

# 統計学 B 1

種村 秀紀 千葉大学理学部数学・情報数理学科

〒 263-8522 千葉県千葉市稲毛区弥生町 1-33

<http://www.math.s.chiba-u.ac.jp/~tanemura/index.html>

平成 22 年 4 月 9 日

## 目次

1	記述統計	3
1.1	度数分布表	3
1.2	算術的記述	4
1.2.1	標本平均	5
1.2.2	標本標準偏差	5
1.2.3	偏差値	7
1.2.4	1 次変換を用いた計算	7
1.3	その他の記述的測度	9
1.3.1	位置の測度	9
1.3.2	変動の測度	10
2	確率空間	13
2.1	試行と確率	13
2.2	条件付確率と独立性	16
2.3	ベイズの定理 (条件付確率での原因と結果の考察)	18
3	確率分布 (離散型変数)	20
3.1	離散型確率分布の例	20
3.1.1	ベルヌーイ分布 (Bernoulli distribution)	20
3.1.2	離散一様分布 (Uniform distribution)	21
3.1.3	二項分布 (Binomial distribution)	22
3.1.4	幾何分布 (Geometric distribution)	25
3.1.5	超幾何分布 (Hyper Geometric Distribution)	26
3.1.6	負の二項分布 (Negative binomial distribution)	28
3.1.7	ポアソン分布 (Poisson 分布)	30
4	確率分布 (連続型変数)	34
4.1	連続型確率分布の例	35
4.1.1	一様分布 (Uniform distribution)	35
4.1.2	指数分布 (Exponential distribution)	36
4.1.3	ガンマ分布 ( $\Gamma$ -distribution)	38
4.1.4	ベータ分布 (Beta distribution)	40
4.1.5	正規分布 (Normal distribution)	40
4.1.6	その他の確率分布	44
5	2 次元確率分布	47
5.1	同時確率分布と周辺確率分布	47
5.2	共分散と相関係数	48
5.3	2 次元正規分布	51
6	大数の法則と中心極限定理	54
6.1	大数の法則	54
6.2	中心極限定理	55
6.3	中心極限定理の応用	58

## 参考文献

- [1] 初等統計学, P.G. ホーエル 著, 浅井晃, 村上正康 訳, 培風館.
- [2] 統計解析入門: 篠崎信雄 著, サイエンス社.
- [3] やさしい統計学入門: 田栗正章, 藤越康祝, 柳井晴夫, C.R. ラオ 著, 講談社.
- [4] やさしい統計: 吉原健一, 金川秀也 著, 培風館.
- [5] 統計学入門, 基礎統計学 , 松原望, 縄田和満, 中井検裕 著, 東京大学出版会.
- [6] Introduction to the theory of statistics, A. M. Mood, F.A. Graybill, D.C. Boes 著, Mc.Graw-Hill.

# 1 記述統計

- 母集団 : 調査したい対象
- 標本 : 母集団の中から選んだグループ.(母集団に比べ数は非常に少ない)
- 変数 : 調査項目として選ばれるある 1 つの特性
- 標本データ : 測定値, 変数の調査結果

- 変数の種類

- 質的変数 形, 色, 性別 (数でないもの)
- 量的変数  $\left\{ \begin{array}{l} \text{離散型変数 (個数, 人数, 日数)} \\ \text{連続型変数 (時間, 身長, 重さ)} \end{array} \right.$

## 1.1 度数分布表

- 標本データ

千葉大学男子 70 人の身長データ

169 171 172 167 171 176 164 169 168 164  
 169 168 164 159 161 167 162 174 168 165  
 169 167 165 171 168 170 163 177 162 164  
 172 177 175 173 156 163 159 157 172 174  
 182 161 175 170 175 173 167 154 173 168  
 175 164 169 171 161 163 176 155 166 180  
 168 164 176 168 181 173 159 183 168 166

- 母集団 : 千葉大学の男子学生
- 変数 : 千葉大学の男子学生の身長
- 標本の大きさ :  $n = 70$

最大値 183cm, 最小値 154cm, 範囲 (最大値 - 最小値) =  $183 - 154 = 29$ cm.

階級の幅 3cm, 階級の数 10 で度数分布表を作る.

階級	階級値	度数 $f$	累積度数 $F$	相対度数 $f/n$	累積相対度数 $F/n$
1	153.5 ~ 156.5	155	3	0.043	0.043
2	156.5 ~ 159.5	158	4	0.057	0.100
3	159.5 ~ 162.5	161	5	0.071	0.171
4	162.5 ~ 165.5	164	11	0.157	0.329
5	165.5 ~ 168.5	167	14	0.200	0.529
6	168.5 ~ 171.5	170	11	0.157	0.687
7	171.5 ~ 174.5	173	9	0.129	0.814
8	174.5 ~ 177.5	176	9	0.129	0.943
9	177.5 ~ 180.5	179	1	0.014	0.957
10	180.5 ~ 183.5	182	3	0.043	1.000

度数分布はこれをグラフに表せば直感的にいっそう理解し易くなる. 連続型変数に対してはヒストグラム (柱状図形) と呼ばれるグラフが有効である. 測定値が階級の境界の値になる場合には, 慣例に従って, その測定値は小さいほうの階級に入れることにする.

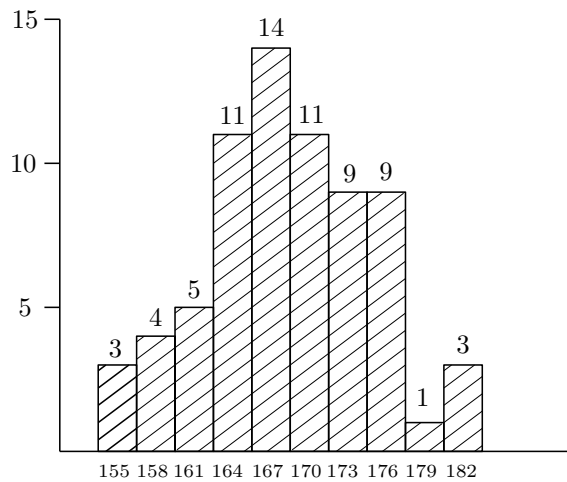


図 1. ヒストグラム (柱状グラフ) 千葉大学の学生 (70 人) の身長

● 離散型変数についての度数分布表

ある街で救急車が一日に何回出動したか記録がある (30 日間).

変数  $X$ : 救急車の一日における出動回数, 標本の大きさ  $n = 30$ .

出動回数 $x$	0	1	2	3	4	5	6
度数 $f$	8	6	5	4	3	3	1
累積度数 $F$	8	14	19	23	26	29	30

離散型変数に対してもヒストグラムを用いることがあるが, ここでは棒グラフを用いる.

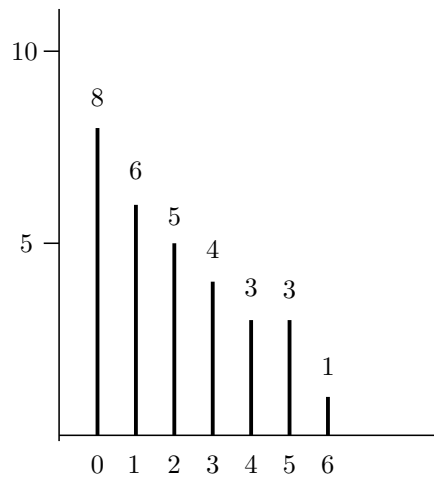


図 2. 棒グラフ 救急車の出動回数

## 1.2 算術的記述

位置の測度 分布の位置を代表させる特性値

(例) 平均, モード, メディアン

変動の測度 分布のばらつきを代表させる特性値

(例) 標準偏差, 四分位範囲

### 1.2.1 標本平均

定義 (標本平均, Sample mean)

(i) データ  $x_1, x_2, \dots, x_n$  が与えられているとき

$$\text{標本平均: } \bar{x} \equiv \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

(ii) 度数分表  $(y_i, f_i), i = 1, 2, \dots, k$ , が与えられているとき

$$\text{標本平均: } \bar{y} \equiv \frac{1}{n}(y_1 f_1 + y_2 f_2 + \dots + y_k f_k) = \frac{1}{n} \sum_{i=1}^k y_i f_i \quad k: \text{階級の数}$$

注:  $\equiv$  は右辺で左辺を定義するという意味で用いられている。

元のデータ

161	→	↘	159.5 ~ 162.5	$\frac{1}{4}(161 + 161 + 163 + 164) = 162.25$
163	→	↗	162.5 ~ 165.5	
164	→			
161	→			

度数分布表

161		2	$\frac{1}{4}(161 \times 2 + 164 \times 2) = 162.5$
164		2	

この例でもわかる様に元のデータを直接用いた標本平均と度数分布を用いた標本平均は一般に一致しない。これは階級の幅があるためで、元のデータが異なるものでも階級値が同じになる場合があるからである。(上の例では 163, 164 が共に 164 の階級に属している)しかし、標本の数  $n$  が十分大きいときはこの誤差は無視できるほど小さくなる。

### 1.2.2 標本標準偏差

定義 (標本標準偏差, Sample standard deviation)

(i) データ  $x_1, x_2, \dots, x_n$  が与えられているとき

$$\text{標本標準偏差} \quad s_x \equiv \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

(ii) 度数分表  $(y_i, f_i), i = 1, 2, \dots, k$ , が与えられているとき

$$\text{標本標準偏差} \quad s_y \equiv \sqrt{\frac{1}{n-1} \sum_{i=1}^k (y_i - \bar{y})^2 f_i}$$

標本標準偏差に関する計算の注意

度数分布表  $(y_i, f_i)$  を用いたとき等式

$$s_y = \sqrt{\frac{1}{n-1} \left\{ \left( \sum_{i=1}^k y_i^2 f_i \right) - \frac{1}{n} \left( \sum_{i=1}^k y_i f_i \right)^2 \right\}}$$

が成り立つ。(実際の計算をするとき便利な形である.)

(証明)

和に関する性質:  $c$  と  $d$  が定数, つまり  $i$  に関係なく一定の時,

$$\sum_{i=1}^k (ca_i + db_i) = c \sum_{i=1}^k a_i + d \sum_{i=1}^k b_i \text{ (和の線形性という)}$$

度数の性質 :  $\sum_{i=1}^k f_i = n,$

平均の定義 :  $\bar{y} = \frac{1}{n} \sum_{i=1}^k y_i f_i \Rightarrow \sum_{i=1}^k y_i f_i = n\bar{y}$  を用いると,

$$\begin{aligned} (n-1)s_y^2 &= \sum_{i=1}^k (y_i - \bar{y})^2 f_i = \sum_{i=1}^k (y_i^2 - 2y_i\bar{y} + \bar{y}^2) f_i = \sum_{i=1}^k y_i^2 f_i - 2\bar{y} \sum_{i=1}^k y_i f_i + \bar{y}^2 \sum_{i=1}^k f_i \\ &= \sum_{i=1}^k y_i^2 f_i - 2\frac{1}{n} \left( \sum_{i=1}^k y_i f_i \right)^2 + \frac{1}{n} \left( \sum_{i=1}^k y_i f_i \right)^2 = \sum_{i=1}^k y_i^2 f_i - \frac{1}{n} \left( \sum_{i=1}^k y_i f_i \right)^2 \end{aligned}$$

と計算できる. ■

(例) 千葉大学の学生 (70 人) の身長に対して標本平均と標本標準偏差を計算する.

