2008 年度後期　情報統計学　第６回

連続型確率分布と正規分布

ヒストグラム

度数分布や確率分布を図示するために，横軸に階級，縦軸に度数（相対度数）をとり，階級幅を底辺，度数を面積とする柱で各階級の度数を表したグラフを**ヒストグラム**といいます．

ヒストグラムは，図 1 のような棒グラフではなく，図 2 のように，隣りどうしの柱と柱をくっつけて描きます．また，ヒストグラムでは，柱の高さではなく，柱の**面積**で度数を表現しているためです．ですから，ヒストグラムの縦軸はとくに意味はなく，柱の幅が変われば，同じ高さでも表す確率は異なります．

こういう表しかたをするのは，ヒストグラムの横軸は，棒グラフのようにとびとびにはなっておらず，連続した値を表しているからです．すなわち，「階級」や「階級値」は，元々とびとびだった訳ではなく，本来もっと細かいさまざまな数値をとることのできる値（身長や，点数など）を，階級に分けるときの都合でいくつかの段階に分けたものだからです．

また，度数や確率を柱の面積で表しておくと，となりどうしの柱を結合したり分割することができます．図 3 のように，「170 から 175cm の度数」を表す柱と「175 から 180cm の度数」を表す柱をくっつけるだけで，「170 から 175cm または 175 から 180cm の度数」すなわち「170 から 180cm の度数」を表すことができます．このとき，くっつけた柱をならしてしまうと，「170 から 175cm の度数」や「175 から 180cm の度数」はわからなくなりますが，それでも「170 から 180cm の度数」は表されています．逆に，「170 から 175cm の度数」や「175 から 180cm の度数」がわかれば，柱を分割するだけでそれぞれの度数を表すことができます．

連続型確率分布

第 3 回の講義で，度数分布→確率分布→確率変数という順に進めてきた説明では，連続した数値の区間である階級を，ひとつの数値で代表する「階級値」というものをわざわざ導入することで，確率変数は 172.5cm のつぎは 177.5cm というように「とびとび」の値をとる，と考えてきました．

