を目的とする「水文統計学」として発達したものであり、E. J. Gumbelの有名な著書「Statistics of Extremes」[1]もこの分野への適用を念頭に書かれている。確かに、洪水に強い河川工事をどの程度の規模で行うべきかを考えるとき、毎年の平均降水量などは問題ではなく、百年に一度起こるかどうかという豪雨(極値)が問題となるから、まさに極値統計学の舞台であるといってよい[2]。言いかえれば一般の統計学が平常時の統計学とすれば、極値統計学は異常時に対処するための危機管理の統計学ということもできる。

ところで、金属の破壊は川の氾濫と似たような現象である。なぜなら破壊は金属中に含まれる欠陥が起点となり、ここでも欠陥の平均値が問題ではなく、最大サイズの欠陥、すなわち極値が問題となるからである。金属疲労の極値統計的な解析は、古くはWeibullの研究があり、また最近では村上ら[3]の研究がある。以下に極値統計学の元祖である水文統計学の例を挙げながら、金属学への応用を念頭において、極値統計学の基礎を解説する。

2 極値統計に必要なデータ

統計解析の最初の仕事はデータの収集である。例えば水文統計では、過去30年間にある地域で観測された降水量のデータなどである。このデータは、パラパラと降った小雨まで含めると膨大なデータになるし、データが揃っているとは限らない。しかし、極値統計学ではこのような心配はあまりなく、ある値以上の大雨のデータが揃っていれば十分である。さらに言えば、年間最大降雨量のデータがあれば統計処理は可能になる。

同じことは金属中の介在物についても言える。この場合、データは介在物サイズの測定値に相当するが、全介在物のデータは必要でなく、あるサイズ以上のもののみを測定すればよいし、場合によっては測定視野中の最大サイズ(極値)のデータだけを、できるだけ多数の視野について集めればよいことになる。これは、測定視野内の全介在物を測定する従来の方法に比べれば格段に楽である。このことも極値統計解析の大きな利点といえる。

3 確率分布と非超過確率

極値統計に入る前に、一般の統計学の基礎を簡単に述べておく。統計学における変数は、特に確率変数と呼ばれる。降雨量や介在物サイズなどのデータ値が確率変数にあたる。確率変数をxとしたとき、その分布を表す関数$f(x)$は確率密度関数と呼ばれる。また、確率変数のとりうる範囲も数学的には$-\infty$から$+\infty$まで考えられるが、実際問題では介在物の例からもわかるように、正の範囲を考えればよいことが多い。

極値分布を考えるときに特に重要となるものとして非超過確率がある。これは確率変数xがある値z以下である確率であり、図1に示す確率分布で$x \leq z$の部分の面積に相当し、確率変数の取りうる最小値$x=x_0$から$x=z$まで$f(x)$を積分して次式で求められる。

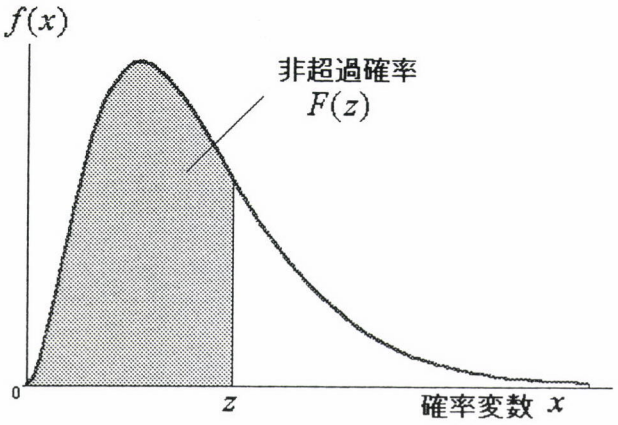

図1　確率分布における非超過確率

$$F(z) = \int_{x_0}^{z} f(x)\,dx \quad \cdots\cdots(1)$$

この$F(z)$は一般の統計学では累積分布関数と呼ばれることが多いが、極値統計学ではその意味を重視して、$x=z$の非超過確率と呼ぶ。さらに、$x=z$を超える確率(超過確率)は、当然確率の性質から$1-F(z)$である。