	階級値 $y_i$	度数 $f_i$	$y_i f_i$	$y_i^2 f_i$
1	155	3	465	72705
2	158	4	632	99856
3	161	5	805	129605
4	164	11	1804	295856
5	167	14	2338	390446
6	170	11	1870	317900
7	173	9	1557	269361
8	176	9	1584	278784
9	179	1	179	32041
10	182	3	546	99372
計		70	11780	1985296

標本平均  $\bar{y} = \frac{1}{70} \sum_{i=1}^{10} y_i f_i = \frac{11780}{70} = 168.3cm$

分散=標本標準偏差の2乗  $s_y^2 = \frac{1}{70-1} \left\{ \sum_{i=1}^{10} y_i^2 f_i - \frac{1}{70} \left( \sum_{i=1}^{10} y_i f_i \right)^2 \right\}$   
 $= \frac{1}{69} \left\{ 1985296 - \frac{1}{70} \times (11780)^2 \right\} = 41.88819884$

標本標準偏差  $s_y = 6.5cm$

注: 有効桁を考慮して計算結果を書く事が大切である. ここでは有効桁を小数点以下一桁とした.

### 1.2.3 偏差値

定義 (偏差値, Equivalent deviate)

(i) データ  $x_1, x_2, \dots, x_n$  が与えられているとき,

$$\text{値 } w \text{ の偏差値} \equiv 50 + \frac{w - \bar{x}}{s_x}$$

(ii) 度数分布表  $(y_i, f_i), i = 1, 2, \dots, k$  が与えられているとき

$$\text{値 } w \text{ の偏差値} \equiv 50 + \frac{w - \bar{y}}{s_y}$$

### 1.2.4 1次変換を用いた計算

度数分布表  $(y_i, f_i), i = 1, 2, \dots, k$  が与えられたとする.

1次変換

$$\left. \begin{array}{l} y_1 \\ y_2 \\ \vdots \\ y_k \end{array} \right\} \rightarrow \left\{ \begin{array}{l} z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1k}y_k + b_1 \\ z_2 = a_{21}y_1 + a_{22}y_2 + \cdots + a_{2k}y_k + b_2 \\ \vdots \\ z_k = a_{k1}y_1 + a_{k2}y_2 + \cdots + a_{kk}y_k + b_k \end{array} \right.$$

で定められる  $y_1, \dots, y_k$  から  $z_1, \dots, z_k$  への変換を 一次変換 という. この節で用いる 1次変換は  $a_{ij} = a, b_j = b$  と  $i, j$  に対して一定である場合,

$$y_i \rightarrow z_i = ay_i + b$$

に限る. このとき公式が成り立つ.

公式

度数分布表  $(y_i, f_i)$  に対する標本平均を  $\bar{y}$ , 標本標準偏差を  $s_y$ , 度数分布表  $(z_i, f_i)$  に対する標本平均を  $\bar{z}$ , 標本標準偏差を  $s_z$  とする.

$$(1) \quad \bar{z} = a\bar{y} + b$$

$$(2) \quad s_z = |a|s_y$$

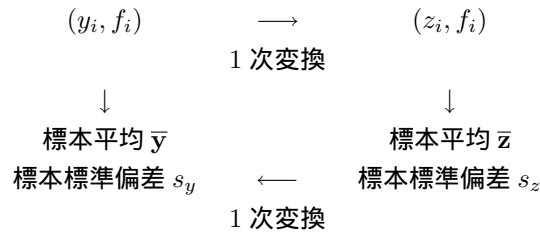
証明

$$(1) \quad \bar{z} = \frac{1}{n} \sum_{i=1}^k z_i f_i = \frac{1}{n} \sum_{i=1}^k (ay_i + b) f_i = \frac{a}{n} \sum_{i=1}^k y_i f_i + b \frac{1}{n} \sum_{i=1}^k f_i = a\bar{y} + b$$

$$(2) \quad s_z^2 = \frac{1}{n-1} \sum_{i=1}^k (z_i - \bar{z})^2 f_i = \frac{1}{n-1} \sum_{i=1}^k (ay_i + b - a\bar{y} - b)^2 f_i = a^2 \frac{1}{n-1} \sum_{i=1}^k (y_i - \bar{y})^2 f_i = a^2 s_y^2$$

■

この公式を用いると、1次変換を用いて標本平均、標本標準偏差の計算を簡単に行うことができる。



(例) 千葉大学の学生(70人)の身長に対して1次変換を用いて標本平均、標本標準偏差を計算してみる。  
 度数は167の時最大であって、階級の幅が3cmであるという理由から、

$$z_i = \frac{y_i}{3} - \frac{167}{3}, \quad i = 1, 2, \dots, k$$

つまり  $a = \frac{1}{3}, b = -\frac{167}{3}$  という一次変換を用いる。

	階級値 $z_i$	度数 $f_i$	$z_i f_i$	$z_i^2 f_i$
1	-4	3	-12	48
2	-3	4	-12	36
3	-2	5	-10	20
4	-1	11	-11	11
5	0	14	0	0
6	1	11	11	11
7	2	9	18	36
8	3	9	27	81
9	4	1	4	16
10	5	3	15	75
計		70	30	334

まず  $(z_i, f_i)$  に対して標本平均と標本標準偏差を計算する。

$$\begin{aligned}
 \bar{z} &= \frac{1}{70} \sum_{i=1}^{10} z_i f_i = \frac{30}{70} = 0.428571423 \\
 s_z^2 &= \frac{1}{69} (334 - \frac{1}{70} \times 30^2) = 4.654244306 \\
 s_z &= 2.157369765
 \end{aligned}$$

次に1次変換  $y_i = 3z_i + 167, \quad i = 1, 2, \dots, k$  に対して公式を用いると、

$$\begin{aligned}
 \bar{y} &= 3\bar{z} + 167 \doteq \underline{168.3cm} \\
 s_y &= 3s_z \doteq \underline{6.5cm}
 \end{aligned}$$

となり、直接計算した場合と同じ値が得られることが確認できた。

### 1.3 その他の記述的測度

#### 1.3.1 位置の測度

- 最頻値, モード (Mode)

定義 (最頻値)

離散型変数 (または質的変数) が最大度数  $f$  を持つ  $x$  が 1 つあるときその値  $x$  を最頻値 (モード) と呼ぶ. 2 つ以上あるときは, 最頻値なしという.

(例) 救急車の 1 日における出動回数 (30 日間)

出動回数	0	1	2	3	4	5	6
度数 $f$	8	6	5	4	3	3	1
累積度数 $F$	8	14	19	23	26	29	30

最大の度数を持つ測定値は  $x=0$ , モード 0 回

- 中央値, メディアン (Median)

定義 中央値 (離散型変数の場合)

離散型変数データを大きさの順に並べたとき, 標本の数  $n$  が奇数である場合  $\frac{n+1}{2}$  番目の測定値, 偶数である場合  $\frac{n}{2}$  番目と  $\frac{n}{2}+1$  番目の測定値の算術平均を中央値という.

(例) 救急車の出動回数の例では  $n = 30$  の偶数であるから 15 番目と 16 番目のデータでの算術平均 15 番目の値 2 回, 16 番目の値 2 回  $\therefore$  メディアン=2 回

定義 中央値 (連続型変数の場合)

柱状グラフ (ヒストグラム) の面積を半分にする値を中央値と呼ぶ.

(例) 千葉大学 70 人の身長の場合では, 標本数  $n = 70$ , 階級の幅  $3\text{cm}$  ヒストグラムの面積は  $3 \times 70 = 210$ . 従って半分の面積は  $210 \times \frac{1}{2} = 105$ . 累積面積より中央値  $y_M$  は階級 5 (165.5 以上 168.5 未満) に含まれており,

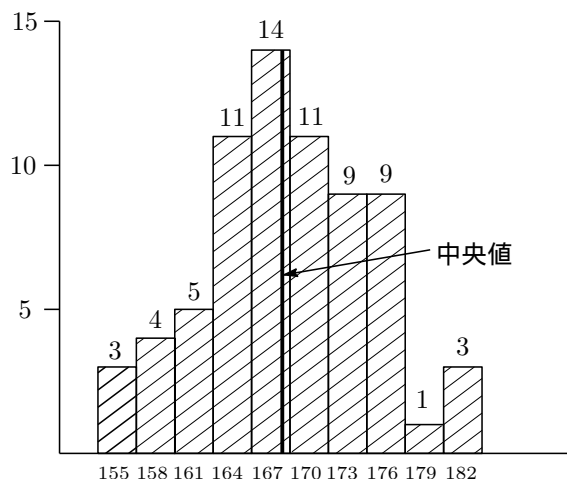
$$69 + 14(y_M - 165.5) = 105$$

を満たす事がわかる. ゆえに,

$$y_M = 165.5 + \frac{105 - 69}{14} = 165.5 + \frac{36}{14} \approx 168.1\text{cm}$$

つまり中央値は 168.1cm である.

階級	度数 $f$	累積度数 $F$	累積面積
1	3	3	9
2	4	7	21
3	5	12	36
4	11	23	69
5	14	37	111
6	11	48	144
7	9	57	171
8	9	66	198
9	1	67	201
10	3	70	210



(標本平均と中央値) : 上の例でもわかるように, 中央値 168.1 と標本平均 168.3 は一般には一致しない. 典型的な例を挙げておく.

データ 1

点	30	60	90
度数	5	10	5

標本平均 60 点, 中央値 60 点

データ 2

点	0	60	90
度数	5	5	10

標本平均 60 点, 中央値 75 点

### 1.3.2 変動の測度

- 範囲, レンジ (Range)

定義 範囲

元のデータ  $x_1, x_2, \dots, x_n$  を用いる.

$$\text{範囲} \equiv (\text{データの最大値}) - (\text{データの最小値})$$

- 四分位範囲 (Interquartile range)

四分位範囲の定義になる第 1 四分位数と第 3 四分位数の定義を述べる. 第 2 四分位数は中央値である.

