データ分布と予測

堀田　敬介

確率変数
確率分布

2006/10/27,Fri. – 11/3,Fri.

試行とは？

- 試行
  - 何かの行為により「偶然による」ひとつの結果を導き出す

〔例〕

さいころ投げ

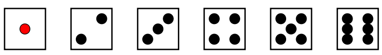
コイン投げ

〔例〕身長の測定，じゃんけん，宝くじを買う，アンケート調査，製品品質検査，etc.

確率変数とは？

試行してみないと何が出るかはわからない！
とりうる値はわかっている．

- 確率変数 random variable
  - それがとる値に対し確率が与えられている変数
    - 例：さいころ投げ



試行結果

$X =$	1	2	3	4	5	6
$P(X = x_k) =$	1/6	1/6	1/6	1/6	1/6	1/6

確率変数の値

確率

確率変数とは？

- 例：コイン投げ

試行結果

$X =$	表	裏
$P(X = x_k) =$	1/2	1/2

確率変数の値

確率

- 一般に，確率変数の確率は以下のように表現される

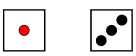
$$P(X = x_k) = p_k \quad (k = 1,2,\cdots)$$

ただし，$p_k \geq 0 \, (k = 1,2,\cdots)$, $\sum_{k=1}^{\infty} p_k = 1$ である．

確率はすべて0以上　　全ての確率を足すと1

演習１

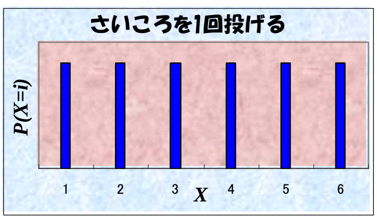
- 確率変数
  - 2個のさいころA, Bを振り出た目の差（Aの目ーBの目）を考える．この確率変数 X のとる値と，その値が出る確率を求めよ．
    - 例）Aが1で，Bが3の時，1-3 = -2



確率分布 probability distribution

- 確率分布
  - 例：さいころを1回投げる

X	1	2	3	4	5	6
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

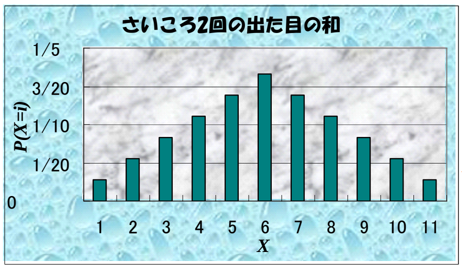


一様分布

確率分布 probability distribution

- 確率分布
  - 例：さいころを2回投げたときの出た目の和

X	2	3	4	5	6	7	8	9	10	11	12
P(X=i)	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36



三角分布

実は
二項分布

確率分布 probability distribution

- 離散（型）確率分布 discrete distribution
  - 可算集合 $\{x_1, x_2, \ldots\}$ の中の値を取る確率変数 X は離散型 discrete type といわれる．このとき，それぞれの値の確率

$$f(x_k) := P(X = x_k) \quad (k = 1, 2, \cdots)$$

を X の確率分布 probability distribution という．

ただし，$\begin{cases} f(x_k) \geq 0 \quad (k = 1, 2, \cdots), \\ \sum_{k=1}^{\infty} f(x_k) = 1 \end{cases}$

確率分布
probability distribution

一般的な定義

確率変数の期待値・分散

- 期待値 expectation, expected value
  - 確率変数 X の期待値
    - 例：コインを3回投げて表が出る回数の確率分布

X	0	1	2	3
P(X=i)	1/8	3/8	3/8	1/8

    - その期待値

$$E(X) = (0 \times \frac{1}{8}) + (1 \times \frac{3}{8}) + (2 \times \frac{3}{8}) + (3 \times \frac{1}{8}) = \frac{3}{2}$$

コインを3回投げると，
平均して1.5回表が出る
ことが期待される

  - （離散型）確率変数Xの期待値

$$E(X) = \sum_x x \cdot f(x)$$

演習2

宝くじに関する洒落
LOTTERY: a tax on people who are bad at math

- 期待値を求めよう
  - 宝くじの期待値

H18年オータムジャンボ宝くじ
（新市町村振興 第511回全国自治宝くじ）1億3千万枚限定販売
〔1千万枚あたりの当たり本数〕

1等	1億5000万円	×2本
前後賞	2500万円	×4本
組違賞	10万円	×198本
2等	1000万円	×2本
3等	100万円	×20本
4等	5万円	×3000本
5等	1万円	×20,000本
6等	3000円	×100,000本
7等	300円	×1,000,000本

確率変数の期待値・分散

- 期待値の基本法則
  - スカラー倍の期待値

$$E(aX) = aE(X)$$

証明：

$$E(aX) = \sum_x ax \cdot f(x) = a\sum_x xf(x) = aE(X)$$

  - 例：さいころを振って出た目の1000倍円もらえる．

aX	1000	2000	3000	4000	5000	6000
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

$$E(aX) = \frac{1}{6} \times 1000 + \frac{1}{6} \times 2000 + \cdots + \frac{1}{6} \times 6000 = 3500$$

X	1	2	3	4	5	6
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

$$aE(X) = 1000\left(\frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \cdots + \frac{1}{6} \times 6\right) = 3500$$

確率変数の期待値・分散

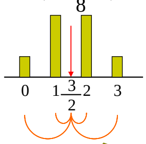
- 分散 variance
  - 確率変数 X の分散

分散（ばらつき）
＝平均（期待値）からのずれ
（の2乗）の平均（期待値）

$$V(X) = E(\{X - E(X)\}^2) = E(X^2) - E(X)^2$$

    - 例：コインを3回投げて表が出る回数の分布の分散は?

