統計基礎

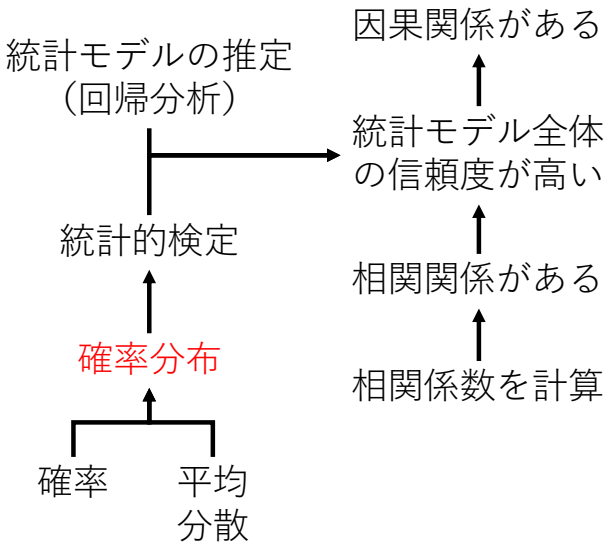
古谷知之

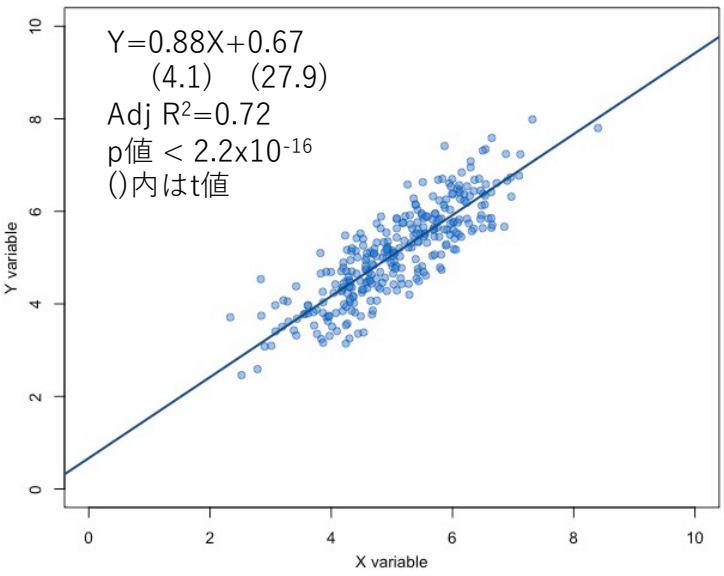
授業概要

＊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





授業内容

- 確率変数と確率分布
- 離散型確率変数の平均・分散・標準偏差
- 一様分布
- 二項分布
- ポアソン分布
- 指数分布
- ベータ分布
- ガンマ分布
- 正規分布、標準正規分布
- χ^2分布