**定義** (離散型変数の場合)

離散型データを小さい順にならべたとき標本の数  $n$  が

$$4 \text{ の倍数のとき} \Rightarrow \begin{cases} \frac{n}{4} \text{ 番目と } \frac{n}{4} + 1 \text{ 番目の値の算術平均を} \\ \text{第 1 四分位数と呼ぶ.} \\ \frac{3n}{4} \text{ 番目と } \frac{3n}{4} + 1 \text{ 番目の値の算術平均を} \\ \text{第 3 四分位数と呼ぶ.} \end{cases}$$

$$4 \text{ の倍数でないとき} \Rightarrow \begin{cases} \left[ \frac{n}{4} \right] + 1 \text{ 番目の値を第 1 四分位数と呼ぶ.} \\ \left[ \frac{3n}{4} \right] + 1 \text{ 番目の値を第 3 四分位数と呼ぶ.} \end{cases}$$

ここで  $\left[ \frac{n}{4} \right]$  は  $\frac{n}{4}$  を超えない最大の整数.

**定義** (連続型変数の場合)

柱状グラフ (ヒストグラム) の面積を 4 等分する値  $y_1, y_2, y_3$  を順に第 1 四分位数, 第 2 四分位数 (=中央値), 第 3 四分位数という.

定義 四分位範囲

$$\text{四分位範囲} \equiv (\text{第 3 四分位数}) - (\text{第 1 四分位数})$$

(例) 救急車の出動回数の場合

第 1 四分位数 = 8 番目のデータ=0 回  
第 3 四分位数 = 23 番目のデータ=3 回  
四分位範囲 = 3 回 - 0 回=3 回

階級	度数 $f$	累積度数 $F$
0	8	8
1	6	14
2	5	19
3	4	23
4	3	26
5	3	29
6	1	30

(例) 千葉大学学生 70 人の身長に対する四分位数と四分位範囲の計算.

- 第 1 四分位数

$y_1$  は左側の面積を  $\frac{210}{4} = 52.5$  にする値であり, 次の等式を満足

$$36 + (y_1 - 162.5) \times 11 = 52.5$$

$$y_1 = 162.5 + \frac{52.5 - 36}{11} \\ = 164.0cm$$

- 第 3 四分位数

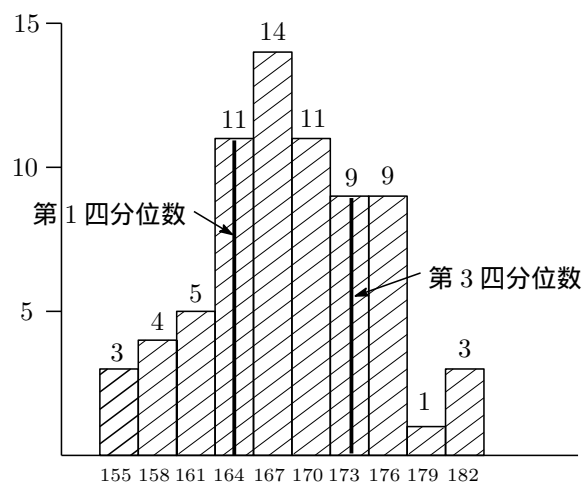
$y_3$  は左側の面積を  $\frac{3}{4} \times 210 = 157.5$  にする値であり次の資料を満足

$$144 + (y_3 - 171.5) \times 9 = 157.5$$

$$y_3 = 171.5 + \frac{157.5 - 144}{9} = 173cm$$

四分位範囲= $y_3 - y_1 = 9cm$

階級	度数 $f$	累積度数 $F$	累積面積
1	3	3	9
2	4	7	21
3	5	12	36
4	11	23	69
5	14	37	111
6	11	48	144
7	9	57	171
8	9	66	198
9	1	67	201
10	3	70	210



## 2 確率空間

- 標本空間 (Sample Space)

試行で得られる結果すべての集まり. 記号は  $S$  を用いる.

- 標本点 (Sample point)

標本空間に含まれる要素を標本点という.

- 事象 (Event)

$S$  の部分集合を事象という.  $A, B, C$  などの記号を用いる. つぎの事象は特別例である.

空事象: 標本点を 1 つも含まない場合も事象であり空事象と呼び  $\phi$  または  $\{\}$  と書く.

全事象: 全ての標本点を含む場合も事象であり全事象と呼び, 標本空間と同じであるので  $S$  と書く.

根元事象: 1 つの標本点からなる事象を根元事象と呼び, 例えば  $i$  を 1 つの標本点とすると  $\{i\}$  となる.

- 確率 (Probability)

各事象に 0 以上 1 以下の数を対応させるもの.

### 2.1 試行と確率

- 試行 1

箱の中に 1~6 の番号がついた玉が各々 1 個ずつ, 計 6 個入っている. この中から無作為に玉を 1 個取り出し, その番号を調べる試行を行う. (無作為に取り出す = 各々の玉が取り出される確率が等しい.)

標本空間  $S = \{1, 2, 3, 4, 5, 6\}$

事象 空事象  $\phi$ , 全事象  $S$ , 根元事象  $\{i\}$ ,  $1 \leq i \leq 6$  のほか, 次の事象がある.  
(計  $2^6 = 64$  個ある.)

2 つの標本点からなる事象  $\{i, j\}$   $1 \leq i < j \leq 6$

3 つの標本点からなる事象  $\{i, j, k\}$   $1 \leq i < j < k \leq 6$

4 つの標本点からなる事象  $\{i, j, k, \ell\}$   $1 \leq i < j < k < \ell \leq 6$

5 つの標本点からなる事象  $\{i, j, k, \ell, m\}$   $1 \leq i < j < k < \ell < m \leq 6$

確率 根元事象の確率は  $\frac{1}{6}$ , つまり,  $P(\{i\}) = \frac{1}{6}, 1 \leq i \leq 6$

$A$  が  $k$  個の標本点から構成されている事象のとき  $P(A) = \frac{k}{6}$

- 試行 2

箱の中に赤球が 3 個, 青球が 2 個, 緑球が 1 個計 6 個入っている. この中から無作為に玉を 1 個取り出し, その玉の色を調べる試行を行う.

標本空間  $S = \{\text{赤}, \text{青}, \text{緑}\}$

事象  $\phi, \{\text{赤}\}, \{\text{青}\}, \{\text{緑}\}, \{\text{赤}, \text{青}\}, \{\text{赤}, \text{緑}\}, \{\text{青}, \text{緑}\}, S$

確率  $\left\{ \begin{array}{l} P(\phi) = 0 \\ P(\{\text{赤}\}) = \frac{3}{6} = \frac{1}{2} \\ P(\{\text{青}\}) = \frac{2}{6} = \frac{1}{3} \\ P(\{\text{緑}\}) = \frac{1}{6} \\ P(\{\text{赤}, \text{青}\}) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} \\ P(\{\text{赤}, \text{緑}\}) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \\ P(\{\text{青}, \text{緑}\}) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2} \\ P(S) = 1 \end{array} \right.$

試行 1 と 2 の対応

計 6 個の玉のうち 1, 2, 3 の番号が付いているものは赤玉, 4, 5 の番号が付いているものは青玉, 6 の番号が付いているものは緑玉とすると,

$$\begin{aligned} P(\{1, 2, 3\}) &= \frac{1}{2} = P(\{\text{赤}\}) \\ P(\{4, 5\}) &= \frac{1}{3} = P(\{\text{青}\}) \\ P(\{6\}) &= \frac{1}{6} = P(\{\text{緑}\}) \end{aligned}$$

が成り立つことがわかる.

● 試行 3(復元抽出法)

箱の中に 1~6 の番号のついた玉が計 6 個入っている. この中から無作為に玉を 1 個取り出し, その番号を調べ元の箱に戻す. そして同じ箱の中から無作為にもう一度玉を取り出しその番号を調べるという試行を行う.

標本空間: 1 回目に  $i$ , 2 回目に  $j$  を取り出したという標本点を  $(i, j)$  と書く. すると,

$$S = \left\{ \begin{array}{cccccc} (1, 1) & (1, 2) & (1, 3) & (1, 4) & (1, 5) & (1, 6) \\ (2, 1) & (2, 2) & (2, 3) & (2, 4) & (2, 5) & (2, 6) \\ (3, 1) & (3, 2) & (3, 3) & (3, 4) & (3, 5) & (3, 6) \\ (4, 1) & (4, 2) & (4, 3) & (4, 4) & (4, 5) & (4, 6) \\ (5, 1) & (5, 2) & (5, 3) & (5, 4) & (5, 5) & (5, 6) \\ (6, 1) & (6, 2) & (6, 3) & (6, 4) & (6, 5) & (6, 6) \end{array} \right\}$$

$$= \{(i, j); 1 \leq i, j \leq 6\}$$

標本空間  $S$  に含まれる標本点の総数  $\#S = 36$

事象: 空事象, 全事象, 根元事象など計  $2^{36}$  個ある.

確率: 根元事象の確率は  $\frac{1}{36}$ , つまり,  $P(\{(i, j)\}) = \frac{1}{36}$ ,  $(i, j) \in S$  であり,  
 $A$  が  $k$  個の標本点から構成されている事象のとき  $P(A) = \frac{k}{36}$

● 試行 4 (非復元抽出法)

箱の中に 1~6 の番号のついた玉が計 6 個入っている. この中から無作為に玉を 1 個取り出す (ただし取り出した玉は箱に戻さない). そして同じ箱の中から無作為にもう一度玉を取り出しその番号を調べるという試行を行う.

標本空間: 1 回目に  $i$ , 2 回目に  $j$  を取り出したという標本点を  $(i, j)$  と書く. すると,

$$S = \left\{ \begin{array}{cccccc} & (1, 2) & (1, 3) & (1, 4) & (1, 5) & (1, 6) \\ (2, 1) & & (2, 3) & (2, 4) & (2, 5) & (2, 6) \\ (3, 1) & (3, 2) & & (3, 4) & (3, 5) & (3, 6) \\ (4, 1) & (4, 2) & (4, 3) & & (4, 5) & (4, 6) \\ (5, 1) & (5, 2) & (5, 3) & (5, 4) & & (5, 6) \\ (6, 1) & (6, 2) & (6, 3) & (6, 4) & (6, 5) & \end{array} \right\}$$

$$= \{(i, j); 1 \leq i, j \leq 6, i \neq j\}$$

標本空間  $S$  に含まれる標本点の総数  $\#S = 30$

事象: 空事象, 全事象, 根元事象など計  $2^{30}$  個ある.

確率: 根元事象の確率は  $\frac{1}{30}$ , つまり,  $P(\{(i, j)\}) = \frac{1}{30}$ ,  $(i, j) \in S$  であり,  
 $A$  が  $k$  個の標本点から構成されている事象のとき  $P(A) = \frac{k}{30}$

一般化について

玉の数を  $N$ , 取り出す回数を  $n$  とする. 試行 3, 4 の一般化を考えてみる.

(試行 3)

$N = 2, n = 5$  表が出る確率が  $\frac{1}{2}$ , 裏が出る確率が  $\frac{1}{2}$  であるコインを 5 回投げて結果を順に記録する.

$N = 6, n = 10$  各目が出る確率が全て等しく  $\frac{1}{6}$  であるサイコロを 10 回投げて結果を順に記録する.

$N = 38, n = 4$  ルーレットを 4 回まわしてその結果を順に記録する.

