データの統計量 (Statistics)

1 データとヒストグラム (Histogram)

大量のデータが与えられたとき，そのデータの状況を見た目で判断しやすくするためには，ヒストグラムを用いると良い．まず N 個のデータが以下のように変数 x で与えられていたとする．

$$(x_1, x_2, \cdots, x_N) \tag{1}$$

これを階級 (Rank) ごとのデータとして整理する．例えば，データの値が $a_0 \sim a_1$ の範囲に何個のデータがあるかカウントする．カウントされたデータ数は，度数 (Frequency) あるいは頻度と呼ばれている．ここで階級は，以下のように n 個の階級で区切られたとする．

$$(a_0 \sim a_1, a_1 \sim a_2, \cdots, a_{n-1} \sim a_n) \tag{2}$$

また各階級における度数は，変数 f を用いて表すと以下のようになる．

$$(f_1, f_2, \cdots, f_n) \tag{3}$$

そして，階級を横軸，度数を縦軸に棒グラフで表したものがヒストグラムと呼ばれている．データの分布状況を捉えるのに便利なグラフである．ここで，$N = f_1 + f_2 + \cdots + f_n$ となる．

種類の異なるデータを度数を基にしたヒストグラムを用いて比較する場合，データ数 N が大きく異なるデータ同士を比較するには，相対度数 (Relative Frequency) を用いてヒストグラム化するのが良い．相対度数は，以下のように各度数をデータ数で除すことで，相対度数の合計が 1 となり，比較が容易になる．この相対度数に 100 を乗ずれば，パーセントの単位となる．

$$\left(\frac{f_1}{N}, \frac{f_2}{N}, \cdots, \frac{f_n}{N}\right) \tag{4}$$

2 平均 (Average)，メディアン (Median)，モード (Mode)

平均計算は，測量データの整理においても極めて重要である．平均値 $\bar{x}$ は，以下の式で計算することができる．

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

ヒストグラムで与えられたデータのみから平均値を計算するには，各階級に対応する代表値と度数を用いて計算する．各階級に対応する代表値が $(m_1, m_2, \cdots, m_n)$ で与えられたとすると，以下の式により計算できる．

$$\bar{x} = \frac{\sum_{i=1}^{n} m_i f_i}{N} \tag{6}$$

メディアンは，中央値と呼ばれる．データ x_i を小さい順に並び替え， $N/2$ 番目のデータが中央値となる．N が奇数の場合は，$N/2$ の値が半端になるので，前後の値を平均化することで求める．ヒストグラムで与えられたデータのメディアンは，階級の中央値となる．階級の刻み幅が大きすぎるヒストグラムの場合は，データから導かれるメディアンとの差が大きく生じてしまい，あまり意味を持たない場合がある．

モードは，最も頻度の大きい階級の代表値である．したがって，モードはヒストグラムで表されていなければ導かれない．モードの場合，階級の刻み幅が小さすぎると，余り意味を持たないときがある．

3 分散 (Variance) と標準偏差 (Standard Deviation)

データのばらつきを判定するのに，分散や標準偏差を用いることが多い．平均値から離れたデータがどれだけ存在するかを判断できる．したがって，各データと平均値との差を基に計算する．単に各データと平均値の差の値（偏差という）を合計すると 0 になるので，偏差の二乗を合計し，データ数で除したものが分散 S^2 である．

$$S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \tag{7}$$

ヒストグラムで与えられたデータのみから分散を計算するには，以下の式となる．

$$S^2 = \frac{\sum_{i=1}^{n} (m_i - \bar{x})^2 f_i}{N} \tag{8}$$

標準偏差は，分散の平方根のことを言い，ここではとりあえず二乗の単位だったのを元に戻して分かりやい値にしたと思って構わない．分散・標準偏差が大きいデータはばらつきの大きいデータと言え，小さいデータはばらつきの小さいデータと言える．同じものを測って測量したデータから標準偏差を計算したとき，標準偏差の小さいデータは精度の高いデータと言える．

4 歪度 (Skew)

歪度は，非対称度とも呼ばれ，ヒストグラムの分布の形が，平均より右よりか左よりかを判定するのに使われる．平均より右か左かを判定するために偏差の符号が重要となる．したがって，偏差の三乗を用いて以下の式により計算される．

$$S_s = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{NS^3} \tag{9}$$

この値が 0 に近いほど左右対称の分布と言え，正の値のときは右寄り，負の値のときは左寄りの分布と判定できる．

5 尖度 (Kurtosis)

