統計学
連続型確率変数

担当：　長倉　大輔

（ながくらだいすけ）

連続型確率変数

- 連続型確率変数

株価収益率、身長、体重、などのように連続的に値をとる確率変数を**連続型確率変数**という。

連続的に値をとるとは、どんなに短い区間にも無限に取りうる値があるという事である。例えば 0.5 と 0.51 の間には 0.501が、0.501 と 0.502の間には 0.5011が…というように取りうる値がびっしりと詰まっている。

連続型確率変数に対しては離散型確率変数のように 1 つ 1 つの値に確率を付与するという事ができない。
よって点ではなく**幅に対して**確率を定義する。

連続型確率変数

- 密度関数

離散型確率変数の確率関数に対応するものとして、連続型確率変数に対して、**密度関数**を定義する。

連続型確率変数 X に対しては X が区間 $[a\,,\,b]$ の中の値を取る確率： $\Pr(a \leq X \leq b)$ を考える。この確率が

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

のようにある関数 $f(x)$ の積分で与えられる時、この $f(x)$ をこの確率変数 X の**密度関数**という。

連続型確率変数

- 密度関数の性質

密度関数は以下の性質を持つ。

(1) $\int_{-\infty}^{\infty} f(x)dx = \Pr(-\infty < x < \infty) = 1.$

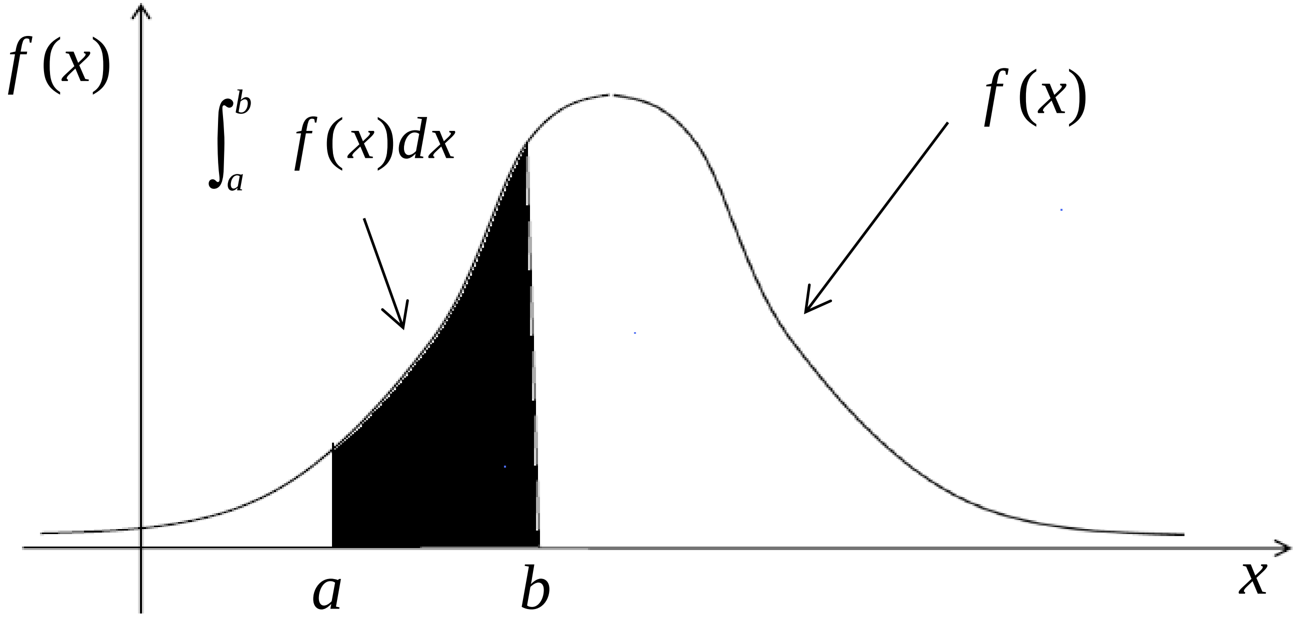
(2) 実現値として取りうる x に対して $f(x) > 0$.

ある連続型確率分布の特徴は全て密度関数の形状によって決まる。密度関数が同じであれば同じ分布である。

連続型確率変数

- 密度関数の図形的意味

確率変数 X に対して $\Pr(a \leq X \leq b)$ は密度関数 $f(x)$ の区間 $[a, b]$ での積分、$\int_a^b f(x)\mathrm{d}x$ で与えられる。図形的には下図の黒い部分の**面積の値**の事である。