(試行 4)

$N = 52, n = 5$  トランプから無作為にカードを順に 5 枚取り出し, 結果を記録する.

$N = 1000, n = 3$  ある学校の生徒 1000 人から無作為に順に 3 人選び, 生徒会長, 生徒会副会長, 書記を決める.

試行 3 と 4 の確率の差に関する注意 ( $n = 2$  の場合)

(試行 3)  $P(\{i, j\}) = \frac{1}{N^2}$

(試行 4)  $P(\{i, j\}) = \begin{cases} \frac{1}{N(N-1)} & (i \neq j) \\ 0 & (i = j) \end{cases}$

2 つのモデルの根元事象の確率の差は

$$\frac{1}{N(N-1)} - \frac{1}{N^2} = \frac{N - (N-1)}{N^2(N-1)} = \frac{1}{N^2(N-1)}$$

例えば  $N = 100$  のとき,

確率は 1 万分の 1 程度  
差は 100 万分の 1 程度  $\Rightarrow$  モデル 3 と 4 は差があるが  $N$  が大きいときは無視できるくらい小さい

● 試行 5

箱の中に 1~6 の番号がついた玉が各々 1 個ずつ計 6 個入っている. この中から一度に 2 個の玉を無作為に取り出す試行を行った.

標本空間: 番号  $i$  と番号  $j$  の玉を取り出すという標本点を  $[i, j]$  と書く ( $i < j$ ). すると,

$$S = \left\{ \begin{array}{cccccc} [1, 2] & [1, 3] & [1, 4] & [1, 5] & [1, 6] & \\ & [2, 3] & [2, 4] & [2, 5] & [2, 6] & \\ & & [3, 4] & [3, 5] & [3, 6] & \\ & & & [4, 5] & [4, 6] & \\ & & & & [5, 6] & \end{array} \right\}$$

$= \{[i, j]; 1 \leq i < j \leq 6\}$

標本空間  $S$  に含まれる標本点の総数  $\#S = 15$

事象: 空事象, 全事象, 根元事象など計  $2^{15}$  個ある.

確率: 根元事象の確率は  $\frac{1}{15}$ , つまり,  $P(\{(i, j)\}) = \frac{1}{15}$ ,  $[i, j] \in S$  であり,  $A$  が  $k$  個の標本点から構成されている事象のとき  $P(A) = \frac{k}{15}$

試行 4 と 5 の対応

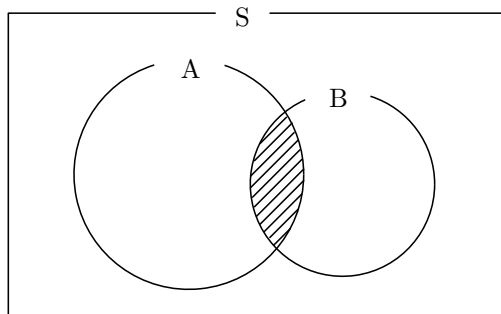
試行 5 で  $i$  番と  $j$  番の玉を一度に取り出す確率は, 試行 4 で 1 回目に  $i$  番, 2 回目に  $j$  番の玉を取り出す確率と, 1 回目に  $j$  番, 2 回目に  $i$  番の玉を取り出す場合を合わせた確率と同じになる.

つまり,  $P(\{(i, j), (j, i)\}) = P(\{[i, j]\})$ ,  $1 \leq i < j \leq 6$  が成り立つ.

## 2.2 条件付確率と独立性

$A$  と  $B$  を事象とする.  $A$  にも  $B$  にも含まれている標本点からなる事象を  $A$  と  $B$  の積事象と呼び  $A \cap B$  と書く.

$A$  かまたは  $B$  に含まれている標本点からなる事象を  $A$  と  $B$  の和事象と呼び  $A \cup B$  と書く.



$A \cap B = \phi$  が成り立つとき  $A$  と  $B$  は互いに素 (または排反) であるという.  $A$  と  $B$  が互いに素であるとき,

$$P(A \cup B) = P(A) + P(B)$$

が成り立つ. この式を加法の公式という.

定義 (条件付確率)

事象  $B$  が起こるという条件の下での事象  $A$  が起こる条件付確率を

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

で定義する. ただし  $P(B) = 0$  の時は定義しない.

定義 (独立性)

事象  $A$  と事象  $B$  が独立であるとは

$$P(A \cap B) = P(A)P(B)$$

が成り立つ事である.

$P(B) \neq 0$  の時  $A$  と  $B$  が独立  $\iff P(A|B) = P(A)$  であることが分かる.

また,

$$P(A \cap B) = P(A|B)P(B)$$

が成り立つ. この式を乗法の公式という.

例) 試行 3 (復元抽出法) を考える.

1 回目に取り出した玉が  $i$  番である事象を  $A(i)$  と書く. つまり,

$$A(i) = \{(i, 1), (i, 2), (i, 3), (i, 4), (i, 5), (i, 6)\}$$

2 回目に取り出した玉が  $j$  番である事象を  $B(j)$  と書く. つまり,

$$B(j) = \{(1, j), (2, j), (3, j), (4, j), (5, j), (6, j)\}$$

$A(i), B(j)$  とその積事象を計算してみると,

$$P(A(i) \cap B(j)) = P(\{(i, j)\}) = \frac{1}{36}$$
$$P(A(i)) = \frac{1}{6}, \quad P(B(j)) = \frac{1}{6}$$

となるので,

$$P(A(i) \cap B(j)) = P(A(i))P(B(j))$$

及びその同値の式

$$P(A(i)) = P(A(i)|B(j))$$

が成り立つ. 従って  $A(i)$  と  $B(j)$  は独立である.

例) 試行 4 (非復元抽出法) を考える. この場合,

$$A(i) = \{(i, j); 1 \leq j \leq 6, \quad i \neq j\}$$
$$B(j) = \{(i, j); 1 \leq i \leq 6, \quad i \neq j\}$$

となるので  $A(i), B(j)$  とその積事象の確率を計算してみると,

$$P(A(i)) = \sum_{\substack{j=1 \\ j \neq i}}^6 P(\{(i, j)\}) = \frac{5}{30} = \frac{1}{6}$$
$$P(B(j)) = \sum_{\substack{i=1 \\ i \neq j}}^6 P(\{(i, j)\}) = \frac{5}{30} = \frac{1}{6}$$
$$P(A(i) \cap B(j)) = P(\{(i, j)\}) = \begin{cases} \frac{1}{30} & i \neq j \\ 0 & i = j \end{cases}$$

となり,

$$P(A(i) \cap B(j)) \neq P(A(i))P(B(i))$$
$$P(A(i)|B(j)) = \begin{cases} \frac{\frac{1}{30}}{\frac{1}{6}} = \frac{1}{5} & i \neq j \\ 0 & i = j \end{cases}$$

が成り立つ. したがって  $A(i)$  と  $B(j)$  は独立ではない.

例) 試行 3 (復元抽出法) を考える.

1 回目の取り出しで偶数番の玉を得たという事象を  $C$  と書く. つまり,

$$C = \left\{ \begin{array}{cccccc} (2, 1) & (2, 2) & (2, 3) & (2, 4) & (2, 5) & (2, 6) \\ (4, 1) & (4, 2) & (4, 3) & (4, 4) & (4, 5) & (4, 6) \\ (6, 1) & (6, 2) & (6, 3) & (6, 4) & (6, 5) & (6, 6) \end{array} \right\}$$

1 回目と 2 回目で取り出した玉の番号の和は 7 以上である事象を  $D$  と書く. つまり,

$$D = \left\{ \begin{array}{cccccc} (1, 6) & (2, 5) & (2, 6) & (3, 4) & (3, 5) & (3, 6) \\ (4, 3) & (4, 4) & (4, 5) & (4, 6) & (5, 2) & (5, 3) \\ (5, 4) & (5, 5) & (5, 6) & (6, 1) & (6, 2) & (6, 3) \\ (6, 4) & (6, 5) & (6, 6) & & & \end{array} \right\}$$

$C, D$  の確率は,

$$P(C) = \frac{18}{36} = \frac{1}{2}, \quad P(D) = \frac{21}{36} = \frac{7}{12}$$

となり, その積事象

$$C \cap D = \left\{ \begin{array}{cccccc} (2, 5) & (2, 6) & (4, 3) & (4, 4) & (4, 5) & (4, 6) \\ (6, 1) & (6, 2) & (6, 3) & (6, 4) & (6, 5) & (6, 6) \end{array} \right\}$$

の確率は,

$$P(C \cap D) = \frac{12}{36} = \frac{1}{3}$$

が成り立つ. 従って,

$$P(C \cap D) = \frac{1}{3} \neq \frac{1}{2} \times \frac{7}{12} = P(C)P(D)$$

となり,  $C$  と  $D$  は独立ではない.

## 2.3 ベイズの定理 (条件付確率での原因と結果の考察)

例 病気の検査の精度

まれにしか起こらない病気を発見するのに有効な検査法がある. この検査法では,

実際に病気の人	⇒	97% の確率で陽性
健康な人	⇒	5% の確率で陽性
風邪の人	⇒	10% の確率で陽性

となることが今までのデータでわかっている. 一方, 多人数からなるある母集団において

実際に病気の人	の割合が 1%
健康な人	の割合が 96%
風邪の人	の割合が 3%

であることがわかっている. この母集団から無作為に選ばれた一人が陽性反応を示した. このときこの人が実際に病気である確率はどのくらいか調べる.