授業ではこれ以外にも、
t分布やF分布といった
確率分布を扱う

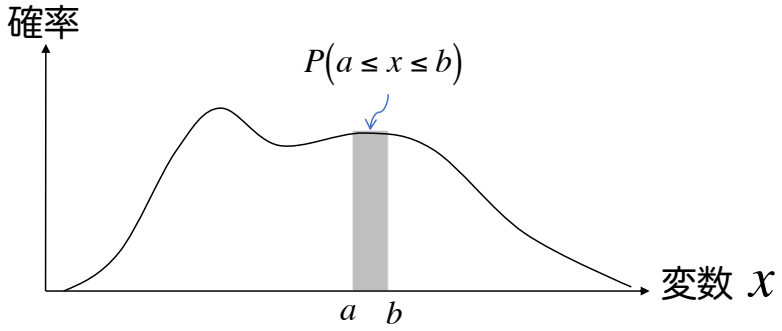
離散型の確率変数と確率分布

- 値の生じ易さに確率値を与えられるような変数を確率変数という
- 確率変数とそれに対応する確率値を確率変数
- 対応関係をまとめた表を確率分布表という

コイントス（確率変数）	表	裏
確率	1/2	1/2

サイコロの目（確率変数）	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

連続的な確率変数と確率分布



- 変数が連続値をとるときなどは確率変数の分布を確率密度関数で表現する

確率分布の種類

- 離散的な確率分布
  - 一様分布
  - 二項分布
  - ポアソン分布
- 連続的な確率分布
  - 正規分布
  - ベータ分布
  - ガンマ分布

確率変数の全体的傾向を把握するには？

- 確率変数の分布について、全体的傾向を把握するために代表値と呼ばれる指標を用いることが多い。

- その代表的な指標として、期待値（平均）や分散・標準偏差がある。

- その他にも、最頻値や中央値と呼ばれる指標も使われる。

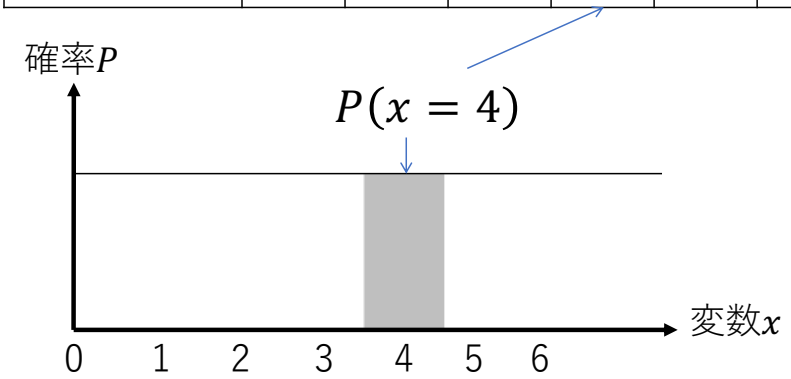
離散型確率分布の平均・分散・標準偏差

確率変数X	x_1	x_2	x_3	$\cdots$	x_n
確率p	p_1	p_2	p_3	$\cdots$	p_n

- 確率変数Xの値x_iと確率p_iが上表のように分布しているとき、確率変数Xの平均μ、分散σ^2、標準偏差σは以下のようになる
- 平均(期待値)$\mu = x_1p_1 + x_2p_2 + \cdots + x_np_n$
  →確率変数がどの辺りに集中しているかを示す
- 分散$\sigma^2 = (x_1 - \mu)^2p_1 + (x_2 - \mu)^2p_2 + \cdots + (x_n - \mu)^2p_n$
- 標準偏差σ
  →分散と標準偏差は、確率変数がどの程度散らばっているかを示す

サイコロの目の確率分布

サイコロの目（確率変数）	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6



サイコロの目の確率分布

サイコロの目（確率変数）	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

- 平均(期待値)$\mu = 1\times(1/6) + 2\times(1/6) + \cdots + 6\times(1/6) = 3.5$
- 分散$\sigma^2 = (1-3.5)^2\times(1/6) + (x_2-\mu)^2\times(1/6) + \cdots + (x_n-\mu)^2\times(1/6) \cong 2.9$
- 標準偏差$\sigma \cong 1.7$

線形変換

- 平均μ、分散σ^2の確率変数Xに対して、線形変換により$X' = cX + d$という新しい確率変数を作る。このときX'の平均$E(X')$と分散$V(X')$は以下のようになる

- $E(X') = \sum_{i=1}^{n}(cx_i + d)p_i = c\sum_{i=1}^{n}x_ip_i + d\sum_{i=1}^{n}p_i = c\mu + d$
- $V(X') = \sum_{i=1}^{n}[(cx_i + d) - (c\mu + d)]^2p_i = c^2\sum_{i=1}^{n}(x_i-\mu)^2p_i = c^2\sigma^2$