4 母集団の分布

極値統計によりデータを処理するとき、もとの生データ(母集団)の分布が重要となる。そこでよく使われる母集団の分布について若干触れておく。それぞれの確率密度関数については付録として巻末にまとめて示す。

4.1 正規分布および対数正規分布

正規分布は、一般の社会現象や自然現象での母集団分布として最もよく用いられるものである。また、対数正規分布は確率変数の代わりにその対数$\log x$をとったときに正規分布に従うものである。実際の応用例では、降雨量などは対数正規分布に近いことが多く、岩井[4]らの詳しい研究がある。また粉末冶金の粉体粒度分布でも確率変数がけた違いに大きい範囲におよぶため、対数正規分布が用いられることが多い[5]。

4.2 指数分布およびアーラン分布

介在物のサイズ分布などを考えると、微小なのものはきわめて多数あるが、サイズが大きいものは急激にその数が少なくなる。このような分布を近似するものとして指数分布がある。指数分布を仮定する最大の利点は数学的な取り扱いが楽であり、確率密度関数の積分も容易なため、非超過確率も簡単に計算できることにある。しかし、介在物の例を考えても明らかのように、大きさ0の介在物は実在しないにもかかわらず、指数分布では$x=0$でも一定の値を持ち、現実とは異なる。ただし、微小介在物は問題にせず、大きいサイズの介在物だけを考えようとする極値統計では指数分布で近似しても十分なことが多い。

アーラン分布は指数分布をより一般化したものである。指数分布は、付録(A6)式で$k=1$としたアーラン分布の特別な場合である。また、図1に示した分布は$k=3$として計算したアーラン分布の例である。

なお、介在物の極値統計解析などでは、多くの場合指数分布に従うものとして行われているが、実際に対象としている母集団の分布が本当に指数分布で近似してよいのかという問題に対しては詳しい研究は少なく、今後の課題である。

5 極値分布

つぎに、最大値に関する極値分布について考える。いま問題としている母集団を、確率密度関数$f(x)$をもつ分布であるとし、その母集団から大きさnの標本をN個とるとする。これは、介在物の顕微鏡観察の例でいえば、介在物がn個含まれるような視野を1視野として、これをN視野観察してデータをとることにあたる。この場合、各視野の最大値(極値)ばかりを集めた$\{z_1, z_2, \ldots, z_N\}$の集合は、もとの母集団とは異なる新しい分布(極値分布)をもつ。そこで、zを確率変数とするこの極値分布の確率密度関数$p(z)$と非超過確率$P(z)$を求めることを考える。

いま、介在物のもとの確率密度関数を$f(x)$とすると、$x=z$を超えない非超過確率は(1)式で示した$F(z)$である。i番目の視野中の最大サイズをz_iとすれば、その視野中のn個の介在物はすべてサイズz_iを超えないのであるから、この視野の代表値z_iの非超過確率は$[F(z_i)]^n$であると考えてよい。各視野ともほぼ等しいn個中の最大値であるなら、これが極値分布の非超過確率$P(z)$を与えるから、

$$P(z) = [F(z)]^n \quad \cdots\cdots(2)$$

よって、極値分布の確率密度関数$p(z)$はこの$P(z)$をzで微分して次式となる。

$$p(z)=\frac{dP(z)}{dz}=n[F(z)]^{n-1}\frac{dF(z)}{dz}=n[F(z)]^{n-1}f(z) \quad \cdots\cdots(3)$$