連続型確率変数

- 1 点の確率

連続型確率変数 X に対して$\Pr(X = a)$の確率はどのような値をとるだろうか？

連続型確率変数の確率は上のように面積で定義されるが、$\Pr(X = a)$ は線であり面積を持たない。つまり

連続型確率変数が 1 点を取る確率は 0 である

と考えるのである ($\Pr(X = a) = 0$)。

連続型確率変数

連続型確率変数のこの性質より

$$\Pr(a \leq X) = \Pr(a < X) + \Pr(X = a) = \Pr(a < X),$$

であり、また同様に

$$\Pr(a \leq X \leq b) = \Pr(a < X < b)$$

となる。

連続型確率変数

- 累積分布関数

連続型確率変数 X の**累積分布関数**も離散型の時と同様に

$$F(x) = \Pr(X \leq x)$$

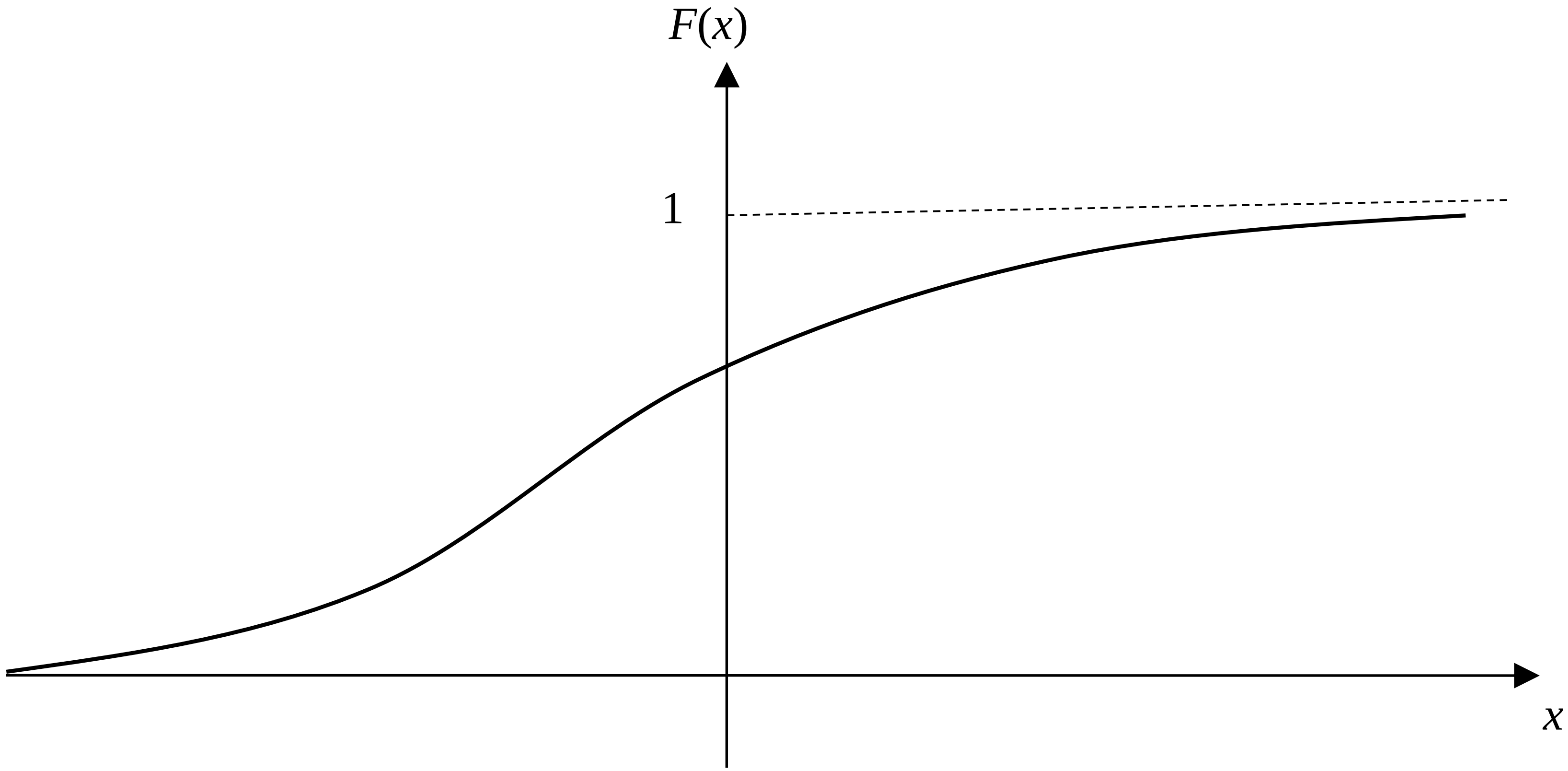
で定義される。連続型確率変数の場合はこれは

$$F(x) = \Pr(X \leq x) = \Pr(-\infty < X \leq x)$$
$$= \int_{-\infty}^{x} f(z)dz$$

のように**区間 $(-\infty, x]$ の積分**で与えられる。

連続型確率変数

- 連続型確率変数の分布関数の形状



連続型確率変数

- 連続型確率変数の平均と分散

連続型確率変数の期待値と分散は密度関数を用いて、以下のように定義される。

(平均)

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)dx$$

(分散)

$$\mathrm{var}(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

連続型確率変数の平均と分散も 離散型確率変数のスライドで述べられた性質を全て満たす。

連続型確率変数

- 連続型確率変数の共分散

連続型確率変数 X, Y の共分散は

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y) f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

と定義される 。ここで $f_{XY}(x, y)$ は**同時密度関数**と呼ばれるもので

(1) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$,

(2) 実現値として取りうる x, y に対して $f_{XY}(x, y) > 0$,

(3) $\Pr(a < X < b,\ c < Y < d) = \int_c^d\int_a^b f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y$,

を満たすx, y の2変量関数である。

連続型確率変数

連続型確率変数の共分散も離散型確率変数のスライドで述べられた性質を全て満たす。ただし性質(1) における $E(XY)$ は

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x, y)\,\mathrm{d}\,x\,\mathrm{d}\,y$$