離散型確率変数と離散型確率分布

確率変数Xが、有限個の実現値$x_1, x_2, \cdots, x_k, \cdots$の値を取りうるとする。このとき$x_k$が実現する確率が

$$Pr(X = x_k) = f(x_k)$$

で与えられるとき、これを離散型確率分布という。ここで

$$f(x_k) \geq 0, \sum_{k=1}^{\infty} f(x_k) = 1$$

である。またこのとき、 Xを離散型確率変数という。

離散型確率変数の期待値

• 確率変数Xの期待値$E[\mathrm{X}]$は以下で定義される。

$$E[X] = \sum_{k=1}^{\infty} x_k f(x_k)$$

$$= \sum_{k=1}^{\infty} x_k Pr(X = x_k)$$

確率変数の期待値の性質

- 確率変数Xの期待値$\mathrm{E}[\mathrm{X}]$の演算は、次の性質を満たす。
  ここで、 X, Yは確率変数、 cは定数とする。

$$\mathrm{E}[c] = c$$

$$\mathrm{E}[c(X)] = cE[X]$$

$$\mathrm{E}[X + Y] = E[X] + E[Y]$$

確率変数の分散・標準偏差の定義

- 確率変数Xの分散$\mathrm{Var}(X)$と標準偏差sdは以下のように定義される

$$\mathrm{Var}(X) = E[(X - E[X])^2] = \sum_{k=1}^{\infty} (x_k - \mu)^2 f(x_k)$$

$$sd = \sqrt{\mathrm{Var}(X)}$$

ここで、 μはx_kの平均値である。

確率変数の分散・標準偏差の性質

- 確率変数Xの分散$Var(X)$は以下のような性質を持つ

$$Var(c) = 0$$

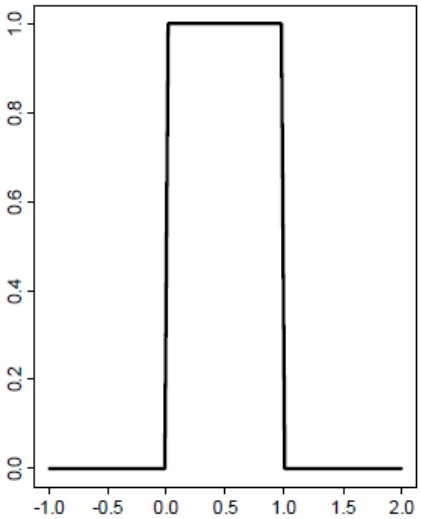
$$Var(X + c) = Var(X)$$

$$Var(cX) = c^2 Var(X)$$

一様分布

- サイコロの目のように、どの変数に対しても同じ確率値をとる確率分布を一様分布という

- 区間$[0, 1]$での一様分布は右図のようになる



コイントスの確率分布

- コイントスで表が出るか裏が出るかという試行は、ある行為が成功するか失敗するかという試行ととらえることができる

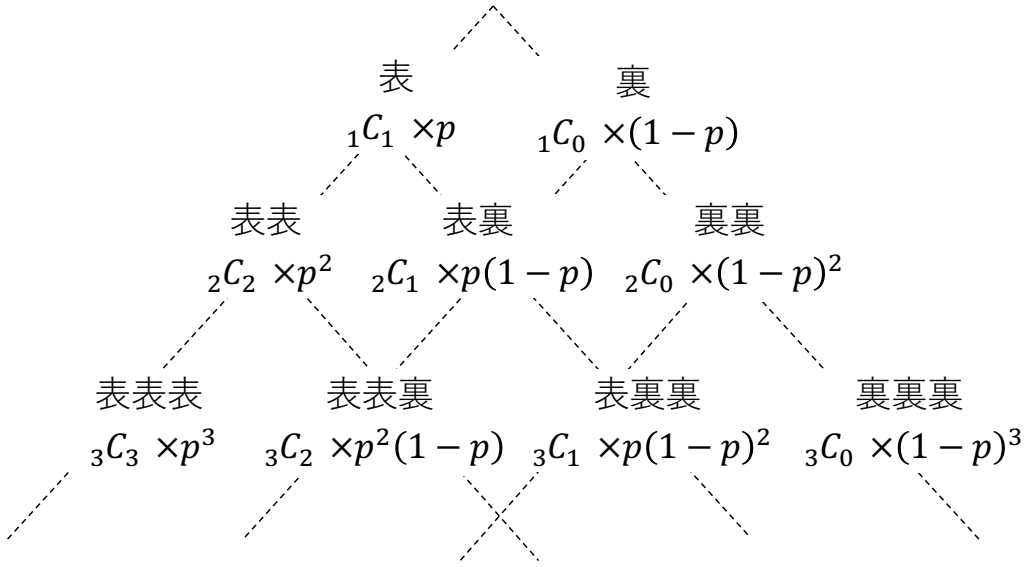
コイントス（確率変数）	表	裏
確率	1/2	1/2

ベルヌーイ試行と二項分布

- 経験的にコイントスをして表が出る確率がpだったとする

- １回目：コイントスで表と裏がでるのは ${}_1C_1$ 回と ${}_1C_0$ 回。従って表と裏が出る期待値は ${}_1C_1 \times p$及び ${}_1C_0 \times (1-p)$となる
- ２回目：コイントスで２回連続表が出る、１回目表(裏)で２回目裏(表）が出る、２回連続裏が出る期待値は、それぞれ ${}_2C_2 \times p^2$ ${}_2C_2 \times p^2$、 ${}_2C_1 \times p(1-p)$、 ${}_2C_2 \times (1-p)^2$
- ３回目：…

コイントスとパスカルの三角形



ベルヌーイ試行と二項分布

- 0か1かしかない試行において、n回の試行でs回成功し、その確率pがわかっているとき、実験が成功する期待値は以下のベルヌーイ試行に従う
- ベルヌーイ試行の確率分布を二項分布といい、その分布は次式の確率密度関数に従う

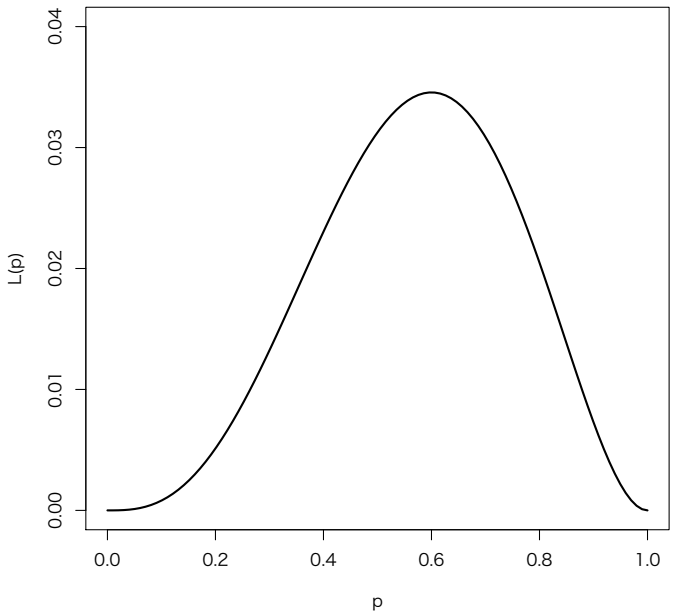
$$Binom(n,p) = {}_nC_s \cdot p^s \cdot (1-p)^{n-s} \approx p^s \cdot (1-p)^{n-s}$$