実際に病気である事象を病
健康である事象を健
風邪である事象を風
陽性反応を示す事象を陽

と書くことにする. この記号を用いて上で述べたことをまとめると,

$P(\text{陽}   \text{病})$	$= 0.97$	$P(\text{病})$	$= 0.01$
$P(\text{陽}   \text{健})$	$= 0.05$	$P(\text{健})$	$= 0.96$
$P(\text{陽}   \text{風})$	$= 0.10$	$P(\text{風})$	$= 0.03$

となり, 求める確率は  $P(\text{病} | \text{陽})$  である. 乗法の公式より,

$$\begin{aligned} P(\text{病} \cap \text{陽}) &= P(\text{病})P(\text{陽} | \text{病}) = 0.01 \times 0.97 = 0.0097 \\ P(\text{健} \cap \text{陽}) &= P(\text{健})P(\text{陽} | \text{健}) = 0.96 \times 0.05 = 0.048 \\ P(\text{風} \cap \text{陽}) &= P(\text{風})P(\text{陽} | \text{風}) = 0.03 \times 0.1 = 0.003 \end{aligned}$$

が成り立ち、加法の公式より、

$$\begin{aligned} P(\text{陽}) &= P(\text{病} \cap \text{陽}) + P(\text{健} \cap \text{陽}) + P(\text{風} \cap \text{陽}) \\ &= 0.0097 + 0.048 + 0.003 \\ &= 0.0607 \end{aligned}$$

が成り立つ。従って、

$$P(\text{病} | \text{陽}) = \frac{P(\text{病} \cap \text{陽})}{P(\text{陽})} = \frac{0.0097}{0.0607} = 0.16$$

従って検査で陽性であっても本当に病気である確率は16%程度である。

次に逆の場合を考えてみる。無作為に選ばれた一人が陰性反応を示したとき、この人が実際に病気である確率  $P(\text{病} | \text{陰})$  である確率を調べる。

$$\begin{aligned} P(\text{陰} | \text{病}) &= 0.03, \quad P(\text{陰} | \text{健}) = 0.95, \quad P(\text{陰} | \text{風}) = 0.9 \\ P(\text{病} | \text{陰}) &= \frac{P(\text{病} \cap \text{陰})}{P(\text{陰})} \\ P(\text{病} \cap \text{陰}) &= P(\text{病})P(\text{陰} | \text{病}) = 0.01 \times 0.03 = 0.0003 \\ P(\text{陰}) &= 1 - P(\text{陽}) = 1 - 0.0607 = 0.9393 \end{aligned}$$

従って、

$$P(\text{病} | \text{陰}) = \frac{0.0003}{0.9393} = 0.000319 \quad (0.0319\%)$$

となり、陰性であれば本当に病気である確率は殆ど無いことが分かる。

上の例は次で示されるベイズの定理の応用例である。

### ベイズの定理

$A_1, A_2, \dots, A_n$  を互いに素である事象列、つまり、

$$A_i \cap A_j = \phi, \quad i \neq j$$

であるとし、さらに、 $\cup_{i=1}^n A_i = S$  を満たすものとする。各  $i$  に対して  $P(A_i) > 0$  であり、 $B$  も  $P(B) > 0$  である事象とすると

$$P(A_k | B) = \frac{P(B | A_k)P(A_k)}{\sum_{i=1}^n P(B | A_i)P(A_i)}$$

が成り立つ。

証明 乗法の公式より、

$$\begin{aligned} P(B | A_k)P(A_k) &= P(B \cap A_k) \\ \sum_{i=1}^n P(B | A_i)P(A_i) &= \sum_{i=1}^n P(B \cap A_i) \end{aligned}$$

が成り立つ。 $\cup_{i=1}^n A_i = S$  であり、 $A_i \cap A_j = \phi, i \neq j$  であることより、加法の公式を用いると、

$$\sum_{i=1}^n P(B \cap A_i) = P(B \cap S) = P(B)$$

となる。従って、

$$\frac{P(B | A_k)P(A_k)}{\sum_{i=1}^n P(B | A_i)P(A_i)} = \frac{P(B \cap A_k)}{P(B)} = P(A_k | B)$$

が得られる。

### 3 確率分布 (離散型変数)

標本空間が有限個の標本点から構成される場合, つまり

$$S = \{x_1, x_2, \dots, x_n\}$$

の場合を考える. ここで  $n$  は自然数である.

定義 (確率密度)

$p$  が  $S$  上の確率密度であるとは,

- (i)  $0 \leq p(x) \leq 1, x \in S$
- (ii)  $\sum_{x \in S} p(x) = p(x_1) + p(x_2) + \dots + p(x_n) = 1$

が成り立つことである.  $S$  の事象  $A$  に対して,

$$(3.1) \quad P(A) = \sum_{x \in A} p(x)$$

で  $P$  を定義すると

- (i)  $0 \leq P(A) \leq 1,$
- (ii)  $P(S) = 1$

が成立し, 加法の公式を満足する.  $P$  を  $p(x)$  を確率密度とする  $S$  上の確率という. 以後  $S$  の元, つまり標本点は実数とする (ほとんどの例は整数).

実数から実数の関数  $f$  が与えられたとき, その期待値は  $\sum_{x \in S} f(x)p(x)$  で定義されるが, その特別な場合として平均, 2次モーメント, 分散, 標準偏差, 積率母関数が次のように定義される.

平均 (mean)  $\mu = \sum_{x \in S} xp(x)$

2次モーメント (second moment)  $m_2 = \sum_{x \in S} x^2 p(x)$

分散 (variance)  $V = \sum_{x \in S} (x - \mu)^2 p(x) = m_2 - \mu^2$

標準偏差 (standard deviation)  $\sigma = \sqrt{V}$

積率母関数 (moment generating function)  $m(t) = \sum_{x \in S} e^{tx} p(x)$

#### 3.1 離散型確率分布の例

##### 3.1.1 ベルヌーイ分布 (Bernoulli distribution)

$$S = \{0, 1\}, \quad p(0) = 1 - p, \quad p(1) = p,$$

(ただし  $p$  は 0 以上 1 以下の実数) の時, 確率密度  $p(x)$  で定まる (離散型) 確率分布をベルヌーイ分布という.

- 対応する試行

表が出る確率が  $p$ , 裏が出る確率が  $1 - p$  であるコインを 1 回投げて, 表が出た回数を調べる.

- 平均

$$\mu = 0 \cdot p(0) + 1 \cdot p(1) = p$$

- 2 次モーメント

$$m_2 = 0^2 \cdot p(0) + 1^2 \cdot p(1) = p$$

- 分散

$$V = m_2 - \mu^2 = p - p^2 = p(1 - p)$$

- 標準偏差

$$\sigma = \sqrt{V} = \sqrt{p(1 - p)}$$

- 積率母関数

$$m(t) = e^{t \cdot 0} p(0) + e^{t \cdot 1} p(1) = 1 \cdot (1 - p) + e^t p = 1 + (e^t - 1)p$$

### 3.1.2 離散一様分布 (Uniform distribution)

$$S = \{1, 2, \dots, n\}, \quad p(x) = \frac{1}{n}, \quad x \in S$$

の時, 確率密度  $p(x)$  で定まる (離散型) 確率分布を離散一様分布という.

- 対応する試行

各々の目が出る確率がすべて等しいサイコロを投げ, 出た目の数を調べる. (この場合  $n = 6$ )

- 平均

$$\mu = \sum_{x=1}^n x p(x) = \frac{1}{n} \sum_{x=1}^n x = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

- 2 次モーメント

$$m_2 = \sum_{x=1}^n x^2 p(x) = \frac{1}{n} \sum_{x=1}^n x^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

- 分散

$$\begin{aligned} V = m_2 - \mu^2 &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n+1}{12} \{2(2n+1) - 3(n+1)\} = \frac{n+1}{12} (n-1) = \frac{n^2-1}{12} \end{aligned}$$

- 標準偏差

$$\sigma = \sqrt{V} = \sqrt{\frac{n^2-1}{12}}$$

- 積率母関数

$$m(t) = \sum_{x=1}^n e^{tx} p(x) = \frac{1}{n} \sum_{x=1}^n e^{tx} = \frac{e^t}{n} \cdot \frac{1 - e^{tn}}{1 - e^t}$$

### 3.1.3 二項分布 (Binomial distribution)

$$S = \{0, 1, 2, \dots, n\}$$

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x},$$

の時, 確率密度  $p(x)$  で定まる確率分布を二項分布という. ここで,  $n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1$  であり  $n$  の階乗という. また  $\binom{n}{x}$  は  $n$  個の異なるものから  $x$  個を選ぶ組み合わせの数という.

- 対応する試行

表が出る確率が  $p$ , 裏が出る確率  $1-p$  であるコインを  $n$  回投げて表が出た回数を調べる.

表が出た回数を  $X$  とおき,  $n = 1, 2, 3$  の時について調べてみる.

$n = 1$  の時 (ベルヌーイ分布と一致)

$$X = 0 \iff \text{裏} \cdots p(0) = 1-p$$

$$X = 1 \iff \text{表} \cdots p(1) = p$$

$n = 2$  の時

$$X = 0 \iff \text{裏} \cdot \text{裏} \cdots p(0) = (1-p)^2$$

$$X = 1 \iff \text{表} \cdot \text{裏 または 裏} \cdot \text{表} \cdots p(1) = 2p(1-p)$$

$$X = 2 \iff \text{表} \cdot \text{表} \cdots p(2) = p^2$$

$n = 3$  の時

$$X = 0 \iff \text{裏} \cdot \text{裏} \cdot \text{裏} \cdots p(0) = (1-p)^3$$

$$X = 1 \iff \text{裏} \cdot \text{裏} \cdot \text{表}, \text{裏} \cdot \text{表} \cdot \text{裏 または 表} \cdot \text{裏} \cdot \text{裏} \cdots p(1) = 3p(1-p)^2$$

$$X = 2 \iff \text{裏} \cdot \text{表} \cdot \text{表}, \text{表} \cdot \text{裏} \cdot \text{表 または 表} \cdot \text{表} \cdot \text{裏} \cdots p(2) = 3p^2(1-p)$$

$$X = 3 \iff \text{表} \cdot \text{表} \cdot \text{表} \cdots p(3) = p^3$$

演習  $n = 4$  の時,  $p(0), p(1), p(2), p(3), p(4)$  を求めよ.

参照として特別な場合の二項分布の確率を記しておく.

$n = 10, p = 0.5$ のとき		$n = 10, p = 0.01$ のとき	
$p(0) = 0.0010$	$p(6) = 0.2051$	$p(0) = 0.9044$	$p(6) = 0$
$p(1) = 0.0098$	$p(7) = 0.1172$	$p(1) = 0.0914$	$p(7) = 0$
$p(2) = 0.0439$	$p(8) = 0.0439$	$p(2) = 0.0042$	$p(8) = 0$
$p(3) = 0.1172$	$p(9) = 0.0098$	$p(3) = 0.0001$	$p(9) = 0$
$p(4) = 0.2051$	$p(10) = 0.0010$	$p(4) = 0$	$p(10) = 0$
$p(5) = 0.2461$		$p(5) = 0$	

さらに一般の場合は二項分布表を見よ.

注意 表が出る確率が  $p$  であるコインを  $n$  回投げる. 表が出た回数の分布は, パラメータ  $(n, p)$  の二項分布  $Bi(n, p)$  であるが, 裏が出た回数の分布は, パラメータ  $(n, 1-p)$  の二項分布  $Bi(n, 1-p)$  となる. 従って, パラメータ  $p \leq 1/2$  の二項分布の値からパラメータ  $p > 1/2$  の二項分布の値が求められることがわかる.

二項分布の平均, 分散および積率母関数を計算するには二項定理を用いる.

二項定理 :  $(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$

$q = 1 - p$  とおく. この二項定理を用いると,

$$P(S) = \sum_{x=0}^n p(x) = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1$$

が直ちに導かれる.

