II-309　二つのピークを持つ度数分布の統計的処理

日本大学工学部　正　員　安　田　禎　輔

まえがき　異なった原因によって生起する事象の度数分布曲線は，複数個のピークを持つ可能性がある。たとえば降雨量はその発生原因である台風，前線，季節風などによって明瞭に差が生じ度数分布曲線にピークが幾つか生ずることがある。このように，水文量の度数分布はときたま二つ以上のピークを持つが，目下その統計的処理方法が見当らない。本報では，二つのピークを有する場合の標本値の統計的処理方法を，正規分布する場合と対数正規分布する場合とについて考える。それには，与えられた大きさnの標本（これを標本nとする）を原因別に二つの標本n_1とn_2に篩い分けることができればよい。超過確率を求める場合には，標本nを原因別に二つの標本n_1とn_2に分け，これらより超過確率を求め加重平均すれば，篩い分けずに求めた超過確率より正確である。

1.1　二つの正規分布の混合

二つのピークを持つ大きさnの標本

$$\{x_i\}=\{x_1,x_2,x_3,\cdots\cdots,x_n\}\cdots\cdots(1)$$

を原因別になんらかの方法で大きさ$n_1 n_2$の二つの標本

$$\{x_{1i}\}=\{x_{11},x_{12},x_{13},\cdots\cdots,x_{1n_1}\}\cdots\cdots(2)$$

$$\{x_{2j}\}=\{x_{21},x_{22},x_{23},\cdots\cdots,x_{2n_2}\}\cdots\cdots(3)$$

に分離できたものとすると

$$n=n_1+n_2$$

$$\{x_i\}=\{x_{1i}\}+\{x_{2j}\}$$

となり，各標本の平均値はそれぞれ次式となる。

$$\bar{x}_1=\frac{1}{n_1}\sum_1^{n_1}x_{1i}\quad,\quad \bar{x}_2=\frac{1}{n_2}\sum_1^{n_2}x_{2j}$$

$$\bar{x}=\frac{1}{n}\sum_1^{n}x_i=\frac{1}{n}\left\{\sum_1^{n_1}x_{1i}+\sum_1^{n_2}x_{2j}\right\}=\frac{1}{n}\{n_1\bar{x}_1+n_2\bar{x}_2\}$$

$$\bar{x}=\alpha_1\bar{x}_1+\alpha_2\bar{x}_2$$

ただし，$\alpha_1=n_1/n$，$\alpha_2=n_2/n$　$\therefore\alpha_1+\alpha_2=1$ ……(4)

標準偏差はそれぞれ

$$\sigma_1=\sqrt{\frac{1}{n_1}\sum_1^{n_1}(x_{1i}-\bar{x}_1)^2}\quad,\quad\sigma_2=\sqrt{\frac{1}{n_2}\sum_1^{n_2}(x_{2j}-\bar{x}_2)^2}$$

となり，σはつぎのようにして求められる。

$$\bar{x}_1<\bar{x}_2\quad,\quad\Delta\bar{x}=\bar{x}_2-\bar{x}_1$$

とおけば

$$\bar{x}=\bar{x}_1+\alpha_2\cdot\Delta\bar{x}$$

$$\bar{x}=\bar{x}_2-\alpha_1\cdot\Delta\bar{x}$$

したがって

$$\begin{aligned}n\sigma^2&=\sum_1^{n}(x_i-\bar{x})^2\\&=\sum_1^{n_1}(x_{1i}-\bar{x})^2+\sum_1^{n_2}(x_{2j}-\bar{x})^2\\&=\sum_1^{n_1}\{(x_{1i}-\bar{x}_1)-\alpha_2\Delta\bar{x}\}^2+\sum_1^{n_2}\{(x_{2j}-\bar{x}_2)+\alpha_1\Delta\bar{x}\}^2\\&=n_1\sigma_1^2+n_2\sigma_2^2+n\cdot\alpha_1\cdot\alpha_2\cdot(\Delta\bar{x})^2\end{aligned}$$

$$\therefore\quad\sigma^2=\alpha_1\sigma_1^2+\alpha_2\sigma_2^2+\alpha_1\cdot\alpha_2\cdot(\Delta\bar{x})^2\cdots\cdots(5)$$