ベルヌーイ試行と二項分布

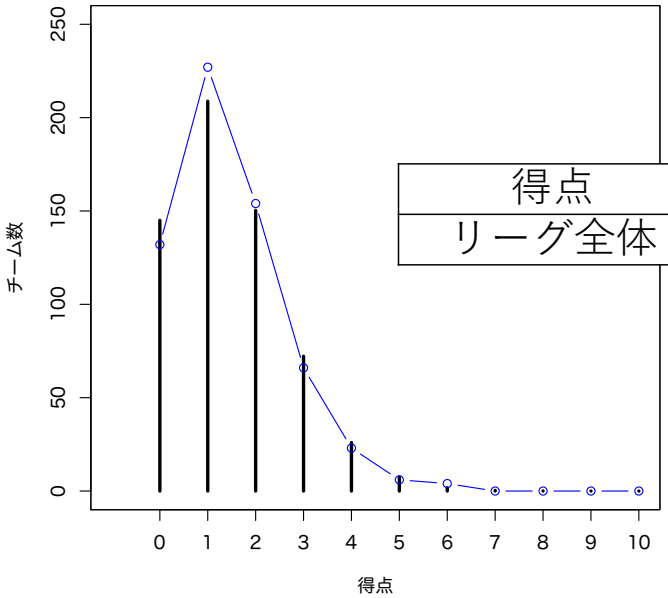
- ベルヌーイ試行の確率変数と確率値を下表のように与えた時、次式が成立する
- 平均：$\mu = np$
- 分散：$\sigma^2 = np(1-p)$

確率変数X	0	1
確率	$1-p$	p

二項分布の例



スポーツの得点分布
2013年サッカーJリーグ(J1)の例



得点	0	1	2	3	4	5	6	7
リーグ全体	132	227	154	66	23	6	4	0

平均得点＝1.44点

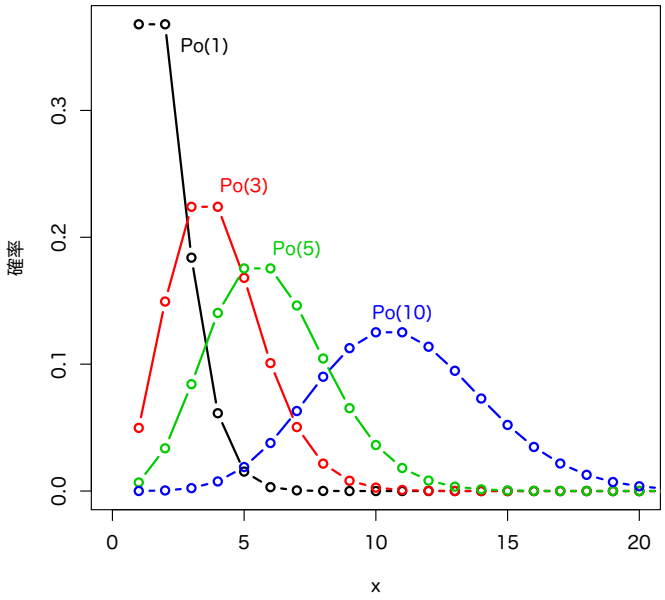
J.League Data Siteより作成

ポアソン分布

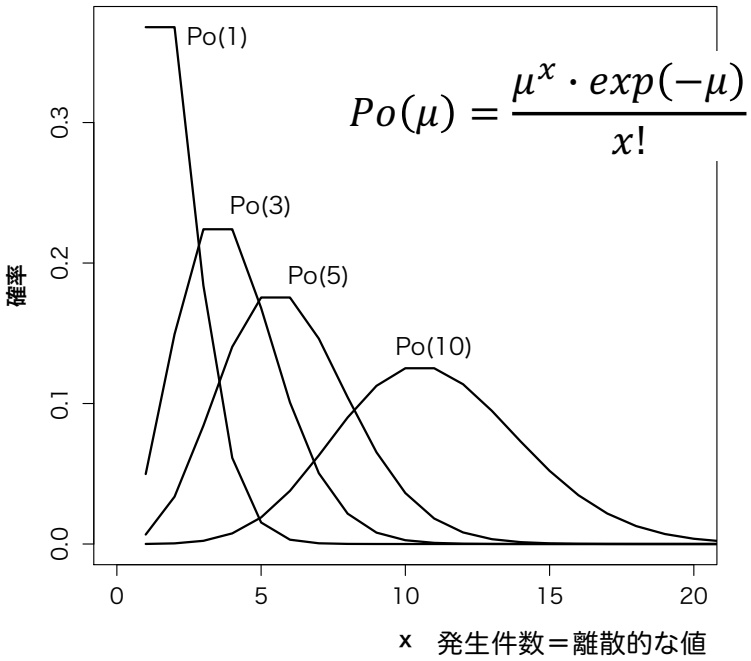
- 非常に多くの観測回数が繰り返されるものの、観測ケースの発生頻度が非常に低い場合に用いられる確率分布
- 一定範囲内（時間、回数、空間）である事象が発生する平均値μをもちいて計算される
- 試行回数n、発生頻度pとすると、 $\mu = np$