• 平均

$$\mu = \sum_{x=0}^n x \cdot p(x) = \sum_{x=0}^n x \times \frac{n!}{x!(n-x)!} p^x q^{n-x} = \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x}$$

$\ell = x - 1, m = n - 1$  とおき,  $n! = n \times (n - 1)! = n \times m!, \quad n - x = m - \ell$  となることに注意すると

$$= \sum_{\ell=0}^m \frac{n \times m!}{\ell!(m-\ell)!} p^{\ell+1} q^{m-\ell} = np \sum_{\ell=0}^m \frac{m!}{\ell!(m-\ell)!} p^{\ell} q^{m-\ell}$$

ここで二項定理を用いると

$$= np$$

を得る. 従って

二項分布の平均:  $\mu = np$

• 分散

$V = m_2 - \mu^2$  よりまず 2 次モーメント  $m_2$  を計算する.

$$\begin{aligned} m_2 &= \sum_{x=0}^n x^2 p(x) = \sum_{x=0}^n x^2 \frac{n!}{x!(n-x)!} p^x q^{n-x} \\ &= \sum_{x=2}^n x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + \sum_{x=1}^n x \frac{n!}{x!(n-x)!} p^x q^{n-x} \\ &= \sum_{x=2}^n \frac{n!}{(x-2)!(n-x)!} p^x q^{n-x} + np \end{aligned}$$

$\ell = x - 2, m = n - 2$  とおき,  $n! = n(n-1)m!, \quad n - x = m - \ell$  に注意すると

$$= \sum_{\ell=0}^m n(n-1)p^2 \cdot \frac{m!}{\ell!(m-\ell)!} p^{\ell} q^{m-\ell} + np$$

となり, ここで二項定理を用いると

$$= n(n-1)p^2 + np$$

を得る. よって,

$$V = \sigma^2 = n(n-1)p^2 + np - (np)^2 = n(p-p^2) = npq$$

となる. 従って

二項分布の分散	:	$V = npq$
二項分布の標準偏差	:	$\sigma = \sqrt{npq}$

● 積率母関数

二項分布の積率母関数は二項定理を用いることにより次のように簡単に計算できる。

$$\begin{aligned} m(t) &= \sum_{x=0}^n e^{tx} p(x) = \sum_{x=0}^n e^{tx} \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x} = (pe^t + q)^n \end{aligned}$$

二項分布の積率母関数:  $m(t) = (pe^t + q)^n$

積率母関数を  $\beta$  階微分してから 0 を代入すると,  $\beta$  次モーメント ( $\beta$  次積率) となる。二項分布の場合に平均について調べてみると

$$\begin{aligned} \frac{d}{dt} m(t) &= \frac{d}{dt} (pe^t + q)^n = n(pe^t + q)^{n-1} pe^t \\ &\quad t=0 \text{ とすれば} \\ &= n(p + q)^{n-1} p = np \end{aligned}$$

となることが確認できる。2 次モーメントについても同様に計算できる。(各自確認するように。) 上の計算では合成関数の微分に関する公式

合成関数の微分

$$\frac{d}{dt} F(g(t)) = F'(g(t))g'(t)$$

を  $F(x) = x^n$ ,  $g(t) = pe^t + q$  に対して適用している。

考察 二項分布は  $n = 1$  という特別な場合はベルヌーイ分布である。そして一般の  $n$  の場合での積率母関数はベルヌーイ分布の積率母関数の  $n$  乗となっている。

**演習**

【1】ある新薬の治癒率は 80 % であるとする。いま, 5 人の患者にこの薬を用いたとき, 治る人数を表す確率変数を  $X$  とする。

- (1)  $X = 4$  となる確率を求めよ。
- (2)  $X$  の確率分布を求め,  $X$  の平均と標準偏差を計算せよ。

【2】男(女)が生まれる確率を  $1/2$  ( $1/2$ ) として, 5 人の子供をもつ家族で次の事象の起こる確率を求めよ。

- (1) 5 人のうち少なくとも 4 人が男である。
- (2) 男と女が少なくとも 1 人は含まれる。
- (3) 5 人とも性別は同じである。

【3】ある自動車部品メーカーでは, 生産される部品の 1 箱 (10 個の部品) には, たかだか 1 個の不良品しか含まれていないことを保証している。過去の経験からこの工場の製造工程では 5 % の不良品を出すことが分かっているとき, 任意の 1 箱がこの保証を満たす確率を求めよ。

### 3.1.4 幾何分布 (Geometric distribution)

$S$  の標本点が無限個つまり

$$S = \{0, 1, 2, 3, \dots\}$$

の場合にも離散型確率分布が定義できる.  $p$  を  $0 < p < 1$  を満たす実数とし

$$p(x) = p(1-p)^x, \quad x \in S$$

で定義される離散型確率密度で定まる (離散型) 確率分布を幾何分布という.

- 対応する試行

表が出る確率が  $p$ , 裏が出る確率が  $1-p$  であるコインを投げる試行を行う. 1 回目の試行で

$$( ) \left\{ \begin{array}{l} \text{表が出た場合そこで終了する.} \\ \text{裏が出た場合もう一度コインを投げる.} \end{array} \right.$$

そして同様に ( ) を表が出るまで繰り返し行う. このとき表がでるまでに裏がでた回数は幾何分布に従う.

全事象の確率が 1 であることは等比級数の和の公式

$$\sum_{x=0}^{\infty} r^x = \frac{1}{1-r}, \quad (\text{ただし } |r| < 1)$$

を用いると

$$P(S) = \sum_{x=0}^{\infty} p(x) = p \sum_{x=0}^{\infty} (1-p)^x = \frac{p}{1-(1-p)} = 1$$

直ちに導かれる.

- 平均

級数の公式

$$\sum_{x=1}^{\infty} x r^x = \frac{r}{(1-r)^2}, \quad (\text{ただし } |r| < 1)$$

を用いると,

$$\mu = \sum_{x=1}^{\infty} x p(x) = \sum_{x=1}^{\infty} x p(1-p)^x = p \times \frac{1-p}{p^2} = \frac{1-p}{p}$$

従って,

$\text{幾何分布の平均 : } \mu = \frac{1-p}{p}$

- 分散

級数の公式

$$\sum_{x=1}^{\infty} x^2 r^x = \frac{r(1+r)}{(1-r)^3}, \quad (\text{ただし } |r| < 1)$$

を用いると,

$$\begin{aligned} V &= \sum_{x=1}^{\infty} x^2 p(x) - \mu^2 \\ &= \sum_{x=1}^{\infty} x^2 p(1-p)^x - \left(\frac{1-p}{p}\right)^2 \\ &= p \cdot \frac{1}{p^3} (1-p)(2-p) - \frac{(1-p)^2}{p^2} \\ &= \frac{1-p}{p^2} \{2-p-(1-p)\} \\ &= \frac{1-p}{p^2}. \end{aligned}$$

幾何分布の分散	:	$V = \frac{1-p}{p^2}$
幾何分布の標準偏差	:	$\sigma = \frac{\sqrt{1-p}}{p}$

- 積率母関数

最初の級数の公式を用いると

$$\begin{aligned} m(t) &= \sum_{x=0}^{\infty} e^{tx} p(x) \\ &= p \sum_{x=0}^{\infty} e^{tx} (1-p)^x = \frac{p}{1-(1-p)e^t}, \quad t < -\log(1-p). \end{aligned}$$

幾何分布の積率母関数:	$m(t) = \frac{p}{1-(1-p)e^t}, \quad t < -\log(1-p)$
-------------	---

### 3.1.5 超幾何分布 (Hyper Geometric Distribution)

$S$  を標本点が有限個つまり

$$S = \{0, 1, 2, 3, \dots, n\}$$

である標本空間であるとする.  $M$  を  $n$  以上の整数とし

$$p(x) = \begin{cases} \frac{\binom{k}{x} \cdot \binom{M-k}{n-x}}{\binom{M}{n}} & x = 0, 1, \dots, n \\ 0 & \text{その他} \end{cases}$$

で定義される離散型確率密度で定まる (離散型) 確率分布を超幾何分布という.

- 対応する試行

$k$  個の赤玉,  $M-k$  個の白玉, 計  $M$  個の玉が箱の中に入っている. この箱の中から  $n$  個の玉を非復元抽出法で取り出したとき赤玉の個数の分布は超幾何分布である.

注意 復元抽出法の場合は  $p = \frac{k}{M}$  である二項分布である.

まず超幾何分布の計算のための公式を準備しておく.

$$\binom{a+b}{j} = \sum_{k=0}^j \binom{a}{k} \binom{b}{j-k}$$

証明 二項定理より  $a, b$  が自然数のとき

$$(1+x)^a = \sum_{k=0}^a \binom{a}{k} x^k, \quad (1+x)^b = \sum_{k=0}^b \binom{b}{k} x^k, \quad (1+x)^{a+b} = \sum_{k=0}^{a+b} \binom{a+b}{k} x^k$$

が成り立つ.  $(1+x)^a(1+x)^b = (1+x)^{a+b}$  に注意すると

$$(3.2) \quad \sum_{k=0}^a \binom{a}{k} x^k \sum_{k=0}^b \binom{b}{k} x^k = \sum_{k=0}^{a+b} \binom{a+b}{k} x^k,$$

が得られる.  $x^j$  の係数を比較すると

$$\begin{aligned} \text{右辺の } x^j \text{ の係数} &= \binom{a+b}{j} \\ \text{左辺の } x^j \text{ の係数} &= \sum_{k=0}^j \binom{a}{k} \binom{b}{j-k} \end{aligned}$$

よって上の関係式が得られる.

- 平均

$$\begin{aligned} \mu &= \sum_{x=0}^n xp(x) = \sum_{x=0}^n x \times \frac{\binom{k}{x} \binom{M-k}{n-x}}{\binom{M}{n}} \\ &= \frac{nk}{M} \sum_{x=1}^n \frac{\binom{k-1}{x-1} \binom{M-k}{n-x}}{\binom{M-1}{n-1}} \\ &= \frac{nk}{M} \sum_{y=0}^{n-1} \frac{\binom{k-1}{y} \binom{(M-1)-(k-1)}{n-1-y}}{\binom{M-1}{n-1}} \end{aligned}$$

ここで 2 番目の等式は次の計算から得られた.

$$x \frac{k!}{(k-x)!x!} \frac{n!(M-n)!}{M!} = \frac{k(k-1)!}{(k-x)!(x-1)!} \frac{n(n-1)!(M-n)!}{M(M-1)!} = \frac{kn}{M} \frac{\binom{k-1}{x-1}}{\binom{M-1}{n-1}}.$$

上述で示された公式を用いると

$$\sum_{y=0}^{n-1} \frac{\binom{k-1}{y} \binom{(M-1)-(k-1)}{n-1-y}}{\binom{M-1}{n-1}} = 1$$

となるので

$$\boxed{\text{超幾何分布の平均} : = \frac{nk}{M}}$$

注意 超幾何分布の平均は,  $p = \frac{k}{M}$  としたときの二項分布と同じ.