また，確率密度関数はそれぞれ

$$f_1(x)=\frac{1}{\sqrt{2\pi}\,\sigma_1}e^{-\frac{(x-\bar{x}_1)^2}{2\sigma_1^2}}$$

$$f_2(x)=\frac{1}{\sqrt{2\pi}\,\sigma_2}e^{-\frac{(x-\bar{x}_2)^2}{2\sigma_2^2}}$$

となり，$f(x)$はつぎのようにして求められる。

確率変数xに対する標本nの度数$nf(x)$は，標本n_1およびn_2の度数$n_1f_1(x)$および$n_2f_2(x)$の和であるから

$$nf(x)=n_1f_1(x)+n_2f_2(x)$$

$$\therefore\quad f(x)=\alpha_1\cdot f_1(x)+\alpha_2\cdot f_2(x)\cdots\cdots(6)$$

となる。標本nの超過確率$W(x)$は，標本n_1および標本n_2の超過確率$W_1(x)$,$W_2(x)$の加重和であり

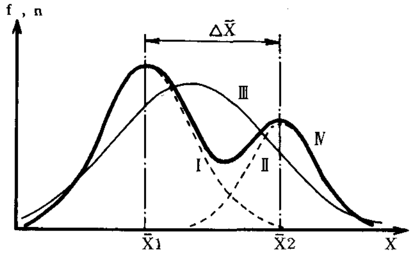
$$W(x)=\alpha_1W_1(x)+\alpha_2W_2(x)\cdots\cdots(7)$$

となる。$W_1(x)$,$W_2(x)$は正規分布の超過確率であり正規分布の数表より計算できる。

1.2　二つの正規分布の混合標本の分離法

標本nを原因別に二つの標本n_1およびn_2とに分離する手順を箇条書きに示す。

① 度数分布曲線または確率密度曲線図-1を描き，ピークの生じる位置を近似的に$\bar{x}_1$，$\bar{x}_2$として，$\Delta\bar{x}=\bar{x}_2-\bar{x}_1$を求める。



図－1

② $\bar{x}=\sum x_i/n$ を計算し，次式によりα_1,α_2を求める。

$$\begin{cases}\bar{x}=\alpha_1\bar{x}_1+\alpha_2\bar{x}_2\\\alpha_1+\alpha_2=1\end{cases}$$

③ $\sigma = \sqrt{\{\Sigma(x_i - \bar{x})^2\}/n}$ を計算する。

④ $\bar{x}_1$および$\bar{x}_2$の外側で，ピーク付近の任意の点xの$f(x)$を図-1より読み取り

$$\begin{cases} f(x) = \alpha_1 \cdot f_1(x) + \alpha_2 \cdot f_2(x) \\ \sigma^2 = \alpha_1 \sigma_1^2 + \alpha_2 \sigma_2^2 + \alpha_1 \cdot \alpha_2 \cdot (\Delta \bar{x})^2 \end{cases}$$

よりσ_1およびσ_2を求める。

⑤ ①～④で求めた$\bar{x}_1$, $\bar{x}_2$およびσ_1, σ_2を用い$f(x)$を計算し，これと①の$f(x)$と比較し，一致するまで①～⑤を繰り返す。

2.1 対数正規分布

対数正規分布は確率変数xを $z=\log_{10} x$ によりzに変換すれば正規分布となる分布である。

$$g(x) = \frac{\log_{10} e}{\sqrt{2\pi}\,\sigma_0} \cdot \frac{1}{x} \cdot e^{-\frac{1}{2\sigma_0^2}\left(\log_{10}\frac{x}{x_0}\right)^2} \quad \cdots\cdots (8)$$

は，置換 $z = \log_{10} x$ により

$$g(x)dx = \frac{1}{\sqrt{2\pi}\,\sigma_0} e^{-\frac{(z-\bar{z})^2}{2\sigma_0^2}} dz = f(z) \cdot dz$$

となる。

$$\therefore f(z) = \frac{1}{\sqrt{2\pi}\,\sigma_0} e^{-\frac{(z-\bar{z})^2}{2\sigma_0^2}} \quad \cdots\cdots\cdots (9)$$

となり，$f(z)$は正規分布関数となるから，$g(x)$は対数正規分布関数である。

ただし，$\bar{z} = \overline{\log_{10} x} = \log_{10} x_0$ $\cdots\cdots\cdots\cdots$ (10)

$\sigma_0 = \sqrt{\frac{1}{n}\Sigma(\log_{10} x_i - \log_{10} x_0)^2}$ $\cdots\cdots\cdots$ (11)