$$V(X) = (0 - 1.5)^2 \times \frac{1}{8} + (1 - 1.5)^2 \times \frac{3}{8} + (2 - 1.5)^2 \times \frac{3}{8} + (3 - 1.5)^2 \times \frac{1}{8} = \frac{3}{4}$$



  - （離散型）確率変数Xの分散

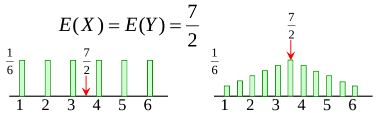
$$V(X) = \sum_x \left(x - E(X)\right)^2 f(x)$$

どの程度
散らばっているか？

確率変数の期待値・分散

- 分散は何故必要か？
  - 例：確率変数Xを，さいころを1回振ったときの目，確率変数Yを，さいころを2回振ったときの目の平均としたとき，それぞれの期待値を求めよ．

$$E(X)=E(Y)=\frac{7}{2}$$



期待値は等しいが，分布の形状は異なる．
期待値は確率変数の重要な指標だが，性質の全てではない！

- 例題のそれぞれの分散の値を求めよ．

$$V(X)\approx 2.917,\quad V(Y)\approx 1.458$$

確率変数の期待値・分散

- 分散の基本法則
  - スカラー倍の分散

$$V(aX)=a^2V(X)$$

証明：

$$V(aX)=E((aX)^2)-E(aX)^2 = a^2E(X^2)-\{aE(X)\}^2 = a^2\{E(X^2)-E(X)^2\}=a^2V(X)$$

- 例：さいころ1個を振り，出目の1000倍円貰える．分散は？

aX	1000	2000	3000	4000	5000	6000
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

確率分布

$$V(aX)=\frac{1}{6}\times(1000-3500)^2+\frac{1}{6}\times(2000-3500)^2+\cdots+\frac{1}{6}\times(6000-3500)^2 = 2{,}916{,}666.66\cdots$$

もし「576倍円貰える」だったらどちらが計算が楽か？

X	1	2	3	4	5	6
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

$$aV(X)=1000^2\left\{\frac{1}{6}\times(1-3.5)^2+\frac{1}{6}\times(2-3.5)^2+\cdots+\frac{1}{6}\times(6-3.5)^2\right\} = 1000^2\times 2.9166\cdots = 2{,}916{,}666.66\cdots$$

確率変数の期待値・分散

- 標準偏差 standard deviation
  - 確率変数Xの標準偏差

$$D(X)=\sqrt{V(X)}$$

- 例：コインを3回投げて表が出る回数の分布の標準偏差は?

$$D(X)=\sqrt{V(X)}=\sqrt{\frac{3}{4}}=\frac{\sqrt{3}}{2}$$

補足：確率変数の歪度・尖度

- 歪度 skewness
  - 確率変数Xの確率分布の非対称性の指標

$$\alpha_3=E(X-\mu)^3/\sigma^3 \qquad \left(\mu:=E(X),\ \sigma^2:=V(X)\right)$$

$$\begin{cases}\alpha_3>0 \ldots \text{右の裾が長い}\\ \alpha_3<0 \ldots \text{左の裾が長い}\\ |\alpha_3| \quad \ldots \text{歪みの程度}\end{cases}$$

- 尖度 kurtosis（超過係数 coefficient of excess）
  - 確率変数Xの確率分布の尖り具合を表す指標

$$\alpha_4=E(X-\mu)^4/\sigma^4$$

正規分布は α_4=3なので，これと比較

$$\begin{cases}\alpha_4-3>0 \ldots \text{正規分布より尖っている}\\ \alpha_4-3<0 \ldots \text{正規分布より丸く鈍い形}\end{cases}$$

演習2

- 確率分布を求めよう
  - コインを5回投げて裏が出る回数の確率分布を求めよ．
  - 求めた確率分布をグラフに描画せよ．
  - 期待値を計算しよう．
  - 分散・標準偏差を計算しよう．
  - 歪度・尖度を計算しよう．

Coffee Break!
1から100まで足すといくつ？

Q1. 1から100までの和は？

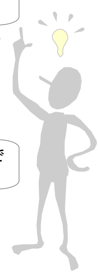
$$1+2+\cdots+100=?$$

326から579までの和は？

Q2. 1から100までの2乗和は？

$$1^2+2^2+\cdots+100^2=?$$

42から283までの2乗和は？



一般に，1からnまでの和と2乗和

$$\sum_{k=1}^{n}k=1+2+\cdots+n=\frac{n(n+1)}{2}$$

$$\sum_{k=1}^{n}k^2=1^2+2^2+\cdots+n^2=\frac{1}{6}n(n+1)(2n+1)$$

$$\left[\begin{array}{l}\because)\ by\ (k+1)^3-k^3=3k^2+3k+1\\ \quad\sum_{k=1}^{n}\{(k+1)^3-k^3\}=\sum_{k=1}^{n}\{3k^2+3k+1\}\end{array}\right]$$

確率分布
probability distribution

離散（型）分布 discrete distribution

- ★（離散）一様分布 uniform distribution
- ★ベルヌーイ分布 Bernoulli distribution
- ★二項分布 binomial distribution
- ★ポアソン分布 Poisson distribution
- ★幾何分布 geometric distribution
- ★負の二項分布 negative binomial distribution
- ★超幾何分布 hypergeometric distribution