いま、nが十分に大きいとし、また極値として最大値を考えているから$F(z)\approx 1$としてよい。対数のテイラー展開

$\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$において$(1+x)=F(z)$

とおくと、$\log F(z)=(F(z)-1)-\frac{1}{2}(F(z)-1)^2+\cdots$

となり、右辺第2項以下を無視すると $\log F(z)\cong F(z)-1$ となる。よって(2)式より次式と書くことができ

$$\log P(z)=n\log F(z)\cong n[F(z)-1]=-n[1-F(z)]\cdots(4)$$

結局、最大値の極値分布では、非超過確率は近似的に次式で表せる。

$$P(z)\cong\exp\{-n[1-F(z)]\}\cdots(5)$$

ここで、$F(z)$は母集団の非超過確率、nは各標本の大きさである。

6 Return Period(再現周期)

つぎに、極値統計で重要なReturn Period(再現周期)について考える。これはある極値がどのくらいの周期で出現するかを示すものである。例えば、ある地方で200mm/hという猛烈な雨が100年に1度の割で起こるとすると、200mm/h降雨の再現周期Tは100年であるという。また、介在物の例でいえば、ある大きさの巨大介在物が1000視野に一度の割で見られるとすれば、このサイズの介在物の再現周期Tは1000視野であるといえる。この再現周期とそのときの極値がわかれば、100年に一度の豪雨にも耐えうる治水工事が可能になるし、巨大介在物による破壊を未然に防止することも出来る。極値統計解析の目的は、再現周期やその極値を推定することにあるといっても過言ではない。

いま介在物データが視野中の最大サイズとして取られているときの例を考えよう。極値データの分布は前節で示した非超過確率$P(z)$をもつから、ある値zを超える確率は$1-P(z)$である。よってz以上の極値が出現するのがT視野に一度とすると、

$$T=\frac{1}{1-P(z)}\cdots(6)$$

となり、これが再現周期である。また、逆に再現周期から$P(z)$を求めたいときは次式を使えばよい。

$$P(z)=1-\frac{1}{T}\cdots(7)$$

7 順序統計量とHazen分率、Thomas分率

次に実際の観測データの極値統計解析方法に入り、非超過確率やReturn Period(再現周期)を求める方法を述べる。いま、ほぼn個の介在物を含む視野を1視野とし、その中の最大の介在物サイズを測定する。これをN視野について行うと、N個の極値の集合が得られる。このデータから極値分布、すなわちその非超過確率$P(z)$を求めようというわけである。いま、N個からなる極値データの集合を、求める極値分布から得られた一標本と考え、N個のデータの中で、ある極値z_iより小さいものの割合(分率)がどのくらいあるかを見積もれば、それが$P(z_i)$を与える。この見積りにはいくつかの方法が提案されているが、主なものとしてHazenの方法とThomasの方法を紹介する。

7.1 Hazenの方法[6]

N個のデータを小さい順(昇順)に最小値(z_1)、2番目(z_2)、…、最大値(z_N)と並び替える。このように、昇順(または降順)に並べ替えたデータを順序統計量と呼ぶ。このとき同じ値があっても連続番号をつけて、最後がN番目となるようにする。このようにしておくと、N個のうちどれか1つが起こる確率は、すべて同じ$1/N$となる。このことは図2に示すように、N個の順序統計量の生起確率が同じになるように、z_iを中心に面積の等しいN本の柱を立てたことになる。確率分布の性質から全面積は1であるから、各柱の面積は$1/N$である。順序統計量において、i番目の統計量z_iの非超過確率は、この値の左側にある面積の総和となるから、$(i-1)$本の柱の面積$\frac{i-1}{N}$とi番目の柱の面積の半分$\frac{1}{2N}$を加えたものになる。すなわちz_iの非超過確率$P(z_i)$は次式で見積もれる。