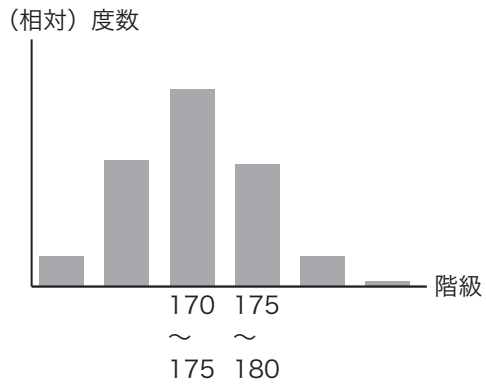

図 1: ヒストグラムはこんなふうには描かない

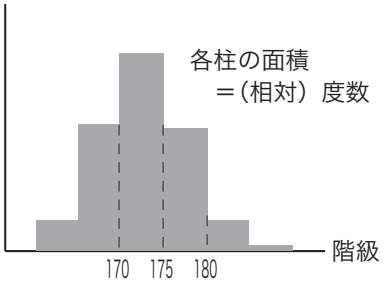

図 2: ヒストグラムはこう描く

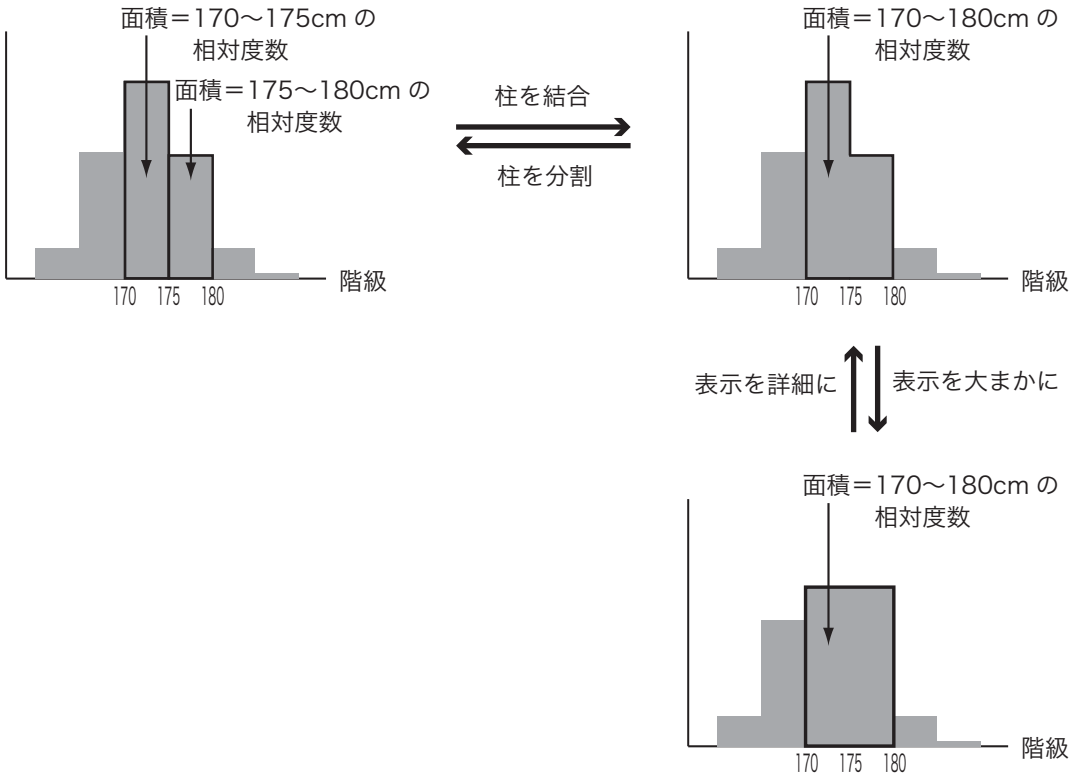

図 3: 確率を柱の面積で描く理由

数学では，とびとびの値をとる数式よりも，連続なグラフになるような数式のほうがずっと簡単なはずです．では，本来連続した数値だったデータを，もっと素直に表現するため，「とびとびではなく連続的な値をとる」確率変数というのは考えられないのでしょうか．

そこで，前節で説明した，「ヒストグラムの柱は，分割することができる」という性質を使います．ヒストグラムの各柱をどんどん細かく分割することで，階級の区切りかたを「十分に」細かくしたとします．このような確率分布は，値がとびとびにならない，「ある範囲内のどんな値にでもなることができる」確率分布と考えることができます（図 4）．このような確率分布を**連続型確率分布**といい，これに対し，２項分布のように，確率変数が（階級に区切ったのではなく）本来とびとびの値（例えば，当たり回数）になるような確率分布を**離散型確率分布**といいます．

連続型確率分布では，確率変数が「ある 1 つの値」をとる確率ではなく，「ある範囲の値」をとる確率を考えます．離散型確率分布で確率変数が「ある範囲の値」をとる確率は，確率変数のある範囲内の値に対応する確率を合計したものです．ヒストグラム上でこれを見ると，ある範囲内にある「柱」の面積を合計したものになります（図 4 の左）．「ヒストグラムで度数を表しているのは柱の高さではなく柱の面積」であるからです．

これを，階級の区切りが見えないほど細かくなったヒストグラムで考えると，柱の境目は見えなくなっているので，灰色の部分の面積がそれに相当します（図 4 の右）．この面積は，数学では「『ヒストグラムの上端をつないだグラフで表される関数』の『ある範囲』での積分」といいます．この「ヒストグラムの上端をつないだグラフで表される関数」を**確率密度関数**といいます．つまり，「連続型確率変数 X が a