離散型分布 discrete distribution

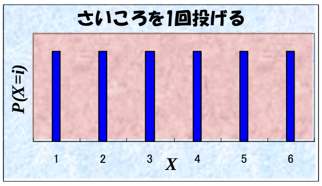
- (離散)一様分布 uniform distribution (of discrete type)
  - すべての確率が等しい分布
    - 例：さいころを1回投げる

X	1	2	3	4	5	6
P(X=i)	1/6	1/6	1/6	1/6	1/6	1/6

  - 確率分布・期待値・分散

$$f(x)=\frac{1}{n},\ (x=1,\cdots,n)$$

$$\begin{cases} E(X)=\dfrac{n+1}{2}, \\ V(X)=\dfrac{(n^2-1)}{12} \end{cases}$$

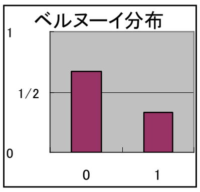


離散型分布 discrete distribution

- ベルヌーイ分布 Bernoulli distribution
  - 試行の結果が2通りしかない確率分布

x	0	1
$P(x)$	p	q

    - 例：コインを1回投げる
      - 表：1/3, 裏：2/3 で出るコインは…

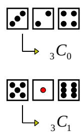


- ベルヌーイ試行
  - 2通りの結果しかない観測があり，これを同条件で独立にn回行うこと．

離散型分布 discrete distribution

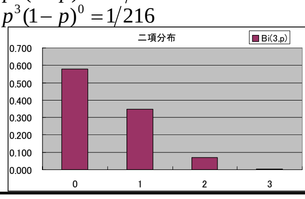
- 二項分布 binomial distribution
  - ベルヌーイ試行で，一つの結果が起こる回数の確率
  - 確率pをもつ事象がn回の施行中x回起こる確率
    - 例：サイコロを3回投げて1の目がx回出る確率は…

0回出る確率 … ${}_3C_0 p^0(1-p)^3=125/216$
1回出る確率 … ${}_3C_1 p^1(1-p)^2=75/216$
2回出る確率 … ${}_3C_2 p^2(1-p)^1=15/216$
3回出る確率 … ${}_3C_3 p^3(1-p)^0=1/216$
$(p=1/6)$



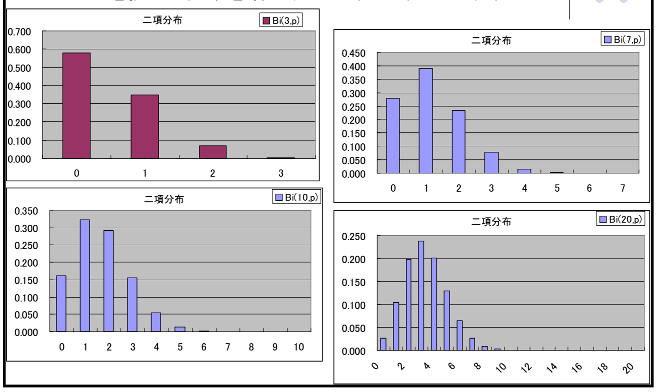
1が0回出る → $\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3$ 3箇所に0個1を置く → ${}_3C_0$

1が1回出る → $\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2$ 3箇所に1個1を置く → ${}_3C_1$



離散型分布 discrete distribution

■サイコロを投げる回数を増やすと1の目がx回出る確率は…



離散型分布 discrete distribution

- 二項分布
  - 確率pをもつ事象がn回の施行中x回起こる確率

$$\boxed{Bi(n,p)}$$

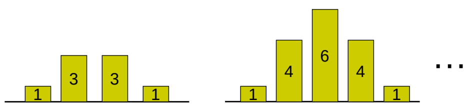
  - 確率分布・期待値・分散

$$f(x)={}_nC_x p^x(1-p)^{n-x},\ (x=0,\cdots,n)$$

$$\begin{cases} E(X)=np, \\ V(X)=np(1-p) \end{cases}$$

Coffee Break! パスカルの三角形と二項係数

$${}_3C_0=1 \qquad {}_4C_0=1$$
$${}_3C_1=\frac{3}{1}=3 \qquad {}_4C_1=\frac{4}{1}=4$$
$${}_3C_2=\frac{3\cdot 2}{2\cdot 1}=3 \qquad {}_4C_2=\frac{4\cdot 3}{2\cdot 1}=6 \qquad \cdots$$
$${}_3C_3=\frac{3\cdot 2\cdot 1}{3\cdot 2\cdot 1}=1 \qquad {}_4C_3=\frac{4\cdot 3\cdot 2}{3\cdot 2\cdot 1}=4$$
$${}_4C_4=\frac{4\cdot 3\cdot 2\cdot 1}{4\cdot 3\cdot 2\cdot 1}=1$$



Coffee Break! パスカルの三角形と二項係数

1	
1 1 ⟷	${}_1C_0 \quad {}_1C_1$
1 2 1 ⟷	${}_2C_0 \quad {}_2C_1 \quad {}_2C_2$
1 3 3 1 ⟷	${}_3C_0 \quad {}_3C_1 \quad {}_3C_2 \quad {}_3C_3$
1 4 6 4 1 ⟷	${}_4C_0 \quad {}_4C_1 \quad {}_4C_2 \quad {}_4C_3 \quad {}_4C_4$
1 5 10 10 5 1 ⟷	${}_5C_0 \quad {}_5C_1 \quad {}_5C_2 \quad {}_5C_3 \quad {}_5C_4 \quad {}_5C_5$
⋮	⋮

