統計基礎

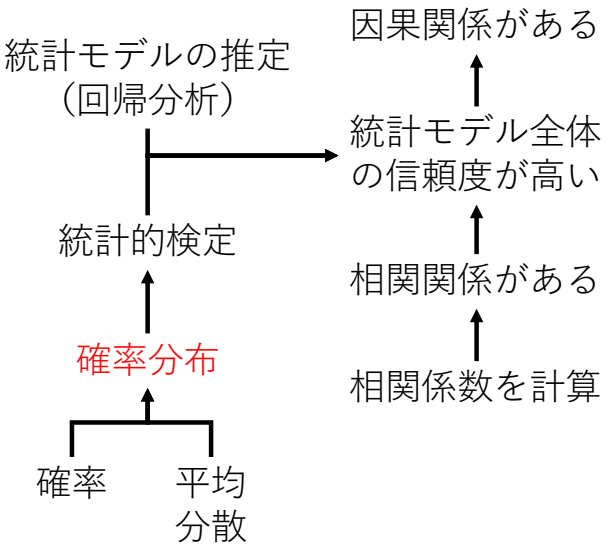
古谷知之

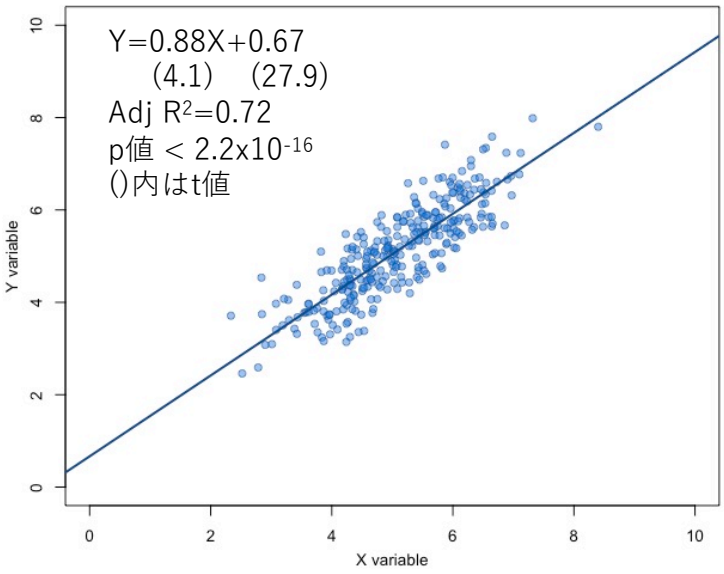
授業概要

＊履修者の状況に応じて変更される場合がありますが、概ね以下のような流れで授業を進めます。

第1回　ガイダンス・確率
第2回　確率変数と確率分布(1)
第3回　確率変数と確率分布(2)
第4回　母集団と平均
第5回　単回帰分析(1)
第6回　単回帰分析(2)
第7回　仮説検定(1)
第8回　仮説検定(2)
第9回　重回帰分析
第10回　R演習(1)
第11回　R演習(2)
第12回　R演習(3)
第13回　R演習(4)
第14回　最終試験