$$P(z_i)=\frac{i-1}{N}+\frac{1}{2N}=\frac{2i-1}{2N}\cdots(8)$$

なお、Hazenの方法は、両端域での精度が悪くなる難点があり[7]、極値統計では多くの場合、つぎに示すThomasの方法が用いられている。

7.2 Thomasの方法[8]

Hazenの方法では、i番目の順序統計量z_iを中心としたN本の柱を考えたが、Thomasの方法ではz_iを側面とする

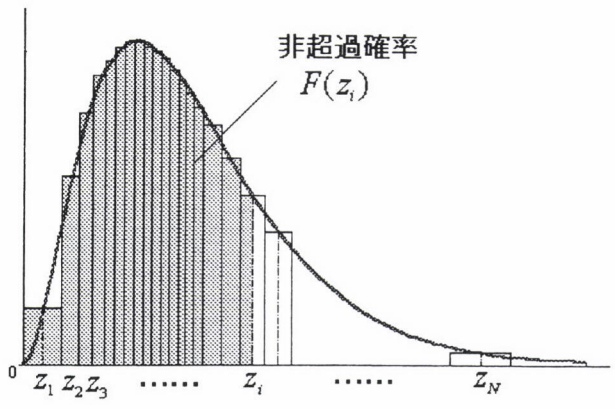

図2 Hazen法による順序統計量の分率の求め方

柱を考える。いまz_iを右の側面とすると、右端に柱のない部分ができ、誤差が生じる。そこで、もう1本柱を加える。このとき、柱の数は$N+1$本となるから、各柱の面積は同じ$1/(N+1)$である。よって、N個のデータを昇順にならべたi番目のデータz_iの非超過確率$P(z_i)$は、i番目の柱を含めてその左側にあるi本の柱の面積であり、次式となる。

$$P(z_i)=\frac{i}{N+1} \quad \cdots\cdots(9)$$

なお、z_iを左側面としたときも同様に考えて、$P(z_i)$を求めるとまったく同じ解が得られるから、(9)式はこの両者の平均であると考えてもよい。

8 極値確率紙(Gumbel確率紙)とGumbelプロット

母集団を指数分布と仮定して、そこから得られた極値分布を解析するために考案された確率紙を極値確率紙、または開発者の名をとってGumbel確率紙と呼ばれている。

Gumbel確率紙では、縦軸に非超過確率$P(z)$を下に示す(13)式の尺度で目盛られており、データを直接プロットできるように工夫されている。しかし、この確率紙は一般に入手困難であり、自分で作るにも労力がいる。そこで、以下ではGumbel確率紙を使わないデータのプロットの方法について解説する。なお、以下に用いる対数はすべて自然対数とする。

平均$1/\lambda$、分散$1/\lambda^2$の指数分布の確率密度関数は付録の(A4)式、累積分布関数は(A5)式で表せる。よって、指数分布をもつ母集団から大きさnの標本をとり、その最大値を集めた極値分布の非超過確率は、(5)式に(A5)式を代入して次式で近似できる。

$$P(z)\approx\exp\{-n[1-F(z)]\}=\exp(-ne^{-\lambda z}) \quad \cdots(10)$$

またこのときの極値分布の確率密度関数は次式となる。

$$p(z)=dP(z)/\mathrm{d}z=n\lambda e^{-\lambda z}\exp(-ne^{-\lambda z}) \quad \cdots\cdots(11)$$

ここで、非超過確率$P(z)$と確率変数zの関係は(10)式の両辺の自然対数を2回とって

$$-\log\{-\log P(z)\}=\lambda z-\log n \quad \cdots\cdots(12)$$

が得られる。この式の左辺をGと置いて、これをGumbel値と呼ぶことにすると、

$$G=-\log\{-\log P(z)\} \quad \cdots\cdots(13)$$

となる。以上をまとめると、Gumbel値Gと極値zの関係は次式の直線関係が得られる。

$$G=\lambda z-\log n \quad \cdots\cdots(14)$$

なお、極値分布におけるi番目の極値z_iの非超過確率をThomas分率で求めるとするとGumbel値は次式で計算すればよい。

$$G_i=-\log\left\{-\log\left(\frac{i}{N+1}\right)\right\} \quad \cdots\cdots(15)$$

