正規分布

1　誤差の分布

前回の資料で示したように，測定には必ず偶然誤差が伴う．この誤差については経験上以下のようなことがわかっている．

- 同じ大きさの正および負の誤差は同じ確率で起こる．
- 値の小さな誤差は大きな誤差よりも高い頻度で起こる．
- ある程度以上の誤差は起こらない．

18 世紀から 19 世紀にかけて活躍したドイツの科学者ガウスはこの条件を公理として以下のような数式化を行った．測定値を x，真の値を m，そして，母標準偏差を σ とすると，

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

であらわされる関数 $f(x)$ で表される確率密度で誤差の分布が表現できる．ここで，

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{2}$$

となるように係数が規格化されている．この関数の概形は図 1 のようになる．この曲線で表される分布を**正規分布**と呼ぶが，導出者の名前を取ってガウス分布と呼ぶことも多い．

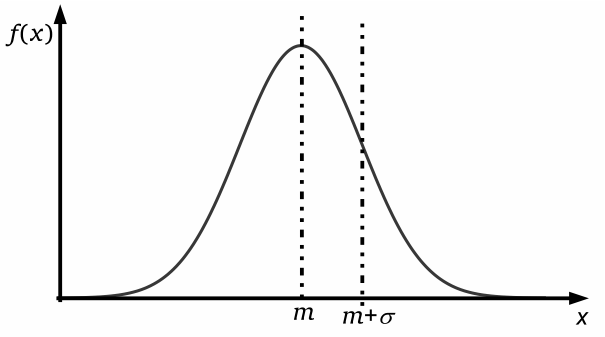

図 1: 正規分布曲線の概形

2　正規分布

前述のように，正規分布曲線の形は平均値（真の値）m と母分散 σ^2 が与えられると決まるので，記号として $N(m, \sigma^2)$ を用いる．また，正規分布においては以下の関係がある．

- $m \pm \sigma$ の区間に測定値の 68.3%が含まれる．
- $m \pm 2\sigma$ の区間に測定値の 95.4%が含まれる．
- $m \pm 3\sigma$ の区間に測定値の 99.7%が含まれる．

正規分布の形が平均値と分散で決まるとしても，異なる測定系の議論でそれぞれ値が異なることは不便である．そこで，式 (1) において変数変換を行う．変数 u として，

$$u = \frac{x - m}{\sigma} \tag{3}$$

を用いると，式 (1) は以下のように書き換えられる．

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{4}$$

これは $m = 0$，$\sigma^2 = 1$ の場合の正規分布 $N(0, 1)$ に相当する．この分布を**標準正規分布**と呼ぶ．

正規分布曲線は，前述のように確率密度を表す曲線である．確率密度ということは，実際の確率は関数を積分して求める必要がある．例えば，測定値が x と $x + \Delta x$ の範囲内に含まれる確率は，

$$\int_x^{x+\Delta x} f(x)dx = \int_x^{x+\Delta x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \tag{5}$$

で与えられるということである．しかしながら，この関数は解析的には積分できない．そこで，昔から標準正規分布について数値表が作成されている．本授業では Calc の関数を使用して値を求めることも可能であるが，必要に応じて表を参照することもあるかもしれない．

表に載せてある数値は $-\infty$ から u までに測定値が含まれる確率を表している．なので，ちょうど中心の $u = 0$ のときが $1/2$ の確率であるので 0.5 となり，標準偏差 σ が 1 であるので，1.0 のときが 0.8413 となる．$1 - 0.8413 = 0.1587$ であるが，それは値が標準偏差以上に外れる確率であり，その 2 倍は 0.3174 であるから，先ほど既述したように，測定値が $\pm\sigma$ の間に入る確率は $1 - 0.3174 = 0.6826$ となり，約 68.3%となるわけである．

なお，最近では表計算ソフトにより正規分布に従う場合の確率の値は簡単に計算できるようになっている．`normdist` もしくは `norm.dist` という関数を用いて，変数の値，平均値，標準偏差，論理値の順に引数を入れることで計算できる．論理値は TRUE としておけば，値が表示される．また，論理値を FALSE とした場合には曲線の高さの値が出力される．

後述の統計的仮説検定などで用いられるが，データを統計的に議論するときにはこの正規分布の両方のすその辺り（確率的に 5%未満となる領域）に値が含まれることはあまりないこととされている．そこで，すそにあたる部分の確率を求めると，

$$\int_{-\infty}^{-u_\alpha} \phi(u)du + \int_{u_\alpha}^{\infty} \phi(u)du \equiv \alpha \tag{6}$$

のように表現できる．ここで α は測定値が平均値から u_α 以上離れている確率を表し，図 2 において斜線を引いた部分になる．この α を危険率と呼び，$\alpha < .05$ となる場合に統計的にあまりない事象と慣習的に判断している．

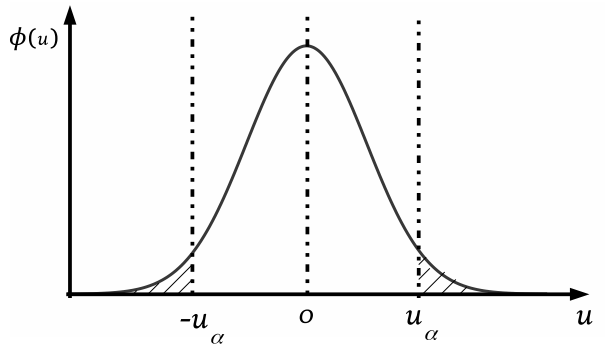

図 2: 標準正規分布 $N(0, 1)$ における危険率 α の考え方

3　パーセンタイル値

標準正規分布表において，記載されている数値はその u の値よりも小さいデータの全体の量に対する割合に相当する．例えば，$u = 0.21$ のときの値は 0.5832 であるが，それは $u \leq 0.21$ の値は全体の 58.32%を占めるということである．ということは，測定値の組において，その測定値が正規分布に従うときには，その測定値が下から数えて何%の位置に来るのかを平均値と標準偏差を用いて計算できるということである．この下位何%に位置するのかを示す値を**パーセンタイル値**と呼ぶ．