で置き換えられる。

連続型確率変数

- 連続型確率変数の相関係数

離散型の場合と同じように

$$\rho_{XY} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

と定義される。$|\rho_{XY}| \leq 1$ を満たす。

連続型確率変数

- 連続型確率変数の独立性

連続型確率変数 X と Y が**独立**であるとは同時密度関数がX と Y の**密度関数の積として表せる**事である。すなわち

$$f_{XY}(x, y) = f_X(x)\, f_Y(y)$$

となる事である。この時 $f_X(x)$と $f_Y(y)$を特に**周辺密度関数**と呼ぶ事もある。

連続型確率変数

- 条件付密度関数

連続型確率変数 X, Y について X の $Y=y$ という**条件付の密度関数**は

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

と定義される。

- 同時密度関数の分解

上の定義式の両辺に $f_Y(y)$ をかけると同時密度関数は

$$f_{X|Y}(x \mid y) f_Y(y) = f_{XY}(x, y)$$

と**条件付密度関数と周辺密度関数の積**として表す事ができる。

連続型確率変数

- 条件付期待値、分散

連続型確率変数 X, Y について X の $Y=y$ という**条件付期待値と分散**はそれぞれ

$$\mu_{X|Y} = E(x\,|\,y) = \int_{-\infty}^{\infty} x\, f_{X|Y}(x\,|\,y)dx$$

および

$$\sigma_{X|Y}^2 = \mathrm{var}(x\,|\,y) = \int_{-\infty}^{\infty} (x-\mu_{X|Y})^2\ f_{X|Y}(x\,|\,y)dx$$

と定義される。周辺密度関数が条件付き密度関数に置き換わっただけである。

連続型確率変数

- 一様分布(uniform distribution)

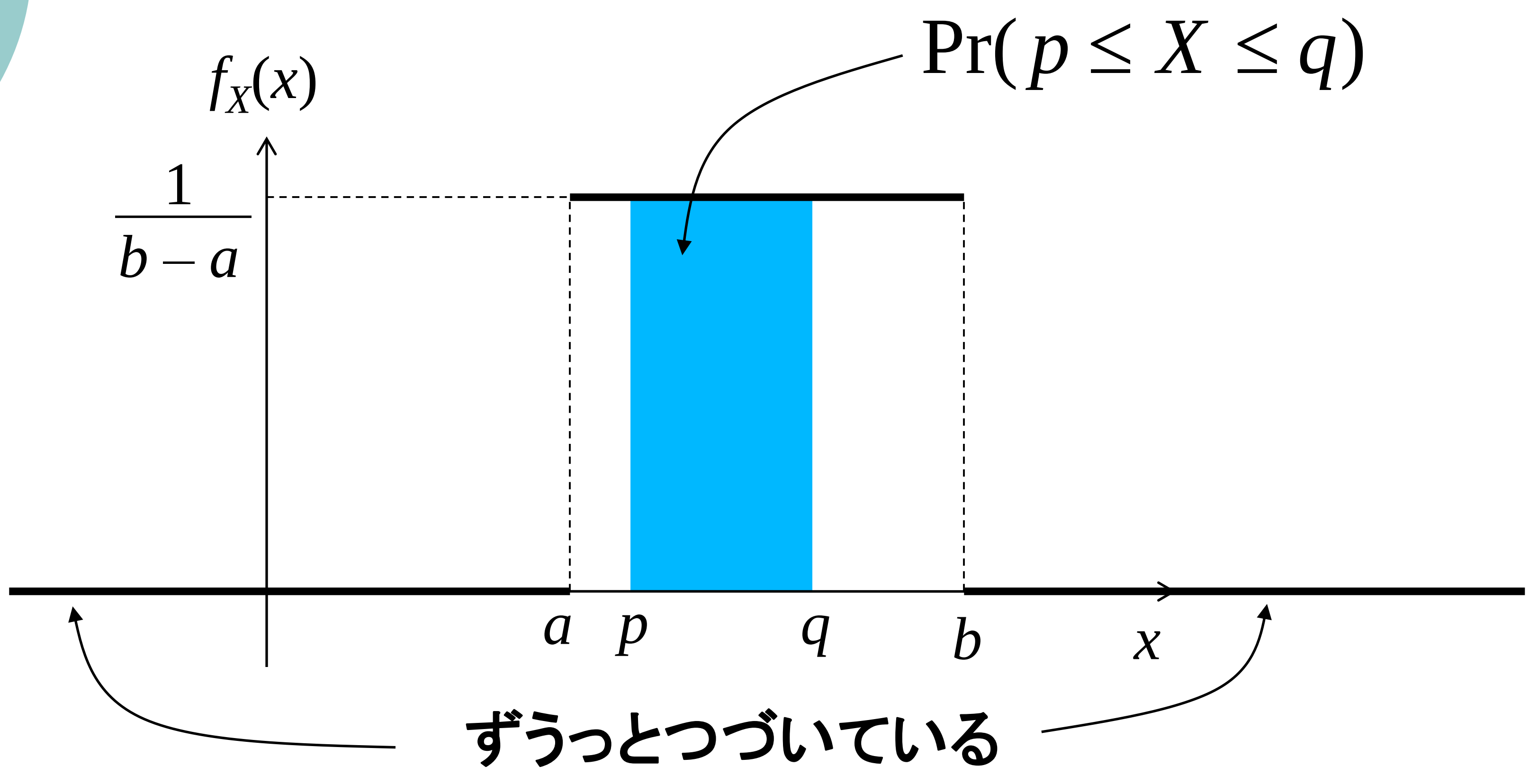
確率変数 X が a から b まで一様に分布している時、その確率変数の分布は**一様分布**と呼ばれる。その密度関数と分布関数は

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{その他} \end{cases} \qquad F_X(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & b < x \end{cases}$$