以上のことから、母集団として指数分布を仮定したときの極値分布は、横軸に確率変数z(降雨量や介在物サイズの極値データ)を、縦軸にGumbel値Gをとってプロットすると直線関係となり、その傾きがλを与える。なお、Gの計算は表計算ソフトなどを用いると簡単であり、最近では優れたグラフソフトがあるのでグラフ化も容易である。

最後に、実際のデータ解析時の注意点を述べる。標本ごとにnが大きく異なるデータを使った場合、解析誤差の原因になる。これは(14)式からも分かるように、データ点が$\log n$だけ左右に振れるためである。このため、nが大きく変わるデータを用いる場合には、各標本ごとのnの値も測定し、(14)式を用い$\log n$だけ補正する必要がある。

9 Return Period(再現周期)の推定例

9.1 降雨データの例

ある地方の73年間の年間最大降雨データ[7]を解析した結果を表1に、またGumbelプロットした結果を図3に示す。この結果をもとに、Return Periodとして50年、100年、500年の降雨量を推定してみよう。非超過確率とReturn Periodの関係を示す(7)式から、$T=50$年に対しては$P(z)=0.98$(非超過確率98%)、100、500年に対しては、それぞれ0.99、

表1　73年間の年間最大降雨量の極値解析結果

順序 (i)	降雨量(z_i)	Thomas分率	Gumbel値
1	56	0.0135	−1.460
2	57	0.0270	−1.284
3	59	0.0405	−1.165
4	60	0.0541	−1.071
5	61	0.0676	−0.991
6	64	0.0811	−0.921
7	66	0.0946	−0.858
8	68	0.1081	−0.800
9	72	0.1216	−0.745
10	73	0.1351	−0.694
11	73	0.1486	−0.645
12	74	0.1622	−0.598
13	76	0.1757	−0.553
14	76	0.1892	−0.510
⋮ 中略	⋮	⋮	⋮
66	161	0.8919	2.168
67	163	0.9054	2.309
68	167	0.9189	2.470
69	175	0.9324	2.660
70	176	0.9459	2.890
71	177	0.9595	3.185
72	184	0.9730	3.597
73	240	0.9865	4.297