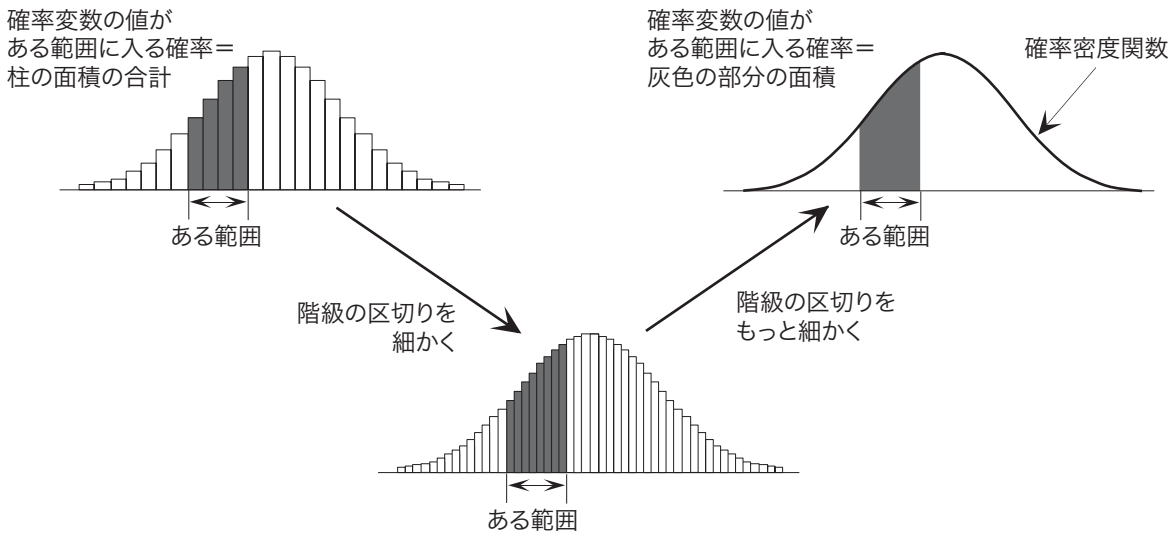

図 4: 連続型確率分布

から b の範囲の値をとる確率」すなわち $P(a \le X \le b)$ は，X の確率密度関数を $f(x)$ とするとき

$$P(a \le X \le b) = \int_a^b f(x)dx \tag{1}$$

となります．

「確率変数がある範囲の値に入る確率」
＝「確率密度関数のグラフの下の部分のうち，この範囲にあたる部分の面積」
＝「確率密度関数のこの範囲での積分」

という関係は，今後の講義でよく出てきますので，よく理解してください．

確率密度関数は確率変数がとりうる各値の「現れやすさ」を表してはいますが，確率そのものではないことに注意してください．「連続型確率変数がある 1 つの値をとる確率」は，確率密度関数の値ではありません．「連続型確率変数がある 1 つの値をとる確率」は，範囲の幅が 0 ですからその範囲に対応するグラフの下の部分の面積も 0 で，すなわち 0 であることに注意しましょう．また，グラフの下の部分全体の面積は，「確率変数の値が，とりうる値の範囲全体のどこかにある確率」ですから 1 (100%) となります．

また，ここまで離散型確率分布を使って説明した期待値や分散，モーメント母関数などは，「確率（頻度関数）の和」を「確率密度関数の積分」に置き換えれば連続型確率分布にも適用できます．例えば，確率変数 X が離散型で，X がある値 x をとる確率 $P(X = x)$ が頻度関数 $f(x)$ で表されるとき，期待値 $E(X)$ は

$$E(X) = \sum_x xf(x) \tag{2}$$

で表されますが，同様に，確率変数 X が連続型でその確率密度関数が $f(x)$ であるとき，期待値 $E(X)$ は

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{3}$$

で表されます．

なお，連続型確率変数 X の確率密度関数が $f(x)$ のとき，「X が x 以下である確率」すなわち $P(X \leq x)$ を $F(x)$ であらわすと，(1) 式から

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{4}$$

となります．この $F(x)$ を**累積分布関数**とよびます．(4) 式から，$F'(x) = f(x)$ となることがわかります．すなわち，確率密度関数と累積分布関数は互いに微分・積分の関係になっています．