$$Po(\mu) = \frac{\mu^x \cdot exp(-\mu)}{x!}$$

ポアソン分布



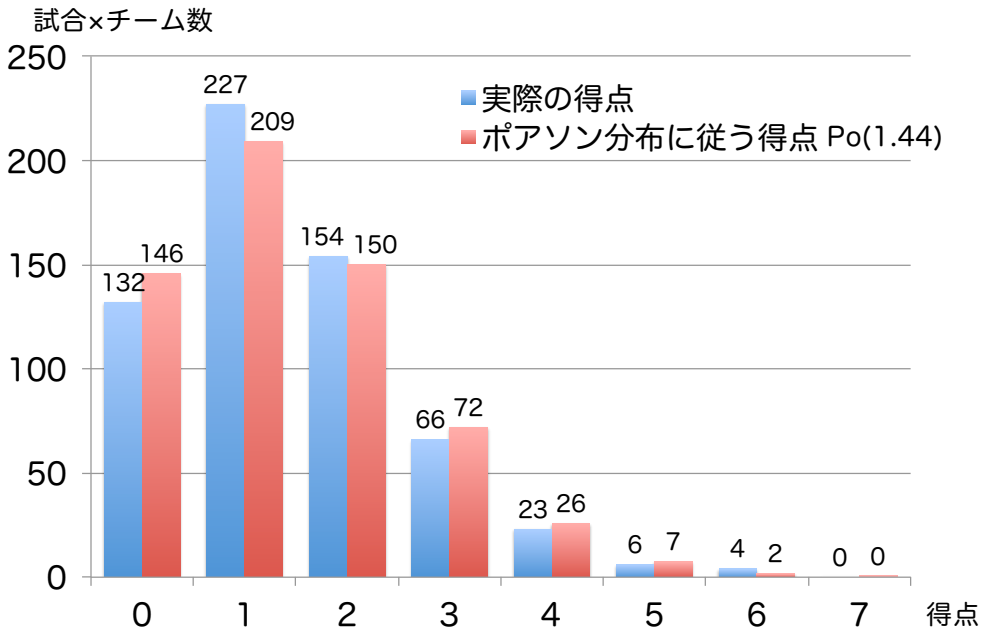
ポアソン分布



理論的な得点分布をポアソン分布で計算すると？

得点	全チーム
0	$612*1.44^0 \times \exp(-1.44)/0! \fallingdotseq 146$
1	$612*1.44^1*\exp(-1.44)/1! \fallingdotseq 209$
2	$612*1.44^2*\exp(-1.44)/2! \fallingdotseq 150$
3	$612*1.44^3*\exp(-1.44)/3! \fallingdotseq 72$
4	$612*1.44^4*\exp(-1.44)/4! \fallingdotseq 26$
5	$612*1.44^5*\exp(-1.44)/5! \fallingdotseq 7$
6	$612*1.44^6*\exp(-1.44)/6! \fallingdotseq 2$
7	$612*1.44^6*\exp(-1.44)/6! \fallingdotseq 0$

実際の得点分布と
ポアソン分布に従う（理論的な）得点分布



指数分布

- ポアソン分布と対をなす分布として指数分布がある
- 指数分布は次式で表される

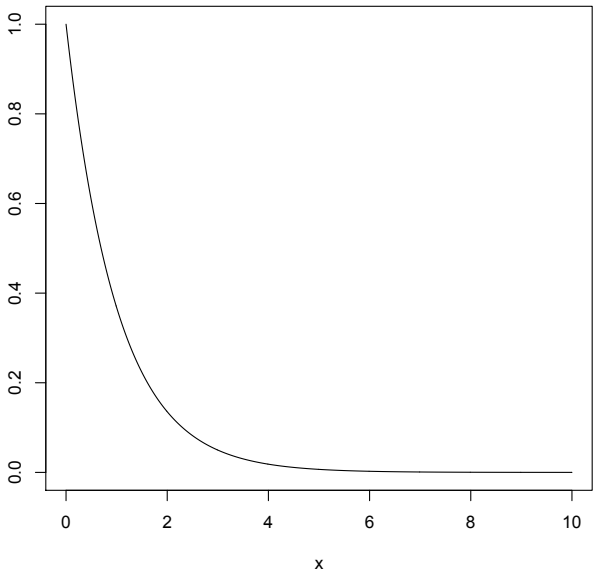
$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

- 指数分布の平均$E(X)$と分散$V(X)$は次の通り

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

指数分布



ベータ分布

- $s = \alpha - 1, n - s = \beta - 1$とすると、二項分布$Binom(n, p)$は次式のベータ関数に従うベータ分布$B(n, p)$に式変形される

$$Binom(n, p) \approx p^s \cdot (1-p)^{n-s}$$
$$= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}$$

$$B(n, p) = k \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$
$$0 < p < 1, 0 < \alpha, 0 < \beta$$