となる。

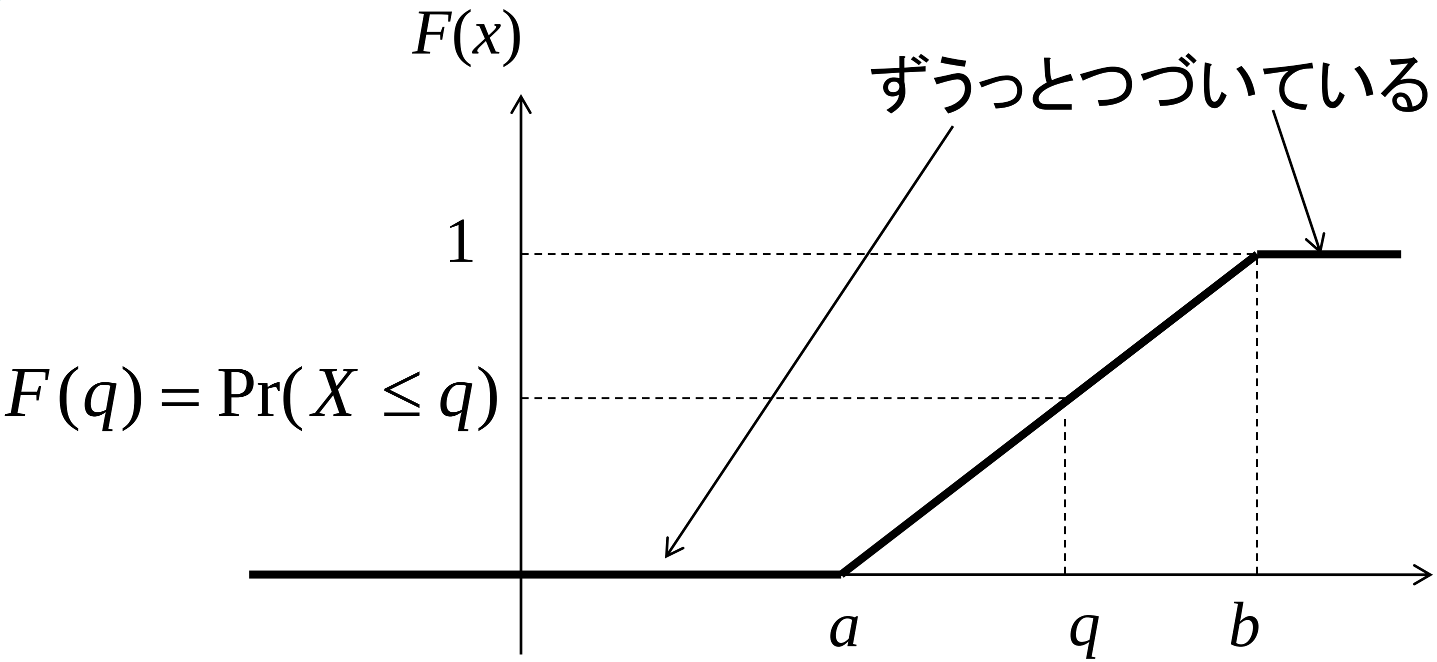
連続型確率変数

- 一様分布の密度関数



連続型確率変数

- 一様分布の分布関数



連続型確率変数

- 一様分布の期待値、分散

(期待値)　a と b の中間の $(b + a)/2$ である。

(分散)　　$(b - a)^2/12$である。

連続型確率変数

- (補足)一様分布の期待値、分散

一様分布の期待値と分散は、期待値と分散の定義より求められる。

(期待値)

$$E(X) = \mu_X = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

(分散)

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^2+ab+a^2}{3}$$ より

$$\mathrm{var}(X) = E(X^2) - \mu_X^2 = \frac{(b-a)^2}{12}$$ と計算すればよい。

連続型確率変数

- (0, 1) の一様分布

一様分布で最もよく使われるのは $a = 0, b = 1$ の場合である。これは**(0, 1) の一様分布**と呼ばれる。この時

(密度関数) $f_X(x) = 1$, (分布関数) $F_X(x) = \Pr(X \leq x) = x$

(期待値) 1/2 (分散) 1/12

となる。重要な特徴は上記分布関数からもわかるように

「X が x 以下になる確率は x である」

という事である。この性質はいろいろな問題を考えるうえで非常に重要になる。

連続型確率変数

- 一様分布の表記

確率変数 X が(a, b) 上の一様分布に従っている時、

$$\boldsymbol{X \sim U\,(a, b)}$$

と表記する。$X \sim U\,(0, 1)$ であれば X は(0, 1)の一様分布に従っているという事である。

連続型確率変数

例題1

確率変数 X が $X \sim U(0, 1)$の時

(a) 確率変数 $Z = 5X$ の分布関数、および密度関数を求めなさい。

(b) 確率変数 $W = X^2$の分布関数を求めなさい。

連続型確率変数

- 正規分布 (normal distribution)

確率変数 X が $-\infty$ から ∞ までの間の実数を取り、その密度関数が

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right], \qquad -\infty < x < \infty$$

で与えられている時、確率変数 X は**正規分布**に従っているという。ここで $\exp[\,a\,]$ は e^a を、e は自然対数の底 ($e = 2.7182818\ldots$)を表している。

連続型確率変数

- 正規分布の期待値と分散

正規分布の期待値と分散は以下の積分を計算する事によって与えられる。(この計算は難しいので略する)。

(期待値)

$$E(X)=\int_{-\infty}^{\infty} x f_X(x)dx=\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx=\mu$$