$(a+b)^n$ の各項の係数

$10=1+3+6$

$10=1+2+3+4$

☆ 組合せ数の和法則

$${}_nC_k={}_{n-1}C_{k-1}+{}_{n-1}C_k \quad (\text{for } 1\le k\le n-1)$$

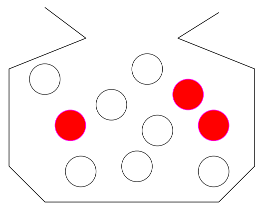
☆ $${}_nC_k=\sum_{n-k-1\le m\le n-1} {}_mC_{m+k+1-n} = \sum_{k-1\le m\le n-1} {}_mC_{k-1}$$

離散型分布 discrete distribution

- 二項分布の例
  - 例1：製品ラインの不良品抜き取り検査
    - 不良率p=0.5%のロットから独立に1個ずつランダム抜取り検査をした時に検出される不良品数x の従う分布
    - 参考）不良率の期待値 $E(x/n)=np/n=p$
  - 例2：袋から球を取り出す
    - 赤玉3，白玉7入っている袋から1つ取り出しては戻すという行為をn回行ったとき，赤玉が5回出る確率は?
  - 例3：サイコロをn回投げて偶数の目が出る回数の従う分布
    - サイコロを5回投げて偶数が出る回数の確率分布を求めよ

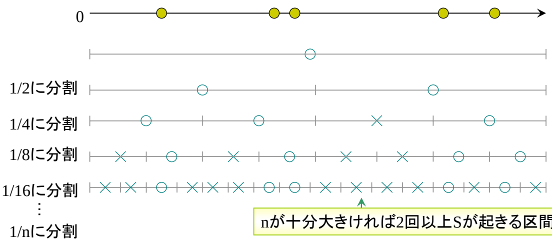
演習3

- 二項分布を求めよう
  - 赤玉3個，白玉7個入っている袋から1つ取り出しては戻すという行為を5回行ったとき，赤球が出る回数の確率分布（二項分布）を求めてみよう！



離散型分布 discrete distribution

- ポアソン分布 Poisson distribution
  - ある時間帯の中で，ある事象が何回起きるか？
    - 例：電話番号案内
      - 事象S＝「通話がある」

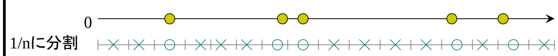


nが十分大きければ2回以上Sが起きる区間が無くなる

⇨ P(ある区間でSが2回以上起きる確率)=0とする

離散型分布 discrete distribution

- ポアソン分布 Poisson distribution



各区間でSが起きる確率　：p
各区間でSが起きない確率：q=1-p
とする

ベルヌーイ試行とみなす

⇨ Sが起きる回数は二項分布 $Bi\,(n, p)$ に従う

確率pの事象がn回の試行の中でS回起こる

ところで，この時間内にSが起きる回数の期待値をλとおくと…

$\lambda = np$　（二項分布の期待値より）

よって，Sがk回起きる確率は，

$$\begin{aligned}&{}_nC_k\, p^k (1-p)^{n-k}\\ &={}_nC_k\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &=\frac{\lambda^k}{k!}\left((1-\frac{1}{n})\cdots\cdots(1-\frac{k-1}{n})\right)(1-\frac{\lambda}{n})^{-k}(1-\frac{\lambda}{n})^n \underset{n\to\infty}{\to} e^{-\lambda}\frac{\lambda^k}{k!}\end{aligned}$$

$$\left(\because \lim_{n\to\infty}(1-\frac{\lambda}{n})^n=e^{-\lambda}\right)$$

離散型分布 discrete distribution

- ポアソン分布 Poisson distribution
  - 2項分布においてある事象が起こる確率が非常に小さい場合に適用できる分布
    - 例：工場の生産ラインでの不良率が1/500のとき，1000個の製品を作ったときx個不良品だった

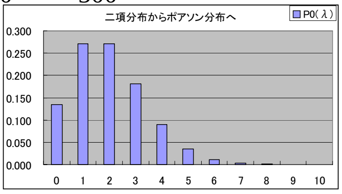
$$f(x) = {}_{1000}C_x \left(\frac{1}{500}\right)^x \left(1-\frac{1}{500}\right)^{1000-x}$$

二項分布

ポアソン分布

$$f(x) = e^{-2}\frac{2^x}{x!}$$

1000個のうち，平均的に2個不良品



離散型分布 discrete distribution

- 二項分布からポアソン分布へ

ポアソンの小数の法則
Poisson's law of small numbers

$np = \lambda$

$${}_nC_x p^x (1-p)^{n-x} \to e^{-\lambda}\frac{\lambda^x}{x!} \quad \begin{pmatrix} n\to\infty, \\ p\to 0 \end{pmatrix}$$

二項分布　ポアソン分布

- 例：工場の生産ラインでの不良率が1/500のとき，1000個の製品を作ったときx個不良品だった

二項分布	ポアソン分布
$f(0)=0.135065,$	$f(0)=0.135335,$
$f(1)=0.270670,$	$f(1)=0.270671,$
$f(2)=0.270942$	$f(2)=0.270671,$
$f(3)=0.180628,$	$f(3)=0.180447,$
$f(4)=0.090223,$	$f(4)=0.090224,$
$\vdots$	$\vdots$

離散型分布 discrete distribution

- 二項分布とポアソン分布
  - 例：単一時間に発生する事故件数は?
    - 一日 m件の事故が発生したとする．これを1時間毎，1分毎，1秒毎と縮めていき，1刻みに1件の事故が発生するようにし(同時刻に2件発生することはないとする)，その刻み数をnとする．
    - すると，この話は n個の刻みの中で 1件事故が発生するかしないかとみなすことができる．即ち，n個の刻みの中から事件が発生した x個の刻みの個数を考えることになる
    - 事故発生率は $p = m / n =$ 一定！