- ただしkは定数

ベータ分布

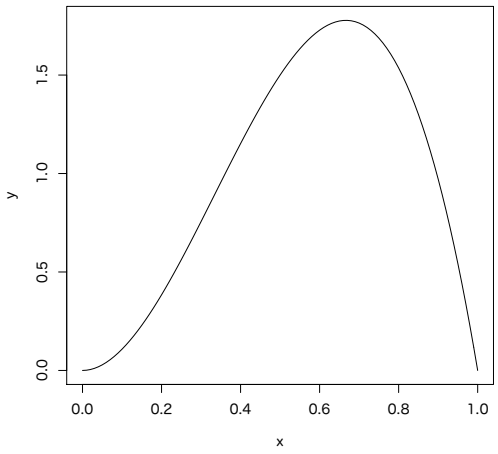
- ベータ分布$B(n, p)$の平均μと分散σ^2は以下のようになる

$$\mu = \frac{\alpha}{\alpha+\beta}$$

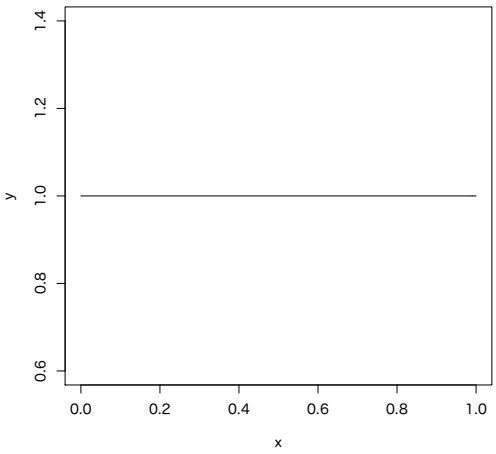
$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

ベータ分布

$B(3,2)$

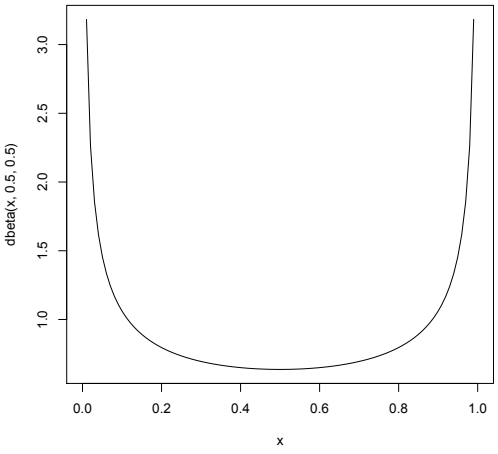


$B(3,2)$=一様分布

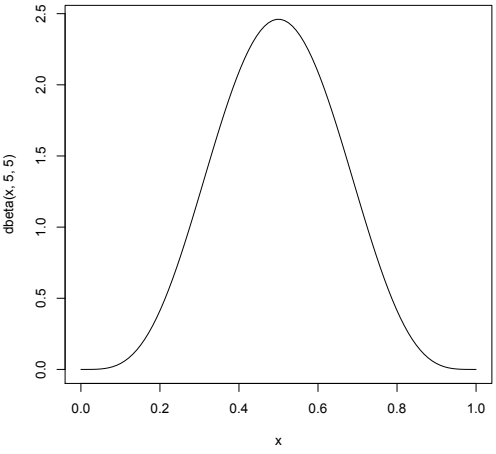


ベータ分布

$B(0.5,0.5)$



$B(5,5)$



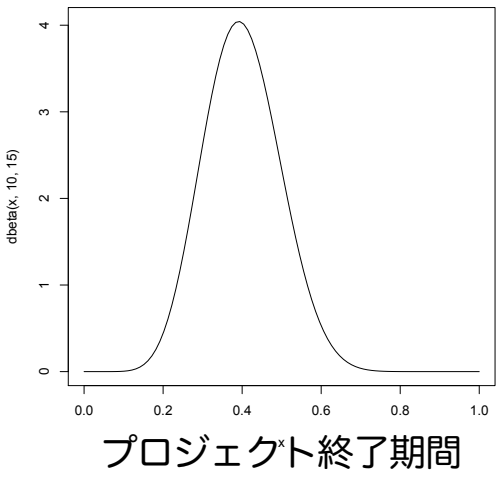
ベータ分布は万能な確率分布

- ベータ分布は以下の分布に変化できる
  - 一様分布、線形分布
  - 単調増加・単調減少分布
  - 単峰分布
  - 左右対称分布

ベータ分布

- 活用例は数少ないが、プロジェクトの時間管理などに用いられることがある

- あるプロジェクトがある期間内に終わることもあれば（確率は低いが）終了までとても時間がかかることがある



ベータ分布とガンマ分布

- ベータ関数の確率密度関数は以下のように式変形できる

$$
\begin{aligned}
Binom(n,p) &\approx p^{\alpha}\cdot(1-p)^{\beta-1} \\
&= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \\
&= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}
\end{aligned}
$$

- ここで$\Gamma(\alpha)$はガンマ関数という

ガンマ関数

- ガンマ関数$\Gamma(\alpha)$は次式で表される

$$
\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}\,dx
$$
$$
\alpha > 0
$$