ところで，さきほど「本来連続した数値だったデータ」と述べましたが，現実のデータは，必ず何桁かの数字で表されるわけですから，どんなに細かく表現してても必ず「デジタル」，すなわち「とびとび（離散的）」です．連続型確率分布というのは，あくまで数式で表しやすくするための手段だと考えてください．

正規分布モデルの計算

代表的な連続型確率分布モデルである**正規分布モデル**は一番応用範囲の広い確率分布モデルで，世の中には正規分布モデルであらわせるような母集団がたくさんあります．これは，中心極限定理という定理があるからです．中心極限定理とは，簡単に言うと「母集団のデータが分布している（ばらついている）原因が，無数の独立な原因の合計になっているときは，母集団分布は概ね正規分布になる」ということです．くわしくは，第 12 回の講義で説明します．

正規分布の確率密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{5}$$

で表され，そのパラメータは期待値（相対度数分布でいえば平均）μ と分散 σ^2 の２つです．ある確率変数 X の確率分布が期待値 μ，分散 σ^2 の正規分布であることを，「確率変数 X は正規分布 $N(\mu, \sigma^2)$ にしたがう」あるいはさらに短く「$X \sim N(\mu, \sigma^2)$」と書きます．

正規分布の確率密度関数のグラフは図 5 のようになります．期待値 μ をとる確率密度がいちばん高く，左右対称に広がっています．

正規分布には，次の大変重要な性質があります[1]．

1. a, b を定数とするとき，確率変数 X が正規分布 $N(\mu, \sigma^2)$ にしたがうならば，$aX + b$ は正規分布 $N(a\mu + b, a^2\sigma^2)$ にしたがいます．

   このとき，$a = 1/\sigma, b = -\mu/\sigma$ とおくと，$a\mu + b = 0, a^2\sigma^2 = 1$ となるので，**X が正規分布 $N(\mu, \sigma^2)$ にしたがうとき，$(X - \mu)/\sigma$ は正規分布 $N(0, 1)$ にしたがう**ことがわかります．この $N(0, 1)$ を**標準正規分布**といいます．

2. 確率変数 X, Y が独立で，いずれも正規分布にしたがうならば，その和はやはり正規分布にしたがいます．これを**正規分布の再生性**といいます．とくに，$X_1, ..., X_n$ が独立で，いずれも正規分布 $N(\mu, \sigma^2)$ にしたがうならば，それらの平均 $(X_1 + ... + X_n)/n$ は $N(\mu, \sigma^2/n)$ にしたがいます．