- 分散

$V = m_2 - \mu^2$  という関係式を用いる.  $\mu$  はすでに計算されているので  $m_2 - \mu$  を計算してみると

$$\begin{aligned}
 m_2 - \mu &= \sum_{x=0}^n x(x-1) \frac{\binom{k}{x} \binom{M-k}{n-x}}{\binom{M}{n}} \\
 &= \sum_{x=0}^n x(x-1) \frac{k!}{(k-x)!x!} \frac{n!(M-n)!}{M!} \binom{M-k}{n-x} \\
 &= \sum_{x=0}^n \frac{k(k-1)(k-2)!}{(k-x)!(x-2)!} \frac{n}{M} \frac{n-1}{M-1} \frac{(n-2)!(M-n)!}{(M-2)!} \binom{M-k}{n-x} \\
 &= n(n-1) \frac{k(k-1)}{M(M-1)} \sum_{x=2}^n \frac{\binom{k-2}{x-2} \binom{M-k}{n-x}}{\binom{M-2}{n-2}} \\
 &= n(n-1) \frac{k(k-1)}{M(M-1)}
 \end{aligned}$$

最後の等式で上述の公式を用いた. したがって

$$\begin{aligned}
 V &= n(n-1) \frac{k(k-1)}{M(M-1)} + n \frac{k}{M} - n^2 \frac{k^2}{M^2} \\
 &= \frac{M(n^2 - n)k(k-1) + nkM(M-1) - n^2k^2(M-1)}{M^2(M-1)} \\
 &= \frac{Mk^2n^2 - Mkn^2 - Mk^2n + Mkn + M^2kn - Mkn - Mk^2n^2 + k^2n^2}{M^2(M-1)} \\
 &= \frac{-Mkn^2 - Mk^2n + M^2kn + k^2n^2}{M^2(M-1)} \\
 &= \frac{nk(-Mn - Mk + M^2 + nk)}{M^2(M-1)} \\
 &= \frac{nk(M-k)(M-n)}{M^2(M-1)} \\
 &= n \frac{k}{M} \frac{M-k}{M} \frac{M-n}{M-1}
 \end{aligned}$$

超幾何分布の分散 :  $V = n \frac{k}{M} \frac{M-k}{M} \frac{M-n}{M-1}$   
 超幾何分布の標準偏差 :  $\sigma = \sqrt{n \frac{k}{M} \frac{M-k}{M} \frac{M-n}{M-1}}$

考察  $p = k/M, q = 1 - p$  とおくと分散は  $npq \times \frac{M-n}{M-1}$  となる. これは二項分布の分散の  $\frac{M-n}{M-1}$  倍になる.

### 3.1.6 負の二項分布 (Negative binomial distribution)

負の二項分布は幾何分布の一般化である. 標本点が無限個である標本空間

$$S = \{0, 1, 2, 3, \dots\}$$

上の離散型確率分布である.  $p$  を 0 以上 1 以下の実数としたとき, (離散型) 確率分布を負の二項分布の離散型確率密度は

$$p(x) = \begin{cases} \binom{r+x-1}{x} p^r q^x & x = 0, 1, 2, \dots \\ 0 & \text{その他} \end{cases}$$

で定義される.

- 対応する試行

表が出る確率が  $p$ , 裏が出る確率が  $1 - p$  であるコインを表が  $r$  回出るまで投げ続ける. このとき裏が出た回数の分布は負の二項分布である.

注意 通常, 組み合わせ  $\binom{n}{k}$  は  $n \geq k \geq 0$  を満たす整数の組  $(n, k)$  に対して定義されているが, すべての整数の組  $(n, k)$  に対しても

$$k \geq 0 \text{ のとき } \binom{n}{k} = \frac{1}{k!} \left( \frac{d}{dy} \right)^k y^n \Big|_{y=1}, \quad k < 0 \text{ のとき } \binom{n}{k} = 0,$$

と定義することにより一般化できる. 定義から  $\alpha$  が負の整数であるとき

$$\binom{\alpha}{k} = (-1)^k \binom{k - \alpha - 1}{k}$$

が成り立つことが分かる.

テイラー展開

$$F(y) = F(y_0) + \sum_{k=0}^{\infty} \left( \frac{d}{dy} \right)^k F(y_0) \frac{(y - y_0)^k}{k!}$$

を  $F(y) = (y + 1)^\alpha$  の場合に組み合わせの記号を用いて書き直すと

$$(y + 1)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} y^k$$

を得る. この等式を  $\alpha = -r, y = -q$  として用いると公式

$$p^{-r} = (1 - q)^{-r} = \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = \sum_{x=0}^{\infty} \binom{x + r - 1}{x} q^x$$

が導かれ, この公式を用いると

$$P(S) = \sum_{x=0}^{\infty} p(x) = p^r \sum_{x=0}^{\infty} \binom{x + r - 1}{x} q^x = 1$$

を得る.

- 平均

$$\begin{aligned} \mu &= \sum_{x=0}^{\infty} x p(x) = \sum_{x=0}^{\infty} x \binom{r + x - 1}{x} p^r q^x \\ &= r \sum_{x=1}^{\infty} \binom{r + x - 1}{x - 1} p^r q^x \\ &= \frac{qr}{p} \sum_{y=0}^{\infty} \binom{r + y}{y} p^{r+1} q^y = \frac{qr}{p} \end{aligned}$$

最後の等式は上述の公式から導いた. 従って

負の二項分布の平均:  $\mu = \frac{qr}{p}$

- 分散

負の二項分布の場合も  $V = m_2 - \mu^2$ ,  $m_2 = \mu + \sum_{x=0}^{\infty} x(x-1)p(x)$  という関係を用いる

$$\begin{aligned} \sum_{x=0}^{\infty} x(x-1)p(x) &= \sum_{x=0}^{\infty} x(x-1) \binom{r+x-1}{x} p^r q^x \\ &= r(r+1) \sum_{x=2}^{\infty} \binom{r+x-1}{x-2} p^r q^x \\ &= \frac{q^2 r(r+1)}{p^2} \end{aligned}$$

となるので

$$V = \frac{q^2 r(r+1)}{p^2} + \frac{qr}{p} - \left(\frac{qr}{p}\right)^2 = \frac{qr}{p^2}$$

負の二項分布の分散	: $V = \frac{qr}{p^2}$
負の二項分布の標準偏差	: $\sigma = \sqrt{\frac{qr}{p^2}}$

- 積率母関数

公式を用いると

$$\begin{aligned} m(t) &= \sum_{x=0}^{\infty} e^{tx} p(x) \\ &= p^r \sum_{x=0}^{\infty} \frac{(r+x-1)!}{x!(r-1)!} (qe^t)^x \\ &= \left(\frac{p}{1-qe^t}\right)^r \end{aligned}$$

が導かれる.

負の二項分布の積率母関数: $m(t) = \left(\frac{p}{1-qe^t}\right)^r$
--

### 3.1.7 ポアソン分布 (Poisson 分布)

ポアソン分布は標本空間

$$S = \{0, 1, 2, 3, \dots\}$$

上の分布であり, その分布密度は

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in S$$

で定義される. ここで  $\lambda > 0$  である.

指数関数のテイラー展開

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots$$

を用いると

$$P(S) = \sum_{x=0}^{\infty} p(x) = e^{-\lambda} e^\lambda = 1$$

が導かれる.

- ポアソン分布と関係のある現象

【1】ある州における1週間あたりの交通事故死亡者数

【2】単位時間当たりの放射性物質の放射の数

【3】ある素材の単位面積あたりの傷の数

- 平均

$$\mu = \sum_{x=0}^{\infty} xp(x) = e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

ポアソン分布の平均:  $\mu = \lambda$

- 分散

$$\sum_{x=0}^{\infty} x(x-1)p(x) = \sum_{x=2}^{\infty} e^{-\lambda} \frac{\lambda^x}{(x-2)!} = \lambda^2 \sum_{\ell=0}^{\infty} e^{-\lambda} \frac{\lambda^{\ell}}{\ell!} = \lambda^2$$

したがって

$$V = \sum_{x=0}^{\infty} x(x-1)p(x) + \mu - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

ポアソン分布の分散 :  $V = \lambda$   
 ポアソン分布の標準偏差 :  $\sigma = \sqrt{\lambda}$

- 積率母関数

$$m(t) = \sum_{x=0}^{\infty} e^{tx} p(x) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} = \exp[e^t \lambda - \lambda]$$

ポアソン分布の積率母関数:  $m(t) = \exp[e^t \lambda - \lambda]$

ポアソン分布に関する重要な定理を紹介する.

【1】長さ  $h(>0)$  の小さな時間の中で正確に一つの出来事が起こる確率は  $\lambda h$  に近似的に等しい. つまり

$$P[\text{長さ } h \text{ の時間の中で出来事が一つ起こる}] = \lambda h + o(h)$$

が成り立つ. ここで

$$\frac{f(h)}{h} \rightarrow 0, h \rightarrow 0 \iff f(h) = o(h)$$

である. たとえば  $h^2, h^3$  は  $o(h)$  である.

【2】長さ  $h(>0)$  の小さな時間の中で出来事が2つ以上起こる確率は  $o(h)$ , つまり無視できる.

【3】重なり合っていない時間における出来事の回数は独立.

定理 3.1 上の3つの仮定が満たされるならば長さ  $t$  の時間で起こる出来事の回数はパラメータ  $\lambda t$  のポアソン分布に従う.

証明  $s > 0$ , にたいして  $P_n(s) \equiv P[\text{長さ } s \text{ の時間の中で出来事が } n \text{ 回起こる}]$  と定義する. まず  $n = 0$  の場合を調べてみる. (iii) より

$$\begin{aligned} P_0(t+h) &= P[\text{区間 } (0, t+h] \text{ で出来事が起こらない}] \\ &= P[\text{区間 } (0, t] \text{ で出来事が起こらない,} \\ &\quad \text{区間 } (t, t+h] \text{ で出来事が起こらない}] \\ &= P_0(t)P_0(h) \end{aligned}$$

を得る. (i),(ii) より

$$\begin{aligned} &P[(t, t+h] \text{ で出来事が起こらない}] \\ &= 1 - P[(t, t+h] \text{ で出来事が 1 つ以上起こる}] = 1 - \lambda h + o(h) \end{aligned}$$

したがって  $P_0(t+h) = P_0(t)\{1 - \lambda h + o(h)\}$  となるので

$$\lim_{h \rightarrow 0} \frac{P_0(t+h) - P_0(t)}{h} = \lim_{h \rightarrow 0} \frac{P_0(t) - \lambda h P_0(t) - P_0(t)}{h} = -\lambda P_0(t)$$

つまり  $\frac{d}{dt}P_0(t) = -\lambda P_0(t)$  となる. したがって

$$\left. \begin{aligned} \frac{d}{dt}P_0(t) &= -\lambda P_0(t) \\ P_0(0) &= 1 \end{aligned} \right\} \Rightarrow \boxed{P_0(t) = e^{-\lambda t}}$$

次に  $n = 1$  の場合を調べてみる.