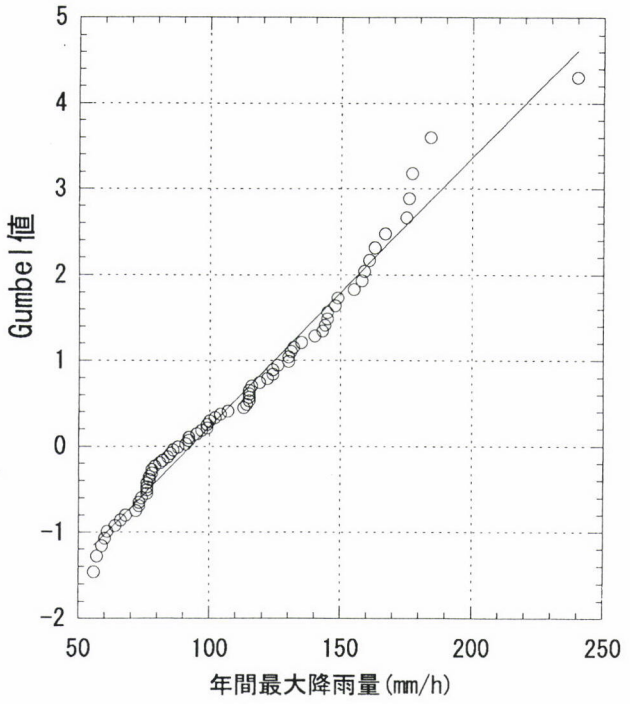

図3　73年間の年間最大降雨量の極値プロット例

0.998となる。Gumbel値は(13)式からそれぞれ、3.90、4.60、6.21となり、それに対応する横軸の値を読み取ればよい。図解的な方法では精度が不十分と考えるときには、最小二乗法を使って回帰直線式を求めて計算してもよい。実際に図3の結果から最小二乗法で求めた回帰式は$z=31.9G+92.8$となるから、結果はReturn Period 50年に対しては時間降雨量217mm/h、100年に対しては240mm/h、500年に対しては291mm/hと求まる。すなわち、この地方では240mm/h程度の雨が100年に一度、また290mm/h程度の豪雨が500年に一度程度は起こると推定される。

9.2　介在物データの例

鋼中の非金属介在物の解析については村上の著書[3]に詳しいので、ここではデータと解析結果のみを例示する。表2は各視野の最大サイズを40視野について測定した結果である(この場合サイズは介在物の面積の平方根$\sqrt{area}$(μm)で表示されている)。この結果をGumbelプロットした結果を図4に示す。このときの回帰式は$z=2.41G+16.95$であるから、Return Period $T=100$視野、500視野に対するGumbel値4.60、6.21を入れると、それぞれ$z=28.0$、31.9と求まる。すなわち、この鋼では500視野相当の領域内には32μm程度の介在物が存在すると考えなければならない。実際の疲労の解析には、部材への応力のかかり方から問題とすべき領域がどの程度かを見積もり、それに対するReturn Periodを決めれば、その領域内にある最大の介在物サイズの予測が出来ることになる。介在物の測定法および極値解析法、また疲労解析への適用については前述の著書に詳しいのでここでは割愛する。

表2　40視野の視野中最大介在物の極値解析結果

順序 (i)	サイズ(z_i)	Thomas分率	Gumbel値
1	13.82	0.024	−1.312
2	14.50	0.049	−1.105
3	14.69	0.073	−0.961
4	14.97	0.098	−0.845
5	15.15	0.122	−0.744
6	15.51	0.146	−0.653
7	15.78	0.171	−0.570
8	16.13	0.195	−0.491
9	16.21	0.220	−0.416
10	16.38	0.244	−0.344
11	16.47	0.268	−0.274
12	16.63	0.293	−0.206
13	16.88	0.317	−0.139
14	16.88	0.341	−0.072
⋮ 中略	⋮	⋮	⋮
33	20.24	0.805	1.528
34	20.78	0.829	1.676
35	21.50	0.854	1.844
36	21.88	0.878	2.040
37	22.87	0.902	2.276
38	23.29	0.927	2.577
39	24.22	0.951	2.996
40	26.51	0.976	3.701

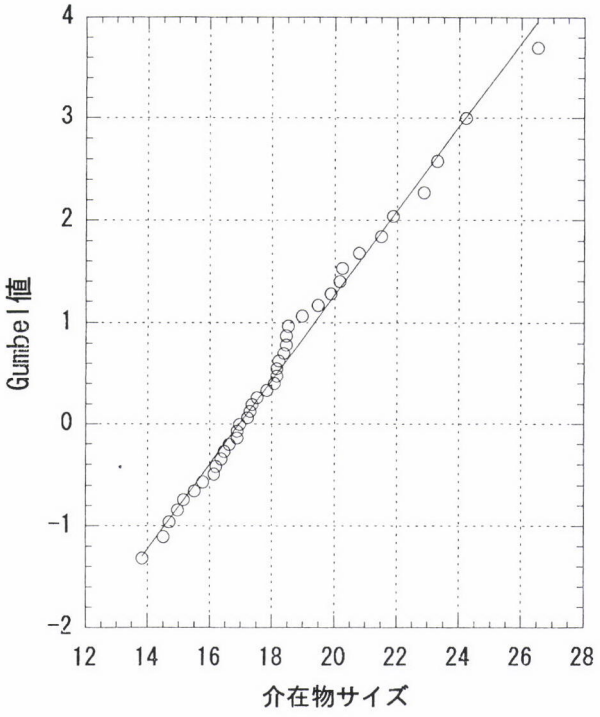

図4　40視野の視野中最大介在物の極値プロット例

10 乱数によるシミュレーション実験

今まで述べた極値分布の考え方を検証するため、多数の乱数を発生させることによるシミュレーション実験を行った。シミュレーション実験はMicrosoft社のExcelの分析ツールとVBA(Visual Basic)モジュールを併用し、グラフ作成にはSynergy Software社のKaleida Graphを用いた。