尖度は，分布のとがり具合を判定することができる．この場合，偏差の四乗を用いて以下の式により計算される．

$$S_k = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{NS^4} \tag{10}$$

この値が 3 に近いほど次節の正規分布に近く，3 より大きいと尖った分布，3 より小さいとなだらかな分布と判定できる．

6 二項分布

ある 2 つの事柄（事象）の起る確率を考えるとき，一方の起る確率を p とすると，他方の起る確率は $1-p$ となる．例えばコインを投げて，表が出る確率が $\frac{1}{2}$ とすると，裏が出る確率は $\frac{1}{2}$ となる．そしてコインを 5 回投げて，5 回とも表の出る確率は，$(\frac{1}{2})^5$ と簡単に計算できる．次にコインを 5 回投げて 1 回だけ表の出る確率は，1 回だけ表の出る組み合わせは 5 通りあるので，$5 \times (\frac{1}{2})^5$ となり，5 回投げて 2 回表の出る確率は，${}_5\mathrm{C}_2$ 通り表の出る組み合わせがあるので，${}_5\mathrm{C}_2 \times (\frac{1}{2})^5$ となる．この場合，表も裏も確率 $\frac{1}{2}$ であるが，表と裏とで確率が異なる場合，n 回投げて r 回の表が出る確率 $f(r)$ は，以下の式で計算することができる．

$$f(r) = {}_n\mathrm{C}_r p^r (1-p)^{(n-r)} \tag{11}$$

これを**二項分布**と呼んでいる．

7 正規分布 (Normal Distribution)

正規分布 (Normal Distribution) は，ランダムな誤差を持つデータの分布を関数で表したもので，平均値を μ，標準偏差を σ としたとき，確率密度関数 $f(x)$ は，以下の式で与えられる．ガウス分布とも呼ばれている．

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{12}$$

なぜ，このような複雑な式が正規分布を表すのか，ここで簡単に述べておくが，詳細は他の書籍を参考にしてほしい．まず左右対称で $x = 0$ のときにピークになるような関数を考えると，$f(x) = Ce^{-h^2x^2}$ が当てはまる，18 世紀にド・モアブル (De Moivre) が二項分布の極限として導き，その後ガウス (Gauss) やハーゲン (Hagen) が言及している．

さて，誤差の三公理は，以下のとおりである．正規分布は，この公理に沿った関数である必要がある．

1. 絶対値の小さい誤差の生じる確率は，大きい誤差の生じる確率よりも大きい．
2. 絶対値の等しい正負の誤差は，同じ確率で生じる．
3. 絶対値の非常に大きい誤差は，ほとんど生じない．

第二公理に従えば，二項分布において，ある事象の起る確率 p は，$\frac{1}{2}$ となる．そして誤差量 x は，無限小の誤差原子 ϵ からなるものと仮定する．つまり誤差量の大きい誤差は，たくさんの誤差原子から成り立っているといえる．したがって，誤差が n 回分全て正の誤差原子 ϵ により構成されて発生したとすると，その誤差量 x は $n\epsilon$ であり，その確率 y は $(\frac{1}{2})^n$ となる．よって，n が十分大きくなれば，第三公理に従い，非常に大きい誤差はほとんど生じないことになる．

n 回のうち，r 回が負の誤差原子より誤差が構成されているとすると，その誤差量 x は以下の式で計算できる．

$$\begin{aligned} x &= (n-r)\epsilon - r\epsilon \\ &= (n-2r)\epsilon \end{aligned} \tag{13}$$

その確率 y は二項分布より以下の式で計算できる．

$$y = {}_n\mathrm{C}_r \left(\frac{1}{2}\right)^n = \frac{n!}{r!(n-r)!}\left(\frac{1}{2}\right)^n \tag{14}$$

したがって，第一公理の条件も満たされている．

次に，微積分を用いて正規分布を誘導するため，$r+1$ 回が負の誤差原子より誤差が構成されているとすると，その誤差量 x_1 は以下の式となる．

$$\begin{aligned} x_1 &= (n-r-1)\epsilon - (r+1)\epsilon \\ &= (n-2r-2)\epsilon \end{aligned} \tag{15}$$

その確率 y_1 は二項分布より以下の式となる．

$$y_1 = {}_n\mathrm{C}_{r+1} \left(\frac{1}{2}\right)^n = \frac{n!}{(r+1)!(n-r-1)!}\left(\frac{1}{2}\right)^n \tag{16}$$

となる．したがって，誤差量の変化 Δx は，以下の式で計算できる．

$$\begin{aligned} \Delta x &= (n-2r)\epsilon - (n-2r-2)\epsilon \\ &= 2\epsilon \end{aligned} \tag{17}$$

一方，確率の変化 Δy は，以下の式で計算できる．

$$\begin{aligned}
\Delta y &= {}_n\mathrm{C}_r \left(\frac{1}{2}\right)^n - {}_n\mathrm{C}_{r+1} \left(\frac{1}{2}\right)^n \\
&= \left(\frac{n(n-1)\cdots(n-r+1)}{r!} - \frac{n(n-1)\cdots(n-r)}{(r+1)!}\right)\left(\frac{1}{2}\right)^n \\
&= \left(\frac{n(n-1)\cdots(n-r+1)}{r!} - \frac{n(n-1)\cdots(n-r)/(r+1)}{(r+1)!/(r+1)}\right)\left(\frac{1}{2}\right)^n \\
&= \left(1 - \frac{n-r}{r+1}\right)\left(\frac{n(n-1)\cdots(n-r+1)}{r!}\right)\left(\frac{1}{2}\right)^n \\
&= \left(\frac{2r-n+r}{r+1}\right){}_n\mathrm{C}_r \left(\frac{1}{2}\right)^n \\
&= \left(\frac{2r-n+r}{r+1}\right) y \\
&= \left(\frac{-2x+2\epsilon}{(n+2)\epsilon - x}\right) y \qquad \text{式 13 より，} r = \frac{n\epsilon - x}{2\epsilon} \text{ を代入} \\
&\approx -\frac{2xy}{n\epsilon} \qquad n \text{ は十分大きく } \epsilon \text{ は十分小さいことを考慮する}
\end{aligned} \tag{18}$$