で，x_0は分布$g(x)$の中央値であり，σ_0は正規分布$f(z)$の標準偏差である。

つぎに，分布$g(x)$の最頻値x_mを求めると

$$x_m = x_0 \cdot 10^{-\sigma_0^2/\log_{10} e} \quad \cdots\cdots\cdots (12)$$

となり

$$g(x_m) = \frac{\log_{10} e}{\sqrt{2\pi}\,\sigma_0} \, \frac{10^{(\sigma_0^2/\log_{10} e)}}{x_0} \cdot e^{-\frac{1}{2}(\sigma_0/\log_{10} e)^2}$$

となる。

2.2 二つの対数正規分布の混合標本の分離法

分離手順を箇条書きに示す。

① 対数正規分布曲線図-2より，近似的に最頻値x_{m1}, x_{m2}を読み取る。x_{m2}はやや精度が悪い。

② 正規分布曲線図-3より，$\bar{z}_1$, $\bar{z}_2$を読み取り，中央値x_{01}, x_{02}を求める。x_{01}は精度が悪い。したがって$\Delta\bar{z}$も精度が悪い。

③ $\bar{z}$, σ_0を計算する。

④ x_{m1}, x_{02}および式(12)より，$\sigma_{02} = \sqrt{\log_{10} e \cdot \log_{10}(x_{02}/x_{m2})}$ を計算する。

⑤ x_{m1}, x_{01}よりσ_{01}を求め，これを第1近似値として以下の計算を繰り返す。

⑥ σ_0, σ_{01}, σ_{02}, $\Delta\bar{z}$の値と次式

$$\begin{cases} \alpha_1 + \alpha_2 = 1 \\ \sigma_0^2 = \alpha_1 \cdot \sigma_{01}^2 + \alpha_2 \cdot \sigma_{02}^2 + \alpha_1 \cdot \alpha_2 \cdot (\Delta\bar{z})^2 \end{cases}$$

より，α_1, α_2を求める。

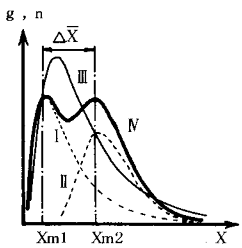
⑦ x_{m1} 付近の点xの$g(x)$を読み取り，これと計算値

$$g(x) = \alpha_1 \cdot g_1(x) + \alpha_2 \cdot g_2(x)$$

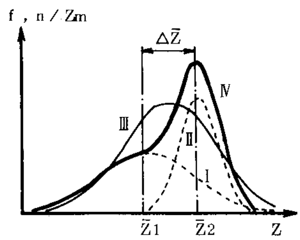
を比較し，一致しなければ，σ_{01}を仮定しなおし，⑥に帰り一致するまで繰り返す。

⑧ 分布曲線$g(x)$を描き，①の分布曲線と比較する。一致しなければ①に帰りx_{m2}②でx_{01}を修正し一致するまで①～⑧を繰り返す。以上のようにして，標本nをn_1, n_2に分離して，確率・超過確率および非超過確率などを計算すれば分離しない場合より精度が高まる。

ここで，Z_m：隣合う階級代表値xの対数変換値の差



図－2



図－3

3. 資料の個数の少ない場合の処理法

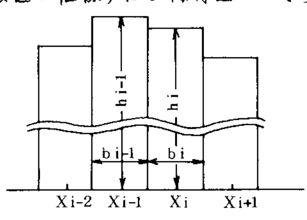
標本の個数が極めて少ない場合には，これらを階級別に分けて度数分布表やヒストグラムを作成することは困難となる。このような場合には，標本値をx軸上に大きさの順に並べ，各標本値x_iを中心とする底辺b_iの面積$1/n$の柱状図を作成し，これより，確率密度曲線を描くことができる。これらの柱状図の幅b_i，高さh_iは図-4より

$$l_i = \frac{b_{i-1} + b_i}{2}$$

$$\left.\begin{aligned} \therefore b_i &= 2l_i - b_{i-1} \\ \text{or } b_{i-1} &= 2l_i - b_i \end{aligned}\right\} \quad \cdots\cdots (13)$$

$$h_i = 1/n b_i$$

となる。



図－4

計算方法は，最初$b_i = b_{i-1}$として最頻値付近より左右に計算を進める。かくして柱状図が描けたら，前節1.2または2.2により処理する。