授業の全体像





確率変数と確率分布

- 連続型確率分布の平均・分散・標準偏差
- 連続型確率分布としての正規分布
- 二変量の確率分布
- 同時確率、周辺確率、条件付き確率

連続型確率分布

確率変数Xのとる値が関数$f(x)$を用いて

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

で与えられる時、 Xを連続型確率分布といい、確率分布を持つ。
ここで、

$$f(x) \geq 0, \int_{-\infty}^{\infty} f(x)dx = 1$$

連続型確率分布の期待値・分散・標準偏差

- 連続型確率分布の期待値$E[X]$は以下のように定義される

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

- 同様に分散$Var[X]$と標準偏差$sd(X)$は以下のように定義される

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

$$sd(X) = \sqrt{Var[X]}$$

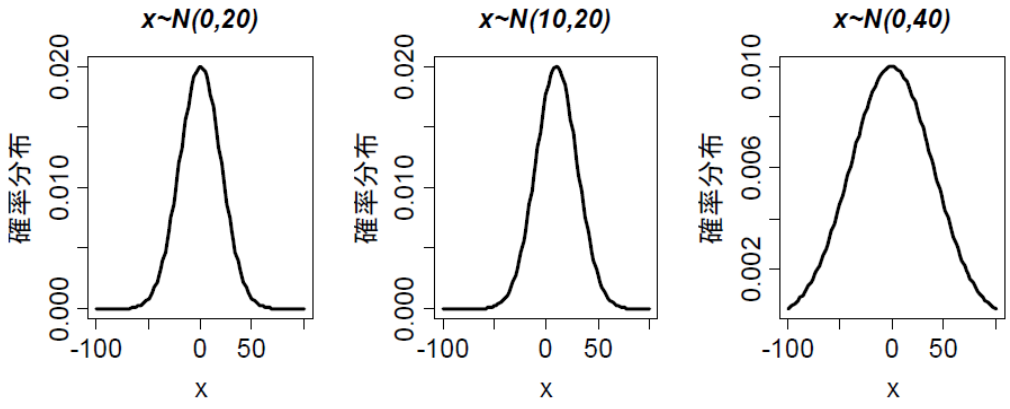
正規分布（ガウス分布）

- 平均μ、分散σ^2となる以下の確率密度関数に従う分布を正規分布という

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

$$E\left(Z\right)=\mu,V\left(Z\right)=\sigma^2$$

正規分布



２つの離散変数に関する確率分布

例：２つのコインをトスする、２つのサイコロを投げる

２つの離散型確率変数をそれぞれX, Yとする。このとき、 $X = x$ かつ$Y = y$となる確率

$$Pr(X = x, Y = y) = f(x, y)$$

を確率変数(X, Y)の同時確率分布という。ここで、

$$f(x, y) \geq 0, \sum_x \sum_y f(x, y) = 1$$

２つの離散変数に関する確率分布

２つの離散型確率変数X, Yに対して、 Xが与えられたときにYが得られる条件付き確率は、以下のようにして定義される。

$$Pr(Y = y|X = x) = \frac{Pr(Y = y, X = x)}{Pr(X = x)}$$

２つの離散変数に関する確率分布

２つの離散型確率変数X, Yに対して同時確率$f(x, y)$が与えられたとする。このとき、

$$f_x(x) = \sum_y f(x, y)$$
$$f_y(y) = \sum_x f(x, y)$$

を確率変数(X, Y)の周辺確率分布という

２つの離散確率変数の共分散

２つの確率変数X, Yに対して、同時確率分布$f(x, y)$が与えられたとする。この時、共分散は以下のように定義される。

$$Cov(X, Y) = E\big[X - E[X]\big]E\big[Y - E[Y]\big]$$
$$= \sum_x \sum_y (x - \mu_x)\big(y - \mu_y\big) f(x, y)$$

ここで、 $\mu_x = E[X], \mu_y = E[Y]$とする。

共分散の性質

共分散は次式のように展開できる

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

【証明】

$$\begin{aligned} Cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E\big[XY - XE[Y] - E[X]Y + E[X]E[Y]\big] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

相関係数

確率変数X, Yの相関係数ρは以下のように定義される

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}$$

相関係数は必ず$-1 \leq \rho \leq 1$となる。

$\rho = 0$のとき、無相関であるという。

スクリーニングテストの例

- 兆候の有無：事象Y
  - 兆候がある：Y_+、兆候がない：Y_-
- 健康状態：事象X
  - 健康である：$X_{健康}$、病気である：$X_{病気}$

健康状態	兆候（Y）		
（X）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

スクリーニングテストの例

- スクリーニングテストを実施した人から任意に一人を選んだとき

- 健康な人である確率は？
- 病気の兆候がある確率は？
- 兆候がありかつ病気である確率は？

スクリーニングテストの例

- 無作為に一人選んだ人が、兆候が＋でかつ病気である確率（同時確率）　$P(X_{病気} \cap Y_{+})$
- 無作為に選んだ人が病気である確率$P(X_{病気})$、無作為に選んだ人が徴候がある$P(Y_{+})$確率（周辺確率）
  無作為に選んだ人が兆候が+のとき病気である確率（条件付き確率）　$P(X_{病気}|Y_{+})$

健康状態	兆候（Y）		
（X）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

S.セン（2005）、p.76

条件付き確率、同時確率、周辺確率

- 同時確率　$P(X_{病気} \cap Y_{+})$ =75/1000
- 周辺確率　$P(Y_{+})$ =175/1000
- 条件付き確率　$P(X_{病気}|Y_{+})$=75/175
- 同時確率＝条件付き確率×周辺確率

健康状態	兆候（Y）		
（X）	ー	＋	合計
健康	800	100	900
病気	25	75	100
合計	825	175	1000

S.セン（2005）、p.76

同時確率分布

健康状態	兆候（Y）		
（X）	−	＋	合計
健康	0.800	0.100	0.900
病気	0.025	0.075	0.100
合計	0.825	0.175	1

S.セン（2005）、p.76

共分散

確率変数X,Yの期待値は、それぞれ以下のようになる。

$$\mu_x = E[X] = \sum_x x\,Pr(X = x) = \sum_x x\,f_x(x) = 0.9\times900 + 0.1\times100 = 820$$

$$\mu_y = E[Y] = \sum_y y\,Pr(Y = y) = \sum_y y\,f_y(y) = 0.825\times825 + 0.175\times175 = 711.25$$

従って確率変数X,Yの共分散$Cov(X,Y)$は以下のようにして得られる。

$$Cov(X,Y) = E\big[X - E[X]\big]E\big[Y - E[Y]\big] = \sum_x\sum_y (x-\mu_x)\big(y-\mu_y\big)\,f(x,y)$$

$$= (800-820)(800-711.25)\times0.8 + (100-820)(100-711.25)\times0.1$$
$$+(25-820)(25-711.25)\times0.025 + (75-820)(75-711.25)\times0.075$$
$$\cong 91779.69$$

二変量の連続型確率変数

２つの連続型確率変数をX,Yとする。このとき、

$$Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dxdy$$

を、確率変数(X,Y)の同時確率密度関数と呼ぶ。ここで、

$$f(x,y) \geq 0, \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

となる。

二変量の連続型確率変数

２つの連続型確率変数をX,Yの同時確率密度関数が$f(x,y)$で与えられたとする。このとき、

$$f_x(x) = \int f(x,y)\,, f_y(x) = \int f(x,y)$$

をそれぞれX,Yの周辺確率密度関数と呼ぶ。

共分散は次式のように定義される。

$$Cov(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y) f(x,y)dxdy$$

ここで、 $\mu_x = E[X], \mu_y = E[Y]$とする。

二変量正規分布の例