[1] 証明は付録を見てください．

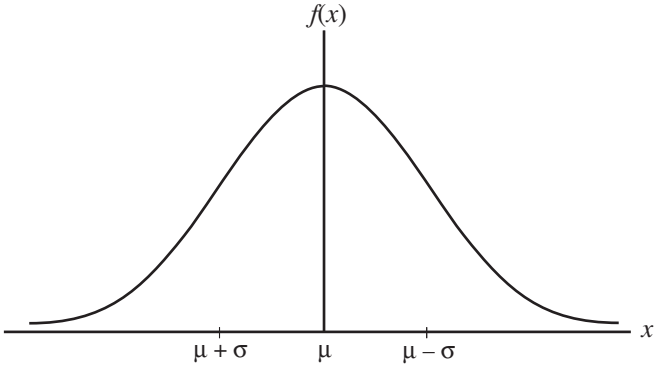

図 5: 正規分布の確率密度関数

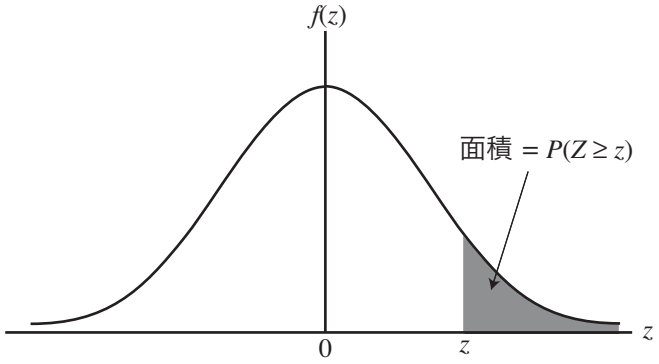

図 6: 標準正規分布の確率密度関数のグラフ上で「確率変数 Z が値 z 以上である確率」$P(Z \geq z)$

これらの性質は，正規分布を用いた統計的推測の技術についての今後の説明で，頻繁に出てきますので，よく理解してください．

正規分布の数表の見方

「標準正規分布にしたがう確率変数が，ある範囲の値をとる」確率は，数表から簡単に知ることができます．配布した数表は「標準正規分布にしたがう確率変数 Z が z 以上である確率」$P(Z \geq z)$ を計算したもので，確率密度関数のグラフにおいては図 6 のグレーの部分の面積になります．標準正規分布の確率密度関数は $z = 0$ に対して左右対称なので，数表は $z \geq 0$ についてのみ掲載されています．

さきほどの「正規分布の性質 1 」を使うと，任意の正規分布において，確率変数がある値の範囲に入る確率をこの数表だけで求めることができます．例えば，$N(50, 10^2)$ にしたがう確率変数 X が $45 \leq X \leq 60$ の範囲に入る確率 $P(45 \leq X \leq 60)$ を求めてみましょう．$Z = (X - 50)/10$ のように変換すると，性質 1 から確率変数 Z は標準正規分布 $N(0, 1)$ にしたがいます．また，

$X = 45$ のとき　$Z = (45 - 50)/10 = -0.5$
$X = 60$ のとき　$Z = (60 - 50)/10 = 1$

ですから，求める確率は $P(-0.5 \leq Z \leq 1)$ です（図 7）．図 8 から，この値は $(0.5 - P(Z \geq 1)) + (0.5 - P(Z \geq 0.5)) = 1 - (P(Z \geq 1) + P(Z \geq 0.5))$ で，数表から $P(Z \geq 1) = 0.15866$，$P(Z \geq 0.5) = 0.30854$ ですから，