10.1 母集団が正規分布のときの極値分布

乱数は平均値$\mu=10$、標準偏差$\sigma=5$の正規分布$N(10,5^2)$に従う、各組30個の正規乱数を1000組、合計30000個($n=30$、$N=1000$)発生させた。まず各組の30個中の最大値をz_iとし、それらからなる極値集合$M\{z_i|i=1,2,3,\cdots,1000\}$を作成した。図5はその極値分布と、前述の(3)式から求めた理論極値分布を比較して示す。乱数実験の結果はほぼ理論通りの分布を示し、(3)式は極値分布の確率密度関数を、また(2)式は極値分布の非超過確率を示す式として適当であることが分かる。

10.2 母集団が指数分布のときの極値分布

今回は、平均値$1/\lambda=10$、標準偏差$1/\lambda=10$の指数分布に従う指数乱数を上と同様$n=30$、$N=1000$、合計30000個発生させた。同様に各組の最大値z_iを極値とする極値分布と理論分布を図6に示す。母集団が指数分布のときは、確率密度関数の数学的取り扱いが容易であり、理論値は直接(11)式から計算できる。この場合も乱数実験の結果はほぼ理論通りの分布を示し、(11)式は極値分布の確率密度関数として適当であるといえる。

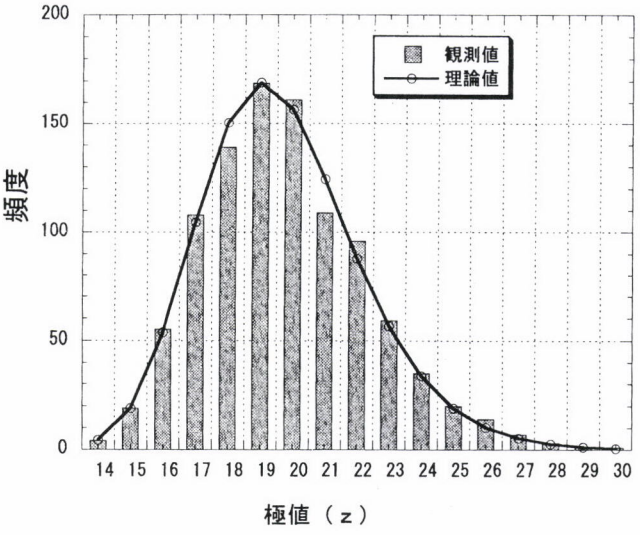

図5 平均値$\mu=10$、標準偏差$\sigma=5$の正規乱数からの極値分布のシミュレーション結果と理論値の比較

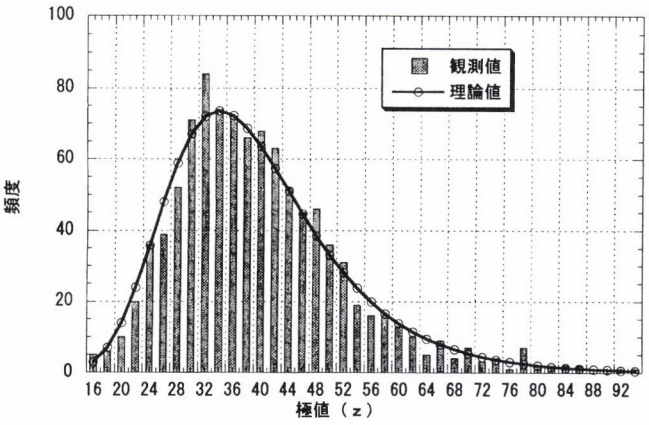

図6 平均値$1/\lambda=10$の指数乱数からの極値分布のシミュレーション結果と理論値の比較

11 おわりに

金属の破壊現象などは、内在する最大欠陥寸法に支配されることはよく知られている。このような現象を取り扱う場合、通常の統計学ではなく極値統計学が役にたつ。極値統計学は土木・建設分野の「水文統計学」や信頼性工学の分野では古くから使われることが多く、これらの分野では優れた教科書[1,7,9]があるが、金属学の分野では適当な解説が少ない。本稿では、極値統計学の基礎的な考え方とその解析法を主として水文統計学の例を引きながら、金属学への応用を念頭において解説した。金属学でのデータ解析の一助となれば幸いである。

(付録)

主要分布の確率密度関数

(1)正規分布

平均値μ、標準偏差σをもつ正規分布$N(\mu,\sigma^2)$の確率密度関数は次式で示される。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \cdots\cdots(A1)$$

標準正規分布$N(0,1^2)$は上式で$\mu=0$、$\sigma=1$であるから、簡単に

$$f(x)=0.399e^{-\frac{x^2}{2}} \quad \cdots\cdots(A2)$$