- ガンマ関数$\Gamma(\alpha)$は以下のような性質を持つ

$$
\Gamma(\alpha) = (\alpha-1)!
$$
$$
\Gamma(\alpha+1) = \alpha\Gamma(\alpha)
$$
$$
\Gamma(1/2) = \sqrt{\pi}
$$

ガンマ分布

- ガンマ分布の確率密度関数$f(x)$は次式のようになる

$$f(x) = Ga(\alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$

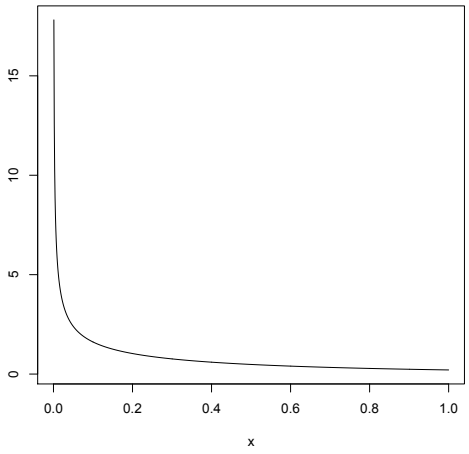
- ガンマ分布の確率変数をXとするとき、平均$E(X)$と分散$V(X)$は以下の通り

$$E(X) = \frac{\alpha}{\lambda}$$

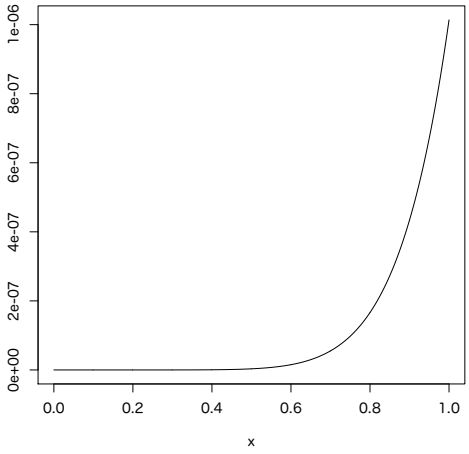
$$E(X) = \frac{\alpha}{\lambda^2}$$

ガンマ分布

$\Gamma(0.5)$



$\Gamma(10)$



正規分布（ガウス分布）

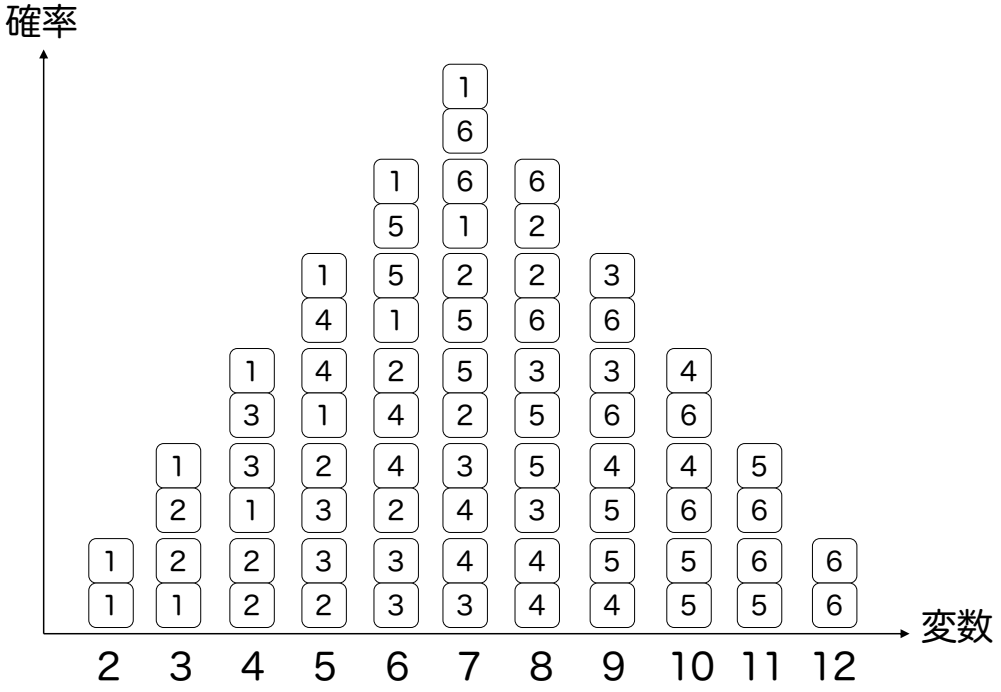
- 平均μ、分散σ^2となる確率変数xについて、以下の確率密度関数に従う分布を正規分布$N(\mu, \sigma^2)$という