      (ポアソン分布の期待値)
    - このときの $m = n\,p$ は二項分布の期待値！

離散型分布 discrete distribution

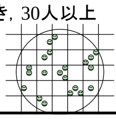
- ポアソン分布 $Po(\lambda)$
  - 確率分布・期待値・分散

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},\ (x = 0,1,2,\cdots)$$

$$\begin{cases} E(X) = \lambda\ \ (= np), \\ V(X) = \lambda \end{cases}$$

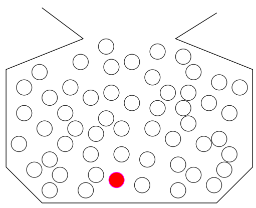
離散型分布 discrete distribution

- ポアソン分布の例
  - 例1：飛行機事故（事故はめったに起きない）
    - 飛行機事故の確率1/10万，飛行機搭乗回数を1万回としたとき，一度も事故にあわない確率は?
  - 例2：大量生産品の不良品数（めったにない）
    - 不良率が1/10000の生産ラインで1万個生産したとき不良品が3個以上出る確率は?
  - 例3：爆撃命中数（めったに当たらない）
    - 第二次大戦中のドイツ軍の砲弾命中精度は λ =0.93のポアソン分布に従うという．1000発打って1発当たる確率は?
  - 例4：薬の副作用
    - 副作用の確率が1/200の薬を5000人が服用したとき，30人以上に副作用が出る確率は?
  - 例5：生物・植物の生態・繁茂状況を示す分布
    - 単位面積あたりのバクテリアの個数



演習4

- ポアソン分布を求めよう
  - 赤玉1個，白玉99個入っている袋から1つ取り出しては戻すという行為を5回行ったとき，赤球が出る回数の確率分布（ポアソン分布）を求めてみよう！

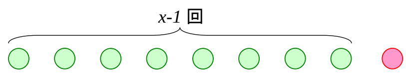


離散型分布 discrete distribution

- 幾何分布 geometric distribution
  - ベルヌーイ試行において，試行回数を決めずに初めてある事象が起こるまでの試行回数をxとすると…

$$f(x) = p(1-p)^{x-1},\ (x = 1,2,\cdots)$$

幾何数列（等比数列）の形なので，幾何分布とよばれる



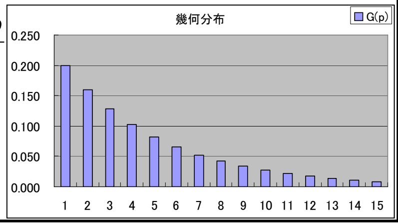
- 幾何分布は，時間を離散的に(1,2,3,…)考えるとき，初めて何かが起こるまで<u>待つ時間の長さの確率分布</u>である〔（離散的な）待ち時間分布〕

離散型分布 discrete distribution

- 幾何分布
  - 確率分布・期待値・分散

$$f(x) = p(1-p)^{x-1},\ (x = 1,2,\cdots)$$

$$\begin{cases} E(X) = \dfrac{1}{p}, \\ V(X) = \dfrac{1-p}{p^2} \end{cases}$$

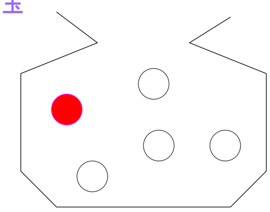


離散型分布 discrete distribution

- 幾何分布の例
  - 例1：災害の到来
    - ある1年に風水害が起こる確率が1/25であるとする．風水害が起こるのは平均何年に1回か?
    - 上記と同じ災害が20年以内に起こる確率は?
  - 例2：袋から…
    - 白玉4つ，赤玉1つが入っている袋がある．1つ取り出して元に戻すという試行を繰り返したとき，10回目に初めて赤玉が取り出される確率は?
  - 例3：ドアを開けられる鍵を見つけよう！
    - n個の鍵束を持っている．かぎ束からひとつ鍵を取り出しドアを開けるとき，何回目で開くか?　ただし，試した鍵は1回毎に鍵束に戻すこととする

演習5

- 幾何分布を求めよう
  - 白玉4つ，赤玉1つが入っている袋がある．1つ取り出して元に戻すという試行を繰り返したとき，初めて赤玉が取り出される回の確率分布（幾何分布）を求めてみよう！　10回目に初めて赤玉が取り出される確率はどれだけか？

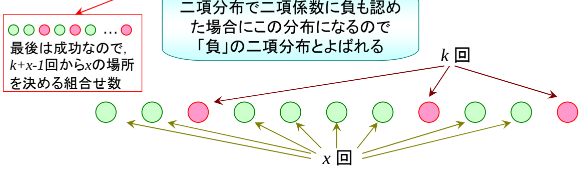


離散型分布 discrete distribution

- 負の二項分布 negative binomial distribution
  - ある事象がk回起こるまでのもうひとつの事象の回数をxとしたときのX=x (x=0,1,2,...)の従う分布

$$f(x) = {}_{k+x-1}C_x\, p^k (1-p)^x,\ (x = 0,1,2,\cdots)$$

二項分布で二項係数に負も認めた場合にこの分布になるので「負」の二項分布とよばれる



- k=1のときは<u>幾何分布に等しい</u>ため幾何分布の一般化となっている

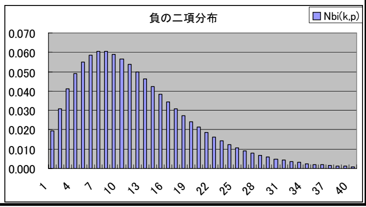
離散型分布 discrete distribution

- 負の二項分布
  - 確率分布・期待値・分散

$$f(x) = {}_{k+x-1}C_x\, p^k (1-p)^x,\ (x = 0,1,2,\cdots)$$

幾何分布のk倍！

$$\begin{cases} E(X) = \dfrac{k(1-p)}{p}, \\ V(X) = \dfrac{k(1-p)}{p^2} \end{cases}$$