4　正規分布の性質

今ある量について母平均を m，母分散を σ^2 とするとき，n 個の標本 x_1，x_2，$\cdots$，x_n の標本平均を $\bar{x}$ を求めた場合，その $\bar{x}$ の母平均 $m_{\bar{x}}$ および母分散 $\sigma_{\bar{x}}^2$ は次のようになることが知られている．

$$m_{\bar{x}} = m, \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

しかし，より重要なことは，元の x がどんな分布であっても標本平均 $\bar{x}$ の分布は正規分布 $N(m, \sigma^2)$ に従う，ということである．これを**中心極限定理**という．そのため，次の変量を使用すると，その分布は n が大きくなるにつれてしだいに標準正規分布 $N(0, 1)$ に近づく．

$$\frac{\bar{x} - m}{\sigma / \sqrt{n}}$$

具体的な例を考える．仮に図 3 のような確率密度であらわされる分布があるとする．このとき，ランダムに 10000 個ほど標本の候補となる値を発生させて母集団を作る．その中からランダムに 20 個の標本を取り出す作業を 100 回行い，その度数分布を調べる作業を行ってみる．プログラミングの知識があれば，簡単にできる作業である．まず，元の確率密度の式は，指数分布として以下のようだとする．

$$p(x) = \exp(-x)$$

変数 x は 0 から 5 までとしているので，この範囲で積分するとほぼ 1 となり，そのまま確率分布に対応しているとしてもかまわない．この確率に従って変数を発生させるプログラムは，指数関数の逆関数である対数関数により 0.01 から 1 までのランダムに発生させた乱数の値 y の対数，すなわち，$x = -\log y$ を 10000 回発生させて求めたものを配列に保存すればよい．実際に行った例をヒストグラム形式にしたものを図 4 に示す．大体良い形になっていると思われる．(実際には e^{-5} は約 0.006 であるので，y の値を 0.01 以上としたために $x > 2.5$ では度数が無いことになってしまっているが，あまり大きな影響はないと思われる．)

その配列の中からランダムに 20 個のデータを取り，算術平均を求める操作を 100 回繰り返して平均値の分布を作成する．実際に試した例を図 5 に示す．データの個数が有限であるのできれいな形にはなっていないが，データ個数を増やしていくと徐々に正規分布に近づいていくことが予想されるであろう．

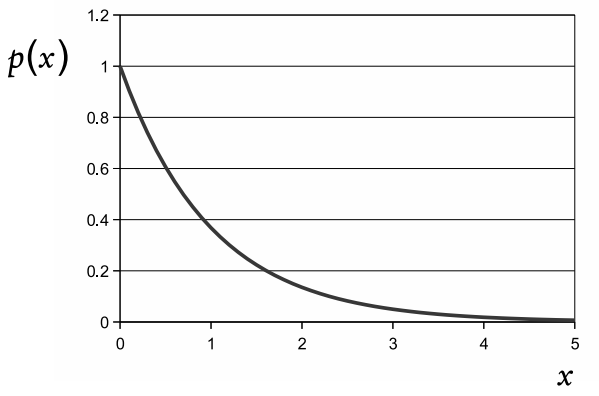

図 3: 指数分布的な分布の例

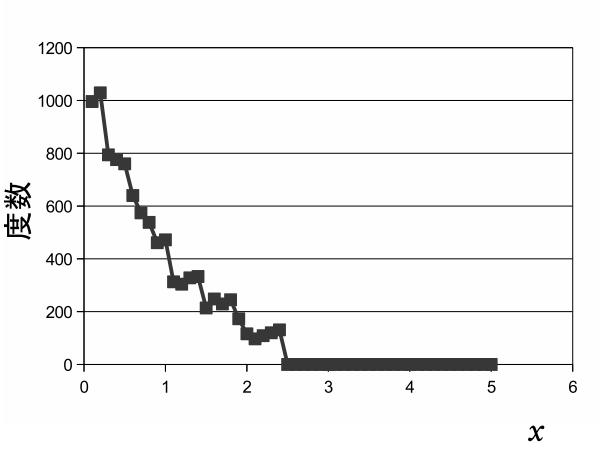

図 4: 乱数により発生させた値の度数分布

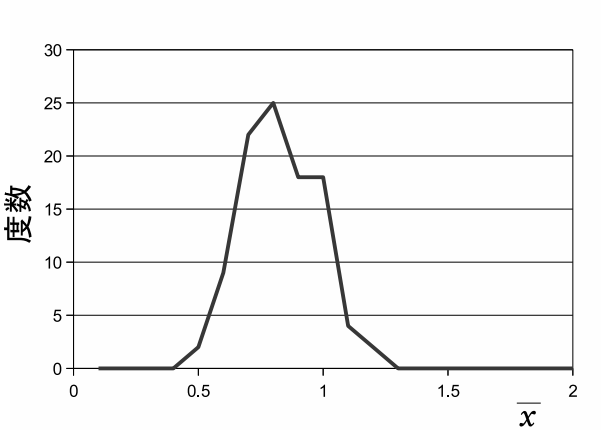

図 5: 指数分布のデータの平均値の度数分布