(分散)

$$\mathrm{var}(X)=\int_{-\infty}^{\infty}(x-\mu)^2\ f_X(z)dz$$

$$=\int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx=\sigma^2$$

密度関数の中の μ と σ^2 が期待と分散になっている。

連続型確率変数

- 正規分布の表記

確率変数 X が平均 μ , 分散 σ^2 の正規分布に従っているとき、

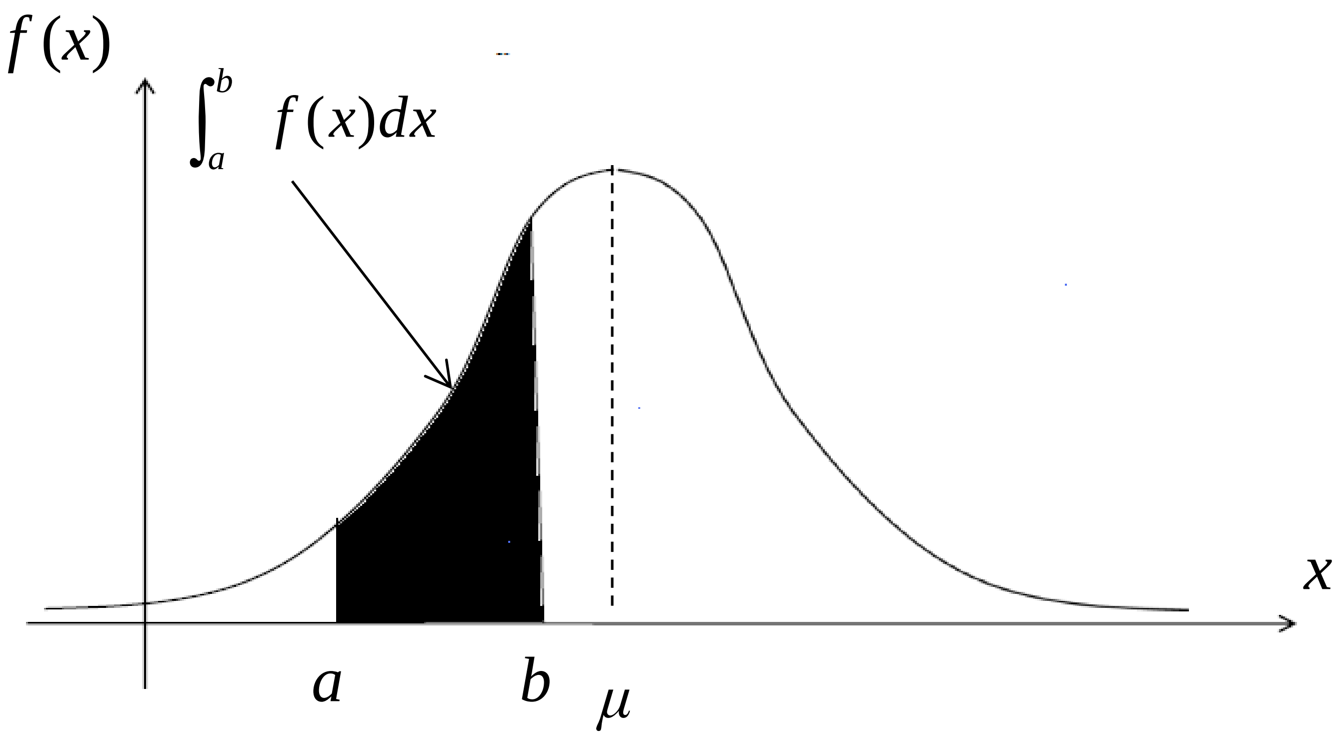
$$X \sim N(\mu, \sigma^2)$$

と表記する。

連続型確率変数

- 正規分布の密度関数

正規分布の密度関数は以下の形状をしている
(平均 μ で**左右対称**になっている)。

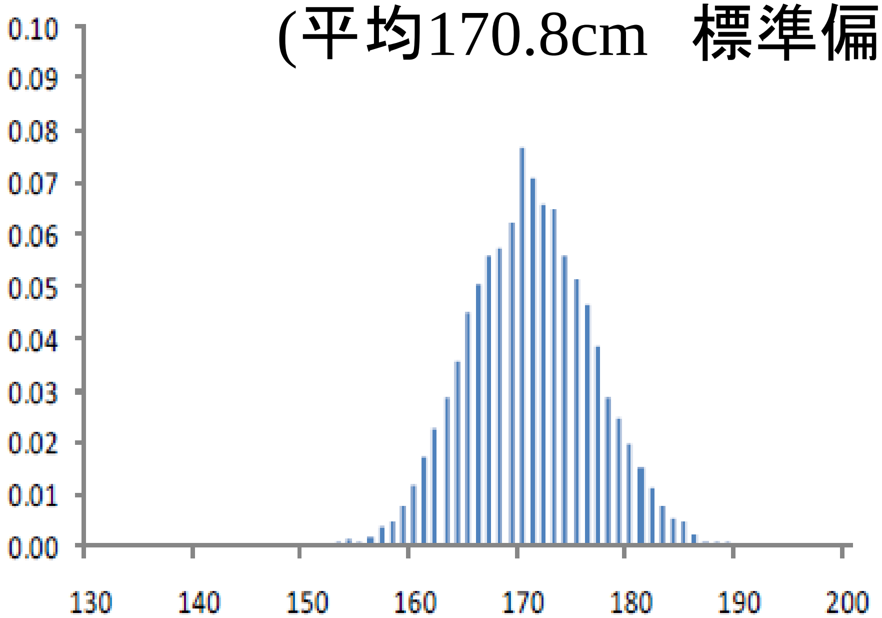


連続型確率変数

- (例)正規分布

身長は正規分布に従っている(厳密に言うと正規分布でよく近似できる)と考えられている。下の図は2005年の17歳の男性の身長データのヒストグラムである

(平均170.8cm　標準偏差 5.81cm)