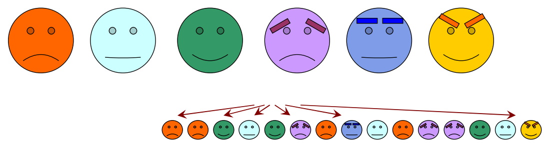


離散型分布 discrete distribution

- 負の二項分布の例
  - 例1：災害の到来
    - ある1年に風水害が起こる確率が1/25であるとする．風水害が起こるのは平均何年に1回か?
    - 上記と同じ災害が20年以内に起こる確率は?
  - 例2：袋から…
    - 白玉4つ，赤玉1つが入っている袋がある．1つ取り出して元に戻すという試行を繰り返したとき，赤玉が3回取り出されるまでに白玉が40回取り出される確率は?
  - 例3：シリーズものコレクター
    - 12種類のキャラクターが売られている．ただし，箱を開けるまで中にどれが入っているかはわからない．あるコレクターが全てのキャラクターを集めるためには何個買わねばならないか?

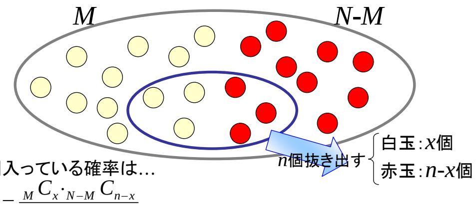
演習6

- 負の二項分布を求めよう
  - お菓子の付録として，6種類のキャラクターがある．ただし，箱を開けるまで中にどれが入っているかはわからない．全てのキャラクターを集めるためには，お菓子を平均何個買わねばならないか？ 負の二項分布を求め，計算しよう！



離散型分布 discrete distribution

- 超幾何分布 hypergeometric distribution
  - 例：白玉がM個，赤玉がN-M個（全部でN個）ある．ここからn個抜き出したとき，白玉がx個入っている確率は?



白玉がx個入っている確率は…

$$f(x)=\frac{{}_M C_x \cdot {}_{N-M} C_{n-x}}{{}_N C_n}$$

n個取り出す組合せのうち，白玉x個，赤玉n-x個取り出す組合せの確率

ただし，xの取り得る範囲は

$$(x=\max\{0, n-(N-M)\},\cdots,\min\{n, M\})$$

離散型分布 discrete distribution

- 超幾何分布
  - 確率分布・期待値・分散

$$f(x)=\frac{{}_M C_x \cdot {}_{N-M} C_{n-x}}{{}_N C_n}$$

$$(x=\max\{0, n-(N-M)\},\cdots,\min\{n, M\})$$

$$\begin{cases} E(X)=\dfrac{Mn}{N}, \\ V(X)=\dfrac{Mn}{N}\cdot\dfrac{N-M}{N}\cdot\dfrac{N-M}{N-1} \end{cases}$$

離散型分布 discrete distribution

- 超幾何分布の性質
  - 非復元抽出（とったものを戻さない）の時に現れる分布
  - 復元抽出の場合，$M/N=p$ とした二項分布となる
  - $N\to\infty$の場合，条件 $M/N\to p$ の元で二項分布となる

離散型分布 discrete distribution

- 超幾何分布の例
  - 例：資源調査「捕獲再捕獲法 capture-recapture method」

標識再捕獲法 (mark-recapture method) ともいう

  - 湖の中の魚の個体数推定など
    - 湖に何匹の魚（N匹）がいるのか知りたい!
    - 動く対象の数え上げで難しい!
    - 再捕獲により度数分布を書いてNを推定
  - 例：ある湖の中に生息している対象魚について，200匹を捕獲し標識をつけた．さてしばらく後，湖から魚を10匹獲ったとき，標識がついている魚が2匹いた．この湖にはこの魚は何匹いると推定されるか?

確率密度関数
p. d. f.
(probability density function)

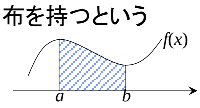
連続（型）分布continuous distribution

- ★（連続）一様分布 uniform distribution
- ★正規分布 normal distribution
- ★標準正規分布 standard normal distribution
- ★指数分布 exponential distribution
- ★ガンマ分布 Gamma distribution（χ^2分布, 指数分布）
- ★ベータ分布 Beta distribution

確率密度関数 p. d. f.

- 連続（型）分布 continuous distribution
  - 確率変数 X の取る値が関数 $f(x)$ により, 以下で与えられている場合, X は連続型の確率分布を持つという

$$P(a \le X \le b) = \int_a^b f(x)dx$$



ただし,

$$\begin{cases} f(x) \ge 0 (-\infty \le x \le \infty), \\ \int_{-\infty}^{\infty} f(x)dx = 1 \end{cases} \longleftrightarrow \quad f(x_k) := P(X = x_k) \ (k = 1,2,\cdots) \quad \begin{cases} f(x_k) \ge 0 \ (k = 1,2,\cdots), \\ \sum_{k=1}^{\infty} f(x_k) = 1 \end{cases}$$

確率密度関数
probability density function

- 累積分布関数 c.d.f., cumulative distribution function

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt \longleftrightarrow F(x) = \sum_{u \le x} f(u)$$

確率密度関数 p. d. f.