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

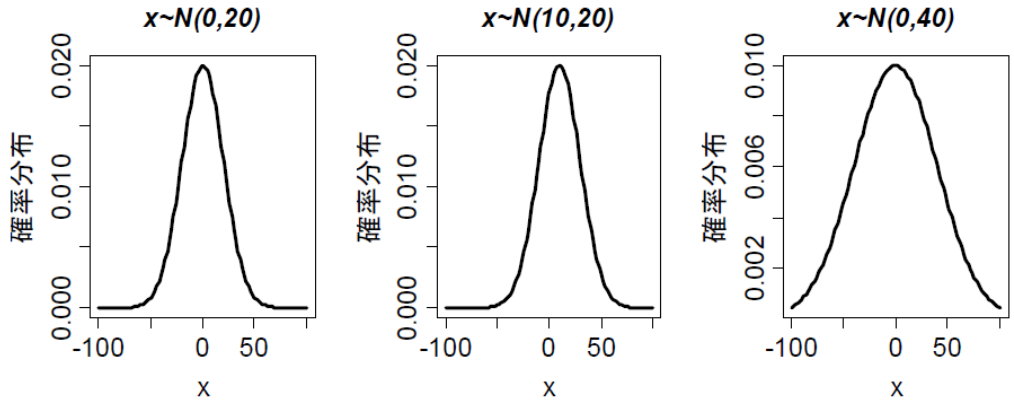
- 正規分布の平均と分散は以下のとおり

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

サイコロ２個の合計値



正規分布



標準正規分布

- 確率変数xを次式により標準化する

$$z = (x - \mu)/\sigma$$

- このとき確率変数zは、平均$\mu = 0$、分散$\sigma^2 = 1$となる標準正規分布$N(0,1)$となる
- 標準正規分布$N(0,1)$の確率密度関数$f(z)$は以下のように表せる

$$N(0,1) = f(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right)$$

二項分布と正規分布

- 二項分布 $Binom(n,p) \approx p^n \cdot (1-p)^{n-s}$
- 二項分布の試行回数nが十分に大きい時、正規分布N(np,np(1-p))に近似できる

$$N\left(np, np(1-p)\right)$$

$$= \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(x-np)^2}{2np(1-p)}\right)$$

χ^2分布

- 確率変数X_iが平均μ、分散σ^2の正規分布に従う$X_i \sim N(\mu, \sigma^2)$とき、その標準化された値$Z_i = \frac{X_i - \mu}{\sigma}$は標準正規分布に従う

$$Z_i \sim N(0,1)$$

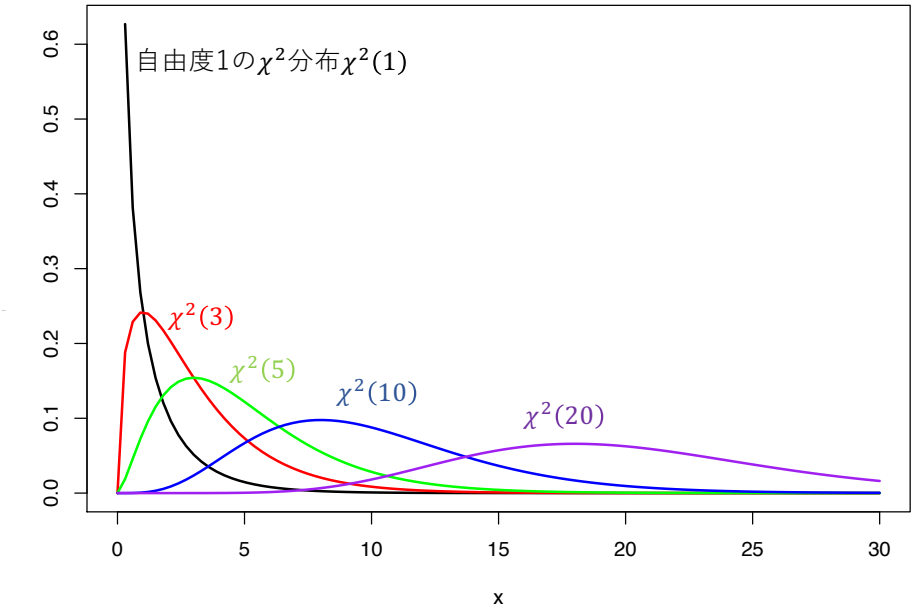
- このときZ_i^2は自由度1のχ^2分布$\chi^2(1)$に従うことが知られている

$$Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$$

- $W = \sum_{i=1}^{n} Z_i^2$は自由度nのχ^2分布$\chi^2(n)$に従う

$$W = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

χ^2分布



χ^2分布

- 自由度kのχ^2分布$\chi^2(k)$の確率密度分布$f(x;k)$は$0 \leq x$の範囲で次式で表すことができる($x < 0$のときは0)

$$f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}$$

- ここで$\Gamma(\alpha)$はガンマ関数
- 自由度kのχ^2分布の確率密度関数について、期待値と分散はそれぞれkおよび$2k$となる

ガンマ分布とχ^2分布

- ガンマ分布$Ga(k/2\,,2)$ は自由度kのχ^2分布$\chi^2(k)$となる

$$Ga\left(\frac{k}{2},2\right)=\chi^2(k)$$

- ここで、$x=z^2$とすると、ガンマ分布$Ga(1/2\,,1/2)$は標準正規分布に式変形できる

$$Ga\left(\frac{1}{2},\frac{1}{2}\right)=\frac{1}{\sqrt{2\pi z^2}}exp\left(-\frac{z^2}{2}\right)$$