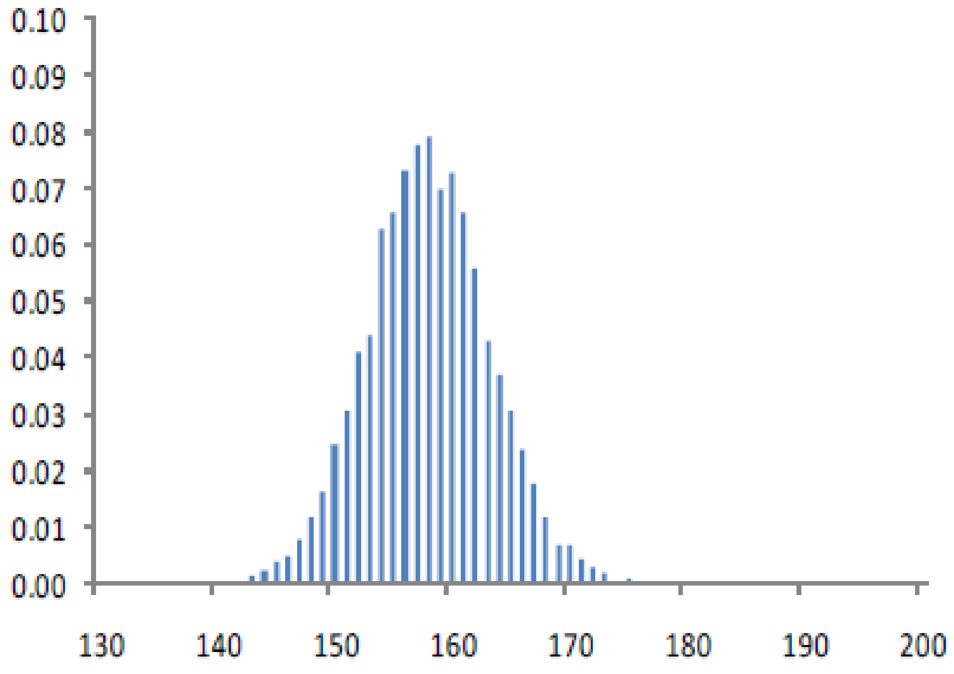


文部科学省の学校保険統計調査より

連続型確率変数

- (例2)正規分布

下の図は 2005 年の17歳の女性の身長データのヒストグラムである(平均158cm　標準偏差5.28 cm)