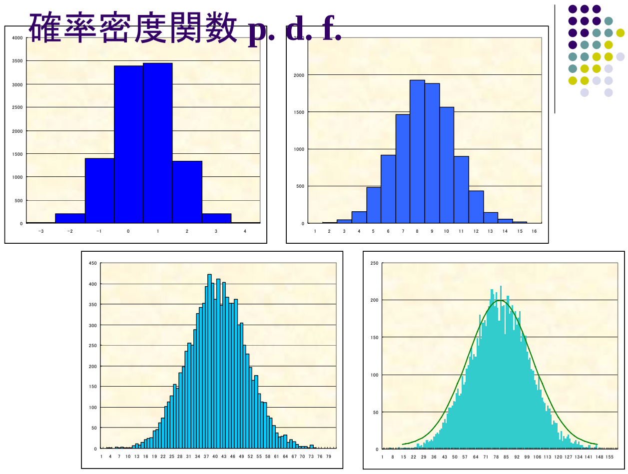
- 連続型確率変数の期待値と分散
  - 連続型確率変数 X の期待値

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \longleftrightarrow \sum_x x \cdot f(x)$$

  - 連続型確率変数 X の分散

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx \longleftrightarrow \sum_x (x - E(X))^2 f(x)$$

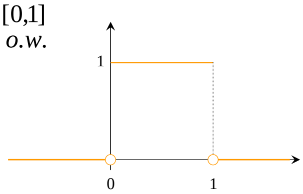
確率密度関数 p. d. f.



連続型分布 continuous distribution

- （連続）一様分布 uniform distribution
  - 確率密度関数

$$f(x) = \begin{cases} 1 & [0,1] \\ 0 & o.w. \end{cases}$$



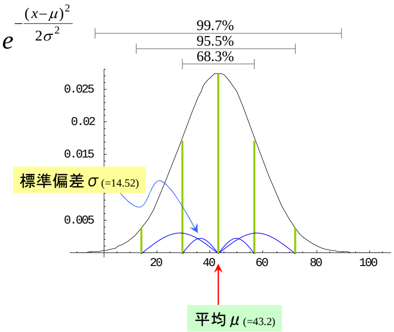
連続型分布 continuous distribution

- 正規分布 normal distribution
  - 確率密度関数

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

平均 μ , 分散 σ^2

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



連続型分布 continuous distribution

- 標準正規分布 standard normal distribution
  - 確率変数の標準化
    - 平均μ，分散σ^2の正規分布に従う確率変数Xについて

$$X \to Z = \frac{X-\mu}{\sigma}$$

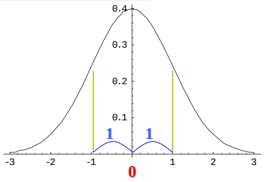
標準正規分布

⇨ 確率変数Zは，**平均0**，**分散1**の正規分布に従う

  - 確率密度関数

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

$N(0,1)$



連続型分布 discrete distribution

- 二項分布から正規分布へ…
  - 試行回数nを大きくすると，二項分布は正規分布に近づく

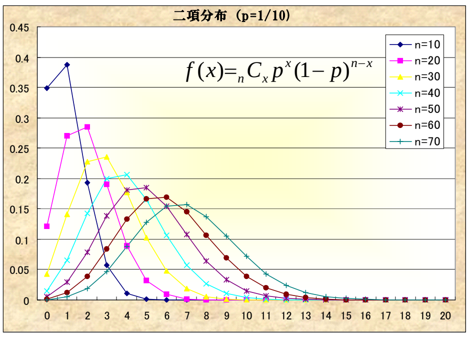
$$Bi(n,p) \xrightarrow{n\to\infty} N(\mu,\sigma^2) \quad \begin{cases}\mu = np \\ \sigma^2 = np(1-p)\end{cases}$$

  - 試行回数nが一定の時に，確率pを0.5に近づけると，二項分布は正規分布に近づく

$$Bi(n,p) \xrightarrow[p\to 0.5]{n=c,} N(\mu,\sigma^2)$$

連続型分布 discrete distribution

- 二項分布から正規分布へ…
  - 試行回数nを大きくすると，二項分布は正規分布に近づく



連続型分布 discrete distribution

- 正規分布による二項分布の近似
  - 例：内閣支持率
    - 500人の人に内閣を支持するかどうか聞いたところ，275人が指示すると答えた．

内閣支持率：$p=\frac{275}{500}=0.55$

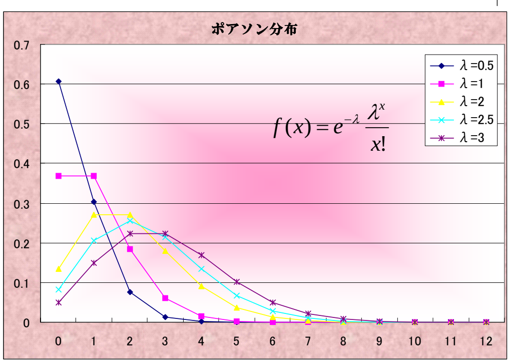
    - 内閣支持率を p (不支持率 q = 1-p) とすると，これは二項分布となる．
    - 点推定では内閣支持率は55%である．正規分布近似を考えると，$\bar{x}=np=500\times 0.55=275$

$\sigma=\sqrt{npq}=\sqrt{500\times 0.55\times 0.45}\cong\sqrt{124}\cong 11$

    - より，95%信頼区間における区間推定では，内閣支持率は $275\pm 1.96\times 11 \Leftrightarrow 253\le x\le 297$ より50.6%～59.4%