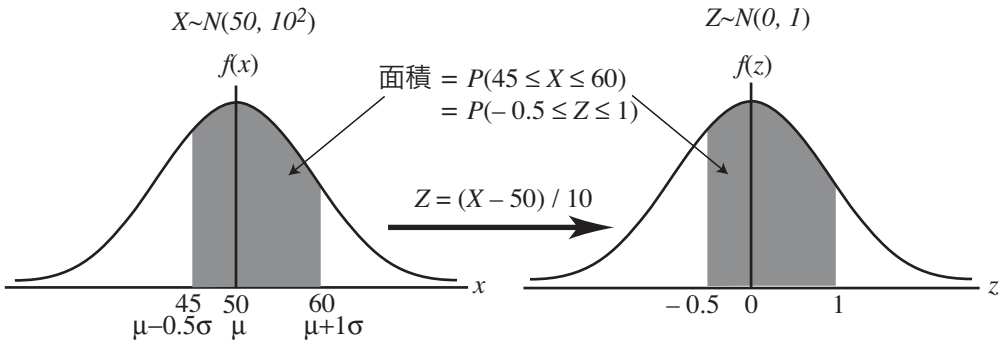

図 7: 任意の正規分布から標準正規分布への変換

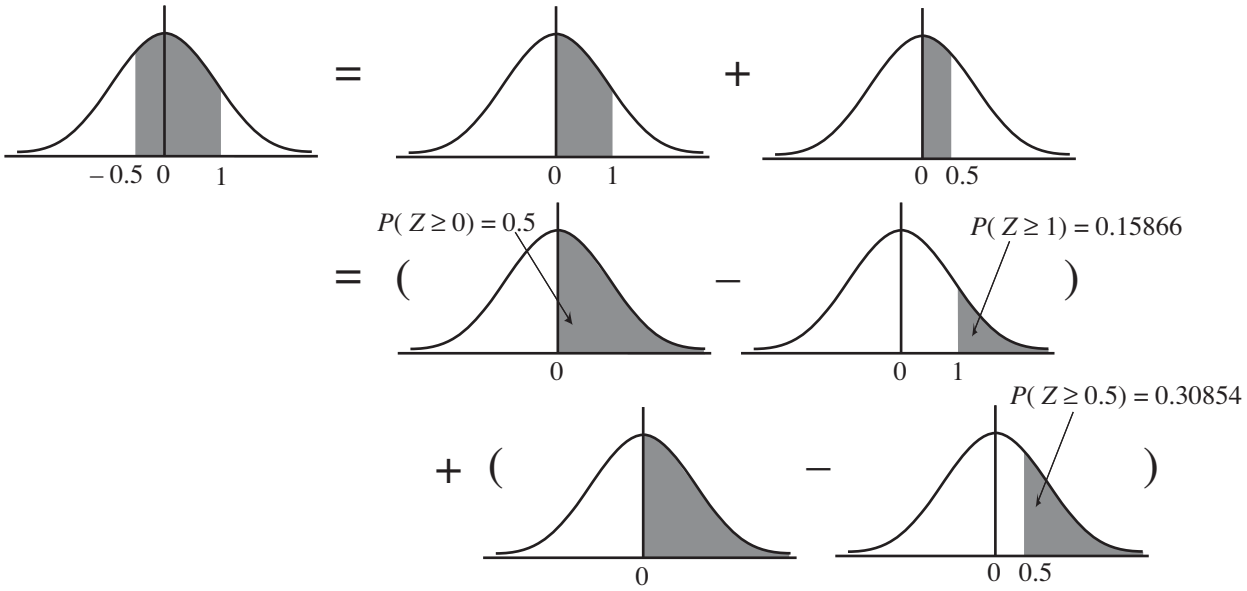

図 8: $P(-0.5 \le Z \le 1)$ を数表から求めるには

$P(45 \le X \le 60) = P(-0.5 \le Z \le 1) = 0.53280$ となります.

付録

正規分布 $N(\mu, \sigma^2)$ のモーメント母関数

モーメント母関数の定義から,

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{2\sigma^2 tx - (x-\mu)^2}{2\sigma^2}\right\} dx \end{aligned} \tag{A1}$$

となります．ここで，exp{} の中の分子の部分を完全平方にすると，

$$\begin{aligned} 2\sigma^2 tx - (x-\mu)^2 &= -\left[x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx\right] \\ &= -\left[x^2 - 2(\mu+\sigma^2 t)x + (\mu+\sigma^2 t)^2 - (\mu+\sigma^2 t)^2 + \mu^2\right] \\ &= -(x-(\mu+\sigma^2 t))^2 + (2\mu\sigma^2 t + \sigma^4 t^2) \end{aligned} \tag{A2}$$

となりますから，(A1) 式は

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-(\mu+\sigma^2 t))^2 + (2\mu\sigma^2 t + \sigma^4 t^2)}{2\sigma^2}\right\} dx \\ &= \exp\left\{\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right\} dx \\ &= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right\} dx \end{aligned} \tag{A3}$$

となります．ここで，式 (A3) の積分は正規分布 $N(\mu+\sigma^2 t, \sigma^2)$ の確率密度関数を全実数で積分したものですから，1 です[2]．よって，モーメント母関数は

$$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \tag{A4}$$

となります．

正規分布 $N(\mu, \sigma^2)$ の期待値・分散がそれぞれ μ, σ^2 であることの証明

(A4) 式から

$$\begin{aligned} M'_X(t) &= (\mu+\sigma^2 t)\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \\ M''_X(t) &= \left\{(\mu+\sigma^2 t)^2 + \sigma^2\right\}\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \end{aligned} \tag{A5}$$