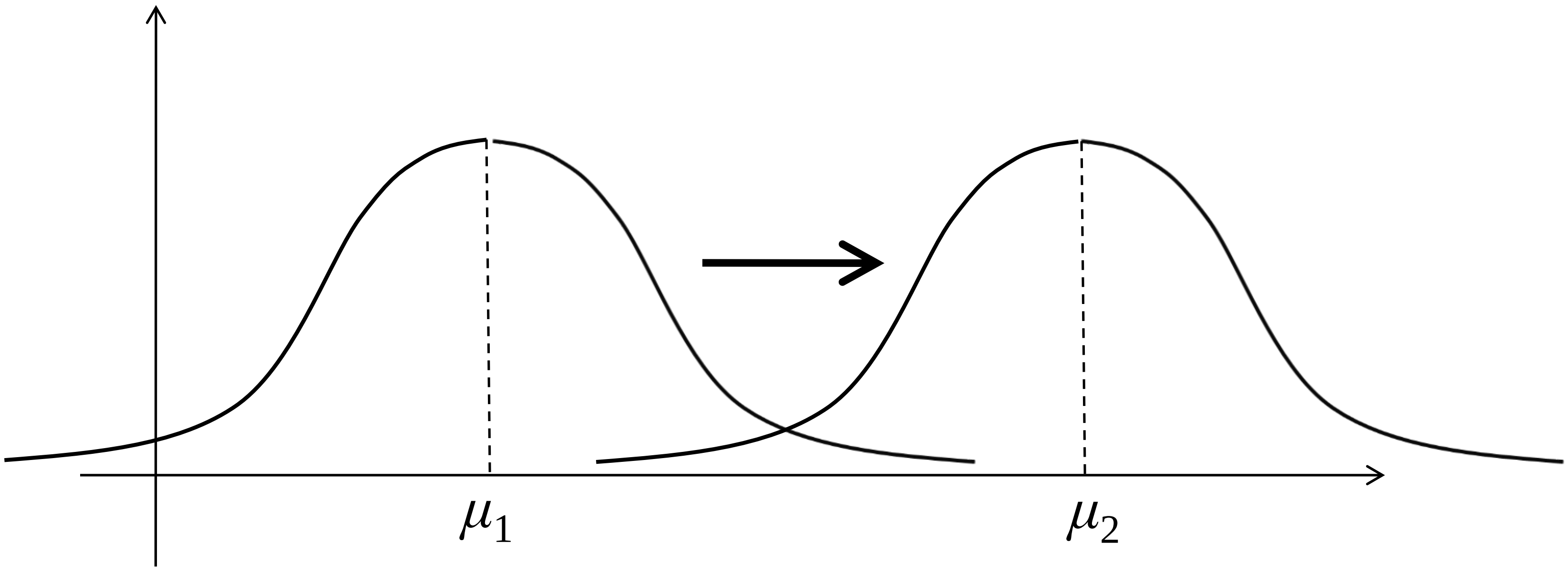


文部科学省の学校保険統計調査より

連続型確率変数

- 正規分布の密度関数と平均 μ の関係

分散 σ^2 を一定にしたまま平均が μ_1 から μ_2 に変化すると

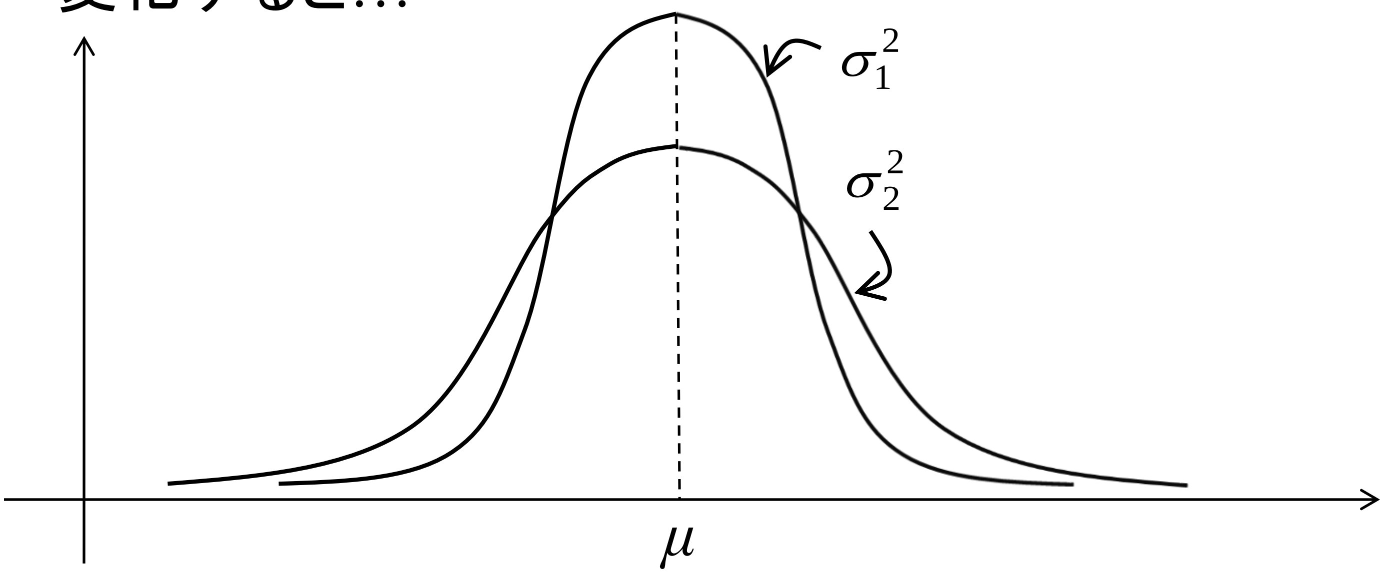


密度関数の形状は同じまま**平行に移動する**。

連続型確率変数

- 正規分布の密度関数と分散 σ^2 の関係

平均 μ を一定にしたまま分散が が σ_1^2 から $\sigma_2^2 (> \sigma_1^2)$ に変化すると…



位置は変わらないが、密度関数の**広がりが大きくなる**。

連続型確率変数

- 正規分布の性質

正規分布には以下の重要な性質がある

(1) (**線形変換しても正規分布**)　$X \sim N(\mu_X, \sigma_X{}^2)$ とすると $Y = aX + b$ も正規分布に従い

$$Y \sim N(a\mu_X + b, a^2\sigma_X{}^2)$$

である。

(2) (**和の分布も正規分布**) $X \sim N(\mu_X, \sigma_X{}^2)$ および $Y \sim N(\mu_Y, \sigma_Y{}^2)$ である時、$Z = X + Y$ も正規分布に従い。

$$Z \sim N(\mu_X + \mu_Y,\ \sigma_X{}^2 + \sigma_Y{}^2 + 2\sigma_{XY})$$

である。ここで $\sigma_{XY} = \mathrm{cov}(X, Y)$ である。

連続型確率変数

- 正規分布の性質

(2) の性質を繰り返して用いれば、一般に、n 個の正規分布に従う確率変数の和もやはり正規分布に従う事がわかる。

(2)の性質は任意の確率変数に対して、一般的には**成り立たない**。

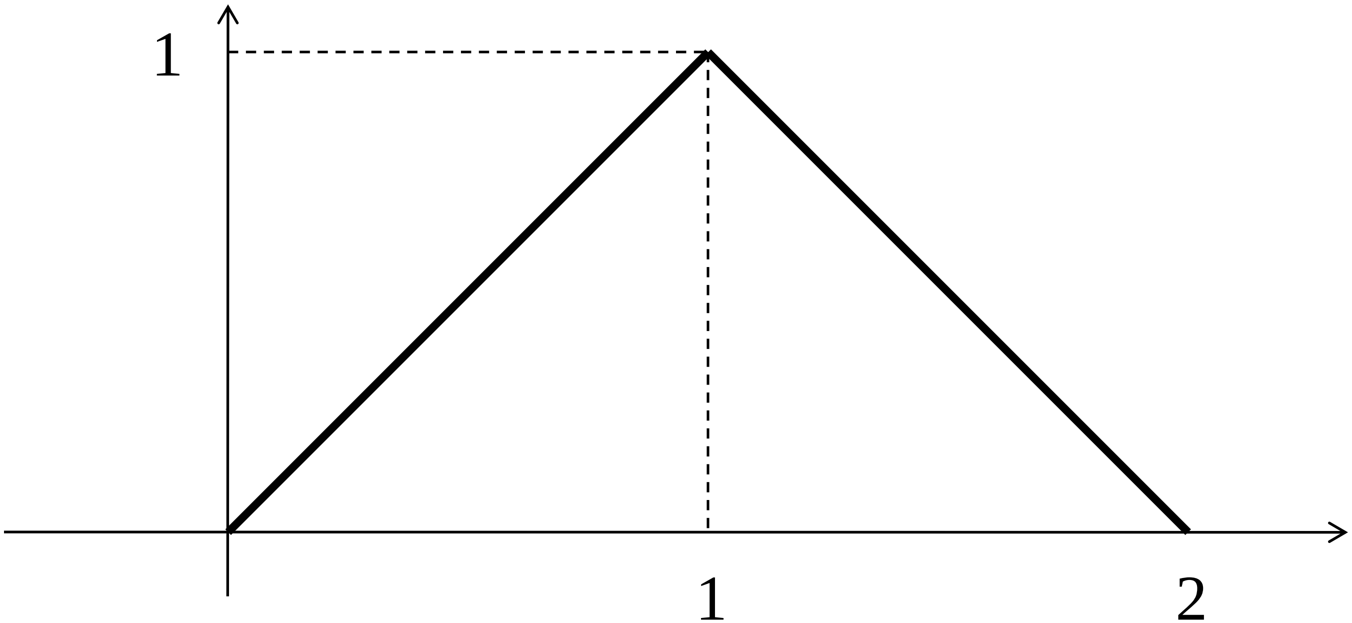
例えば、それぞれが一様分布に従う 2 つの独立な確率変数の和の分布はもはや一様分布**ではない**。