連続型分布 discrete distribution

- ポアソン分布から正規分布へ…



連続型分布 continuous distribution

- 指数分布 exponential distribution $Ex(\lambda)$

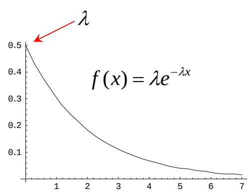
ポアソン分布に従って起きる事象の生起間隔を表現

  - 確率密度関数

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & (x\ge 0) \\ 0 & o.w.\end{cases}$$

  - 累積分布関数

$$F(x)=\begin{cases}1-e^{-\lambda x} & (x\ge 0) \\ 0 & o.w.\end{cases}$$



  - 期待値・分散

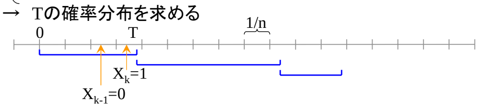
$$E(X)=\frac{1}{\lambda},\ V(X)=\frac{1}{\lambda^2}$$

連続型分布 continuous distribution

- 指数分布 exponential distribution
  - 例：サービスの待ち時間（チケット売場の列）

単位時間をn等分
ある区間kで誰かが購入を終了：$X_k=1$，そうでない：$X_k=0$
チケット販売開始時点：0，最初の人の購入終了時点：T
$X_1,X_2,X_3,X_4,\ldots$はパラメータ$p=\lambda/n$のベルヌーイ試行に従う

→ Tの確率分布を求める



区間N+1で購入終了 → $T≒(1/n)N$ （nが十分大きい時）
任意の $t>0$に対し，kを，$k≦nt<k+1$を満たす整数とする
N+1は幾何分布に従う

$$\rightarrow\ P(0\le T\le t)\approx P(0\le N\le nt)=\sum_{i=0}^{k}p(1-p)^i$$

$$=1-(1-p)^{k+1}\approx 1-(1-\lambda/n)^{nt}\xrightarrow[n\to\infty]{}1-e^{-\lambda t}=\int_0^t \lambda e^{-\lambda x}dx$$

累積分布関数

確率密度関数

連続型分布 continuous distribution

- ガンマ分布 Gamma distribution $Ga(\alpha,\lambda)$
  - 確率密度関数

$$f(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}\quad (x\ge 0)$$

  - ガンマ関数 Gamma function

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

$$\Gamma(n)=(n-1)!\qquad (n\in \boldsymbol{N})$$

nが自然数の時

⇒ $\begin{cases} Ga(n/2,1/2) & \text{：自由度 n の}\chi^2\text{分布} \\ Ga(1,\lambda) & \text{：指数分布} \end{cases}$

連続型分布 continuous distribution

- ベータ分布 Beta distribution $Be(\alpha,\beta)$
  - 確率密度関数

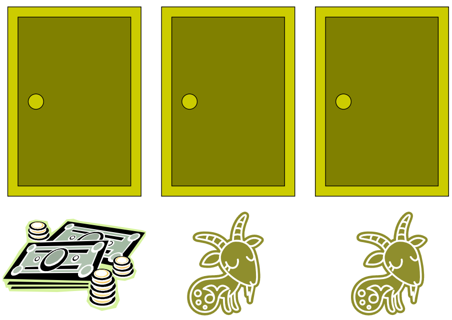
$$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\quad(0<x<1)$$

  - ベータ関数 Beta function

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx\quad(\alpha,\beta>0)$$

$$=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Coffee Break!
Monty-Hole Dilemma
確率的直感



3枚の扉の向こうに
•百万ドル（当たり）
•山羊（はずれ）
•山羊（はずれ）
が隠されているよ．あなたは扉を1つだけ選んでいいのよ．

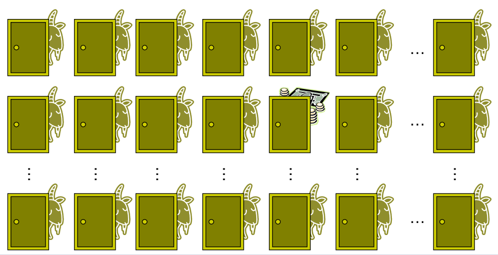
ところで，あなたが選ばなかった2つの扉のうち，山羊の扉を開くから，それを見た後で，開く扉を変えてもいいよ．

さぁ，どうする？

Coffee Break!
Monty-Hole Dilemma
確率的直感

どうしても納得いかない人のため，扉の数を増やしてみましょう！

最初に選ぶ扉が100万もあったらどうかしら？



100万の扉からあなたが1つを選んだ後で，残り99万9999の扉のうち山羊（はずれ）の99万9998の扉を開いて見せます．

それでもあなたは，最初の選択を変えない？　あなたの最初の選択は神懸かり的な幸運に恵まれているのかしら？

参考文献

- 東京大学教養学部統計学教室編 「統計学入門」東京大学出版会（1991）
- 東京大学教養学部統計学教室編 「自然科学の統計学」東京大学出版会（1992）
- 白石修二「例題で学ぶ Excel統計入門」森北出版（2001）
- 村上雅人「なるほど統計学」海鳴社（2002）
- 丹慶勝市「図解雑学 統計解析」ナツメ社（2003）
- J.Matousek, J.Nesetril / 根上生也・中本敦浩 訳「離散数学への招待 上」シュプリンガー・フェアラーク東京（2002）
- 徳山豪「工学基礎 離散数学とその応用」数理工学社（2003）
- B.Schechter / グラベルロード訳「My Brain is Open」共立出版（2003）