となりますから，

$$\begin{aligned} E(X) &= M'_X(t)\big|_{t=0} = \mu \\ V(X) &= E(X^2) - (E(X))^2 \\ &= M''_X(t)\big|_{t=0} - \mu^2 \\ &= \mu^2 + \sigma^2 - \mu^2 = \sigma^2 \end{aligned} \tag{A6}$$

が得られます．

確率変数 X が $N(\mu, \sigma^2)$ にしたがうとき，$aX+b$ が $N(a\mu+b, a^2\sigma^2)$ にしたがうことの証明

X のモーメント母関数が $M_X(t)$ であるとき，$Y = aX + b$ のモーメント母関数 $M_Y(t)$ は

$$\begin{aligned} M_Y(t) &= M_{aX+b}(t) \\ &= E(e^{t(aX+b)}) \\ &= e^{bt}E(e^{(at)x}) = e^{bt}M_X(at) \end{aligned} \tag{A7}$$

[2]確かに 1 になることの証明は省略します．求める積分の２乗を重積分で表して，さらに変数を極座標に変換すると求められます．解析学の本を参考にしてください．

となるので，(A4) 式から

$$\begin{aligned} M_Y(t) &= e^{bt}\exp\left\{\mu(at)+\frac{\sigma^2(at)^2}{2}\right\} \\ &= \exp\left\{(a\mu+b)t+\frac{(a^2\sigma^2)t^2}{2}\right\} \end{aligned} \tag{A8}$$

となります．(A8) 式と (A4) 式を比べると，(A8) 式は $N(a\mu+b, a^2\sigma^2)$ のモーメント母関数になっていることがわかります．

正規分布の再生性の証明

確率変数 X, Y が互いに独立のとき，$X+Y$ のモーメント母関数を求めると

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \tag{A9}$$

となります．X, Y が互いに独立のとき e^{tX} と e^{tY} も独立なので，$E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY})$ という関係がなりたちます[3]．したがって，(A9) 式から $M_{X+Y}(t) = M_X(t)M_Y(t)$ という関係が導かれます．そこで，X, Y が互いに独立で，それぞれ $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ にしたがうとすると，(A4) 式のモーメント母関数を用いて

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp\left\{\mu_1 t+\frac{\sigma_1^2t^2}{2}\right\}\exp\left\{\mu_2 t+\frac{\sigma_2^2t^2}{2}\right\} \\ &= \exp\left\{(\mu_1+\mu_2)t+\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}\right\} \end{aligned} \tag{A10}$$

となり，これは $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ のモーメント母関数になっていますから，$X+Y$ も正規分布にしたがうことがわかります．

さらに，$X_1, \ldots, X_n$ が互いに独立でいずれも $N(\mu, \sigma^2)$ にしたがうとすると，(A4) 式のモーメント母関数を用いて

$$\begin{aligned} M_{X_1+\cdots+X_n}(t) &= M_{X_1}(t)\ldots\dot{M}_{X_n}(t) \\ &= \left[\exp\left\{\mu t+\frac{\sigma^2t^2}{2}\right\}\right]^2 \\ &= \exp\left\{n\mu t+\frac{(n\sigma^2)t^2}{2}\right\} \end{aligned} \tag{A11}$$

となり，これは $N(n\mu, n\sigma^2)$ のモーメント母関数になっていますから，$X_1+\cdots+X_n$ は $N(n\mu, n\sigma^2)$ にしたがうことがわかります．「X が $N(\mu, \sigma^2)$ にしたがうとき，$aX+b$ は $N(a\mu+b, a^2\sigma^2)$ にしたがう」ことは証明済ですので，X を $X_1+\cdots+X_n$ におきかえて，$a = 1/n, b = 0$ とおくと，$(X_1+\cdots+X_n)/n$ は $N(\mu, \sigma^2/n)$ にしたがうことがわかります．

[3] この証明には多次元確率分布の知識が必要ですので省略します．例えば，「講義の案内」で紹介した「基礎統計学 I　統計学入門」の第 7 章を参考にしてください．