X_1 と X_2 は独立で $X_1 \sim U(0, 1)$, $X_2 \sim U(0, 1)$とすると、
$Y = X_1 + X_2$の密度関数は (次ページ)

連続型確率変数

- 一様分布の和の密度関数

(Y の密度関数)



となる(証明は結構難しい)。明らかに一様分布ではない

連続型確率変数

- 標準正規分布

正規分布において期待値 μ を $\mu = 0$, 分散 σ^2 を $\sigma^2 = 1$ としたものには特に**標準正規分布**という特別な名前が与えられており、その密度関数は

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2} \right], \quad -\infty < x < \infty$$

となる。期待値 μ 分散 σ^2 の正規分布に従う確率変数 X に対して、新しく確率変数 Z を

$$\boldsymbol{Z = (X - \mu)/\sigma}$$

と定義すると、正規分布の性質より、Z は標準正規分布に従うことを示すことができる。

連続型確率変数

- 正規分布の分布関数

正規分布の分布関数は**明示的には得られない**。

(つまり分布関数の $F_X(x)=\int_{-\infty}^{x} f_X(z)dz$ という積分が明示的に解けない)。

しかしながら、非常に高精度の近似式がたくさんあり、応用上はそれで十分である。

例えば Excel であれば 「 NORMDIST( )」というExcel関数が正規分布の関数を計算してくれる。

連続型確率変数

- 分布表を使った確率の計算

また、様々な実際の数値に対して、標準正規分布の分布関数の値を表にしたものを（標準正規分布の）正規分布の分布表といい、統計学の教科書の巻末に載っていることがおおい（例えば、この講義の参考文献としてあげた本にも載っている）。

それを用いれば、ある具体的な数字 a が与えられた時に、任意の正規分布に従う確率変数 $X \sim N(\mu, \sigma^2)$に対して $\Pr(X \leq a)$ $(= \Pr(X < a)$ である事に注意) を求めることができる。

連続型確率変数

分布表によって分布関数の値を以下のように求めることができる。

まず $X \sim N(\mu, \sigma^2)$の時、$\Pr(X \le a)$ という確率は

$$\begin{aligned} \Pr(X \le a) &= \Pr(X - \mu \le a - \mu) \\ &= \Pr((X - \mu)/\sigma \le (a - \mu)/\sigma) \\ &= \Pr(Z \le (a - \mu)/\sigma) \end{aligned}$$

と、$Z \sim N(0, 1)$ を用いて表す事ができる。標準正規分布表には、ある z に対して$\Pr(Z < z)$ の値が載っているので、$z = (a - \mu)/\sigma$ のところの値を見ればよい。
(参考書では $\Pr(Z < z)$ という表記の代わりに $P\{Z < z\}$ という表記になっているが同じものを表す)。

連続型確率変数

■ 分布表の見方

講義Webにある標準正規分布表（参考文献のものと同じ）から、$P(Z < 0.22)$ と $P(Z < 1.64)$を探してみよう。

z	0.00	0.01	0.02	0.03	0.04	….	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160		
0.1	0.5398	:	:	:	:		
0.2	0.5793		0.5871				
:	:						
1.6					$\boxed{0.9495}$		

これは $\Pr(Z < 0.22) = 0.5871$, $\Pr(Z < 1.64) = 0.9495$
という事を示している。

連続型確率変数

(1) この分布表には Pr $(Z < z)$ しか載っていないが、
Pr $(Z > z)$ は $1 - \Pr(Z < z)$によって計算できる。

(2) z が 0 以上の場合しか載っていないが、標準正規分布の密度関数は **0 で左右対称**なので、どのような $a > 0$ に対しても

$$\Pr(Z < -a) = \Pr(Z > a) = 1 - \Pr(Z < a)$$

が成り立つので、この関係から0以下の場合も計算できるので問題ない(例えば Pr $(Z < -0.5)$ は $\Pr(Z < -0.5) = 1 - \Pr(Z < 0.5) = 0.3085$ と計算できる)

連続型確率変数

例題 2 (分布表を使った確率の計算)

(a) $X \sim N(7, 4)$の時、$\Pr(X \leq 10)$ を求めなさい。

(b) ある国の男子の身長は $N(171, 6^2)$ であるとする
(単位はcm)。

168cm以上180cm以下 の男子は全体の何%をしめるか?

演習問題

問題1 確率変数 X が $X \sim U(0, 1)$ の時 $Z = a + (b - a)X$ はどのような分布になるか答えなさい。ただし、$b > a$ とする。

問題2 確率変数 X が $X \sim N(12, 4)$ の時、次の確率を求めよ

(a) $X > 10$, (b) $X < 8$, (c) $|X - 12| < 3$

問題3 $X \sim N(0, \sigma^2)$, $Y = X^2$とする。cov(X, Y) を求めなさい(ヒント: この時 $E(X^3) = 0$ である)。