と書ける。

(2)対数正規分布

対数正規分布では正規分布の確率変数xを$\log x$と置きかえるから、確率密度関数は$f(\log x)$となり、確率関数の性質からそれぞれ区間dxと区間$d(\log x)$の面積(生起確率)

は変わってはならないので$f(x)\,dx=f(\log x)\,d(\log x)$である。いま、$d(\log x)/dx=1/x$を考慮すると、対数正規分布の確率密度関数は次式で表される。

$$f(x)=\frac{1}{\sqrt{2\pi}\delta}\cdot\frac{1}{x}\exp\left\{-\frac{(\log x-m)^2}{2\delta^2}\right\} \quad \cdots\cdots(A3)$$

ここで、m,δはそれぞれ確率変数を$\log x$としたときの母平均と母標準偏差である。

(3)指数分布

平均値$1/\lambda$、分散$1/\lambda^2$の指数分布の確率密度関数は、確率変数$x\geq 0$に対して

$$\begin{aligned} f(x)&=\lambda e^{-\lambda x} \quad (x>0)\\ &=0 \qquad (x=0) \end{aligned} \quad \cdots\cdots(A4)$$

で表され、累積分布関数は解析的に簡単に求めることができ、次式となる。

$$F(z)=\int_0^z f(x)\,dx=1-e^{-\lambda z} \quad \cdots\cdots(A5)$$

(4)アーラン分布

アーラン分布は、$\lambda>0, k$を正の整数としたとき、確率変数$x(\geq 0)$に対して確率密度関数は次式で示される。

$$f(x)=\frac{\lambda^k}{(k-1)!}e^{-\lambda x}x^{k-1} \quad \cdots\cdots(A6)$$

引用文献

1) E. J. Gumbel：Statistics of Extremes, Columbia University Press, NY, (1958)
2) E. J. Gumbel：The Return Period of Flood Flows, Ann. Math. Statistics, Vol. XXII, 2 (1941)
3) 村上敬宣：金属疲労－微小欠陥と介在物の影響，養賢堂，(1993)
4) 岩井重久：土木学会論文集，1，2 (1946)
5) 水渡英二：粉体－理論と応用，丸善，(1962)，78.
6) A. Hazen：Flood Flows, Wiley, NY, (1930)
7) 岩井重久，石黒政儀：応用水文統計学，森北出版，(1970)
8) H. A. Thomas：Trans AGU, (1938)
9) 信頼性工学入門，真壁肇編，日本規格協会，(1991)

(1999年1月18日受付)