したがって，次式を得る．

$$\frac{\Delta y}{\Delta x} = -\frac{2xy}{2n\epsilon^2} \tag{19}$$

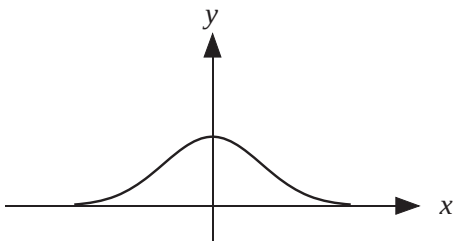
ここで，n は無限大へ，ϵ は 0 に限りなく近づいたとき，式 19 の分母 $2n\epsilon^2$ が $\frac{1}{h^2}$ に近づくとすると，次式が成り立つ．

$$\begin{aligned}
\frac{dy}{dx} &= -2h^2xy \\
\frac{dy}{y} &= -2h^2x dx
\end{aligned} \tag{20}$$

次に両辺を積分する

$$\begin{aligned}
\int \frac{1}{y} dy &= \int -2h^2x dx \\
\ln y &= -h^2x^2 + C \\
y &= C\mathrm{e}^{-h^2x^2}
\end{aligned} \tag{21}$$

$f(x) = e^{-x^2}$ 式のグラフを描くと，下図のようになる．



確かに左右対称の偶関数で中心 $x = 0$ においてピークがあり，中心極限定理も満たしている．

さて，この関数における定数 C, h を求めなければならない．正規分布は，確率密度関数なので，$-\infty$ から $-\infty$ まで積分した値が 1 となる必要がある．すると，$C = \frac{h}{\sqrt{\pi}}$ が導かれ，次式を得る．

$$y = \frac{h}{\sqrt{\pi}} \mathrm{e}^{-h^2x^2} \tag{22}$$

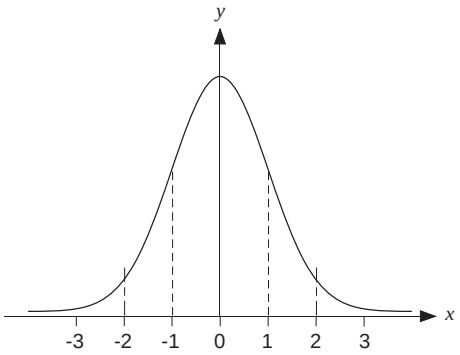
次にこの式の二階微分を計算する．

$$\begin{aligned} \frac{dy}{dx} &= -\frac{2h^3}{\sqrt{\pi}} \mathrm{e}^{-h^2x^2} x \\ \frac{d^2y}{dx^2} &= -\frac{2h^3}{\sqrt{\pi}} \mathrm{e}^{-h^2x^2} + \frac{4h^5}{\sqrt{\pi}} \mathrm{e}^{-h^2x^2} x^2 \\ &= -\frac{2h^3}{\sqrt{\pi}} \mathrm{e}^{-h^2x^2} (1 - 2h^2x^2) \end{aligned} \tag{23}$$

したがって，この関数は $1 - 2h^2x^2 = 0$ を満たす x において変曲点を持つ．つまり $x = \pm\frac{1}{\sqrt{2}h}$ において変曲点が存在する．この変曲点が標準偏差 σ に相当する．したがって，標準偏差 $\sigma = \frac{1}{\sqrt{2}h}$ より，$h = \frac{1}{\sqrt{2}\sigma}$ を式 22 に代入すれば，次式が得られる．

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x}{\sigma})^2} \tag{24}$$

これは，平均値が 0 の時の式なので，平均値 μ を考慮すると，式 12 が導かれる．下図は平均値が 0，標準偏差が 1 の時の正規分布のグラフである．



上図の正規分布の面積全体において，標準偏差 $-\sigma$〜$+\sigma$ の範囲が占める割合は約 0.6827 で，標準偏差 -2σ〜$+2\sigma$ の範囲が占める割合は約 0.9545，標準偏差 -3σ〜$+3\sigma$ の範囲が占める割合は約 0.9973 である．計測機器の精度を標準偏差で表しているものもあるが，その機器で測った場合，標準偏差を越える誤差で測ってしまう確率は，1-0.6827，つまり 3 割程度の確率で発生するので，注意が必要である．したがって誤差は，標準偏差の 2 倍，3 倍を見積もったうえで測らなければならない．

測量データの誤差は，次節で述べるように過失誤差・系統誤差・偶然誤差に分類されるが，偶然誤差は正規分布に従うものである．つまり，同じものを繰り返し計測して得られたデータの分布が正規分布にならない場合は，偶然誤差以外の要因が含まれると推察される．