$$\begin{aligned} P_1(t+h) &= P[\text{区間 } (0, t+h] \text{ で出来事が一つ起きる}] \\ &= P[\text{区間 } (0, t] \text{ で一つ起き, } (t, t+h] \text{ で一つも起きない}] \\ &\quad + P[\text{区間 } (0, t] \text{ で一つも起きない, } (t, t+h] \text{ で一つ起きる}] \\ &= P_1(t)P_0(h) + P_0(t)P_1(h) \\ &= P_1(t)e^{-\lambda h} + e^{-\lambda t}P_1(h) \end{aligned}$$

従って  $P_1(t+h) - P_1(t) = P_1(t)(e^{-\lambda h} - 1) + e^{-\lambda t}(\lambda h + o(h))$  であるので  $e^{-\lambda h} = 1 - \lambda h + o(h)$  を考慮すると

$$\frac{d}{dt}P_1(t) = \lim_{h \rightarrow 0} \frac{P_1(t+h) - P_1(t)}{h} = -\lambda P_1(t) + \lambda e^{-\lambda t}$$

となる. したがって

$$\left. \begin{aligned} \frac{d}{dt}P_1(t) &= -\lambda P_1(t) + \lambda e^{-\lambda t} \\ P_1(0) &= 0 \end{aligned} \right\} \Rightarrow \boxed{P_1(t) = \lambda t e^{-\lambda t}}$$

ここで微分方程式は定数変化法で解いている. つまり  $P_1(t) = C_1(t)e^{-\lambda t}$  とおくと

$$\begin{aligned} \frac{d}{dt}P_1(t) &= \frac{dC_1(t)}{dt}e^{-\lambda t} - \lambda C_1(t)e^{-\lambda t} \\ &= -\lambda P_1(t) + \frac{dC_1(t)}{dt}e^{-\lambda t} \end{aligned}$$

したがって

$$\lambda e^{-\lambda t} = \frac{dC_1(t)}{dt}e^{-\lambda t} \Rightarrow \lambda = \frac{dC_1(t)}{dt} \Rightarrow C_1(t) = \lambda t + C'$$

ゆえに

$$P_1(t) = (\lambda t + C')e^{-\lambda t}$$

$P_1(0) = 0$  より  $C' = 0$  したがって

$$P_1(t) = \lambda t e^{-\lambda t}$$

が示された. 一般の  $n \geq 2$  の場合は

$$\frac{d}{dt}P_n(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$$

が得られ

$$P_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

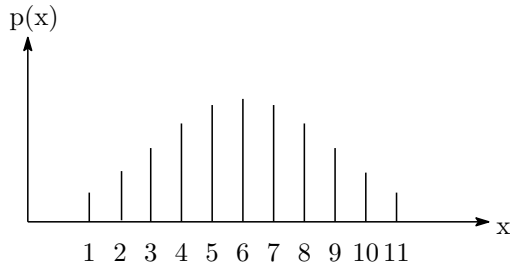
が得られる.

### 章末演習問題

- 【1】  $X$  が  $P[X = 0] = P[X = 1]$  を満たすポアソン分布を持つ確率変数であるとき, 平均  $E[X]$  を求めよ.
- 【2】  $X$  を  $n, p$  をパラメータとする 2 項分布とする. 平均  $E[X] = 5$ , 分散  $V[X] = 4$  が成り立つとき,  $n, p$  の値を求めよ.
- 【3】 平均  $E[X] = 10$ , 標準偏差  $\sigma_X = 3$  が成り立つとき,  $X$  は負の 2 項分布をもつことが可能であるかを調べよ.
- 【4】  $X$  を,  $n = 100, p = 0.1$  をパラメータとする 2 項分布をもつ確率変数とする. このとき,  $P[X \leq \mu_X - 3\sigma_X]$  の値を求めよ.
- 【5】  $X$  がポアソン分布に従い,  $P[X = 0] = \frac{1}{2}$  のとき, 平均  $E[X]$  を求めよ.
- 【6】  $X$  がパラメータ  $n, p$  の 2 項分布とする.  $n$  が固定されていると仮定した場合, 分散  $V[X]$  が極大化されるのは,  $p$  がどんなときかを調べよ.
- 【7】 (a)  $X$  が  $P[X = 1] = P[X = 2]$  のポアソン分布に従うとき,  $P[X = 1]$  または  $2]$  を求めよ.  
(b)  $X$  が平均 1 のポアソン分布をもつ場合,  $E[|X - 1|] = 2\sigma_X/e$  となることを示せ.
- 【8】  $X$  は  $n, p$  をパラメータとする 2 項分布をもち,  $Y$  は  $r, p$  をパラメータとする負の 2 項分布をもつとする.  $F_X(r - 1) = 1 - F_Y(n - r)$  であることを示せ.

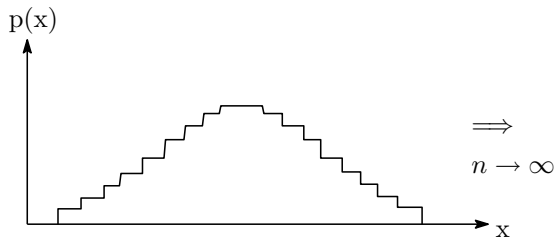
## 4 確率分布 (連続型変数)

離散型分布 (離散型確率変数) との関係

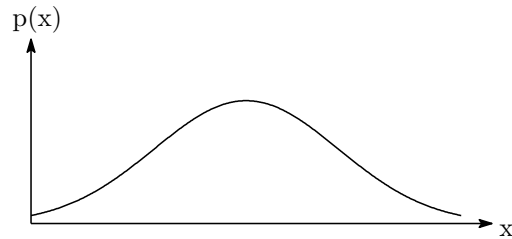


$$\sum_{i=1}^n p(x_i) = 1$$

$$\begin{matrix} x_1 & x_2 & x_3 & \dots & x_n \\ p(x_1) & p(x_2) & p(x_3) & & p(x_n) \end{matrix}$$



離散型分布



連続型分布

期待値 :

$$E[g(X)] = \sum_{x=1}^n g(x)p(x) : \text{和の計算}$$

期待値 :

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)dx : \text{積分の計算}$$

定義 (連続型確率密度)  $p$  が  $\mathbb{R} = (-\infty, \infty)$  上の確率密度であるとは

$$\begin{aligned} (i) \quad & p(x) \geq 0, x \in \mathbb{R} \\ (ii) \quad & \int_{-\infty}^{\infty} p(x)dx = 1 \end{aligned}$$

が成り立つ事である.  $A$  を  $\mathbb{R}$  上の事象としたとき

$$P(A) = \int_A p(x)dx$$

と定義すると  $P$  は確率となる.

連続型確率分布の平均, 2次モーメント, 分散, 標準偏差は次で定義される.

- 平均

$$\mu = \int_{-\infty}^{\infty} xp(x)dx$$

- 二次モーメント

$$m_2 = \int_{-\infty}^{\infty} x^2p(x)dx$$

- 分散

$$V = \int_{-\infty}^{\infty} (x - \mu)^2p(x)dx = m_2 - \mu^2$$

- 標準偏差

$$\sigma = \sqrt{V}$$

- 積率母関数

$$\mu(t) = \int_{-\infty}^{\infty} e^{tx} p(x) dx$$

積率（モーメント）と積率母関数との関係

$$(1) \left. \frac{d}{dt} m_t(x) \right|_{t=0} = \mu \quad (2) \left. \frac{d^2}{dt^2} m_t(X) \right|_{t=0} = m_2$$

(1) の証明を与える. (2) の証明は同様のできるの各自確認しておくように.

$$\begin{aligned} \frac{d}{dt} m_t(X) &= \frac{d}{dt} \left\{ \int_{-\infty}^{\infty} e^{tx} p(x) dx \right\} \\ &= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} p(x) dx \\ &= \int_{-\infty}^{\infty} x e^{tx} p(x) dx \end{aligned}$$

ここで  $t = 0$  とおくと

$$= \int_{-\infty}^{\infty} x p(x) dx = \mu.$$

## 4.1 連続型確率分布の例

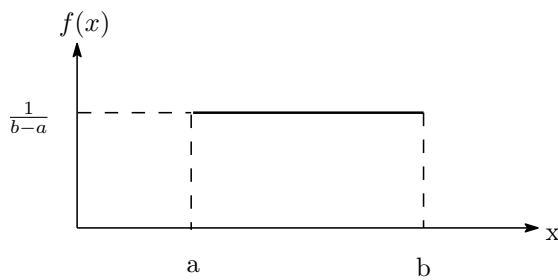
### 4.1.1 一様分布 (Uniform distribution)

$-\infty < a < b < \infty$  に対して  $S = (a, b)$ ,

$$p(x) = \begin{cases} \frac{1}{b-a} & (a < x < b) \\ 0 & (\text{その他}) \end{cases}$$

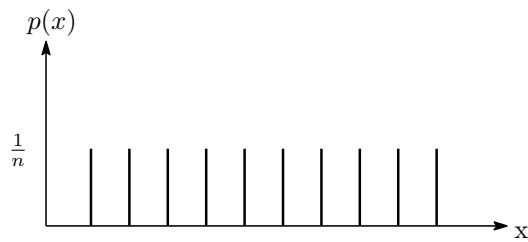
と定義したとき,  $p(x)$  を確率密度とする分布を区間  $(a, b)$  上の一様分布という.

#### 【1】離散一様分布との比較



一様分布

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= \int_a^b \frac{1}{b-a} dx \\ &= \frac{1}{b-a} [x]_a^b \\ &= \frac{b-a}{b-a} \\ &= 1 \end{aligned}$$



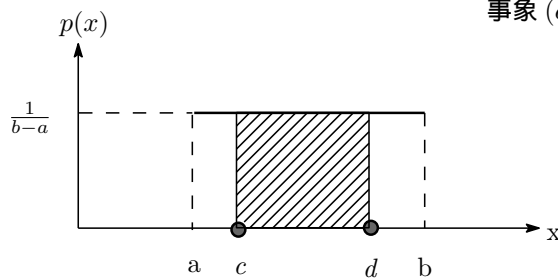
離散一様分布

$$S = \{1, 2, \dots, n\}$$

$$p(x) = \frac{1}{n}, \quad x \in S$$

$$\sum_{x=1}^n p(x) = 1$$

例)  $a < c < d < b$  の時, 区間  $(c, d)$  の確率を計算してみる.



事象  $(c, d)$  の確率 =  $P(c < X < d)$

$$= \int_c^d p(x) dx$$

$$= \frac{1}{b-a} \int_c^d dx$$

$$= \frac{d-c}{b-a}$$

一様分布の平均, 2次モーメント, 分散, 標準偏差の計算

- 平均

$$\mu = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{b-a} \int_a^b xdx = \frac{1}{2} \times \frac{b^2 - a^2}{b-a} = \frac{a+b}{2}$$

- 2次モーメント

$$m_2 = \int_{-\infty}^{\infty} x^2 p(x) dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \times \frac{b^3 - a^3}{3} = \frac{a^2 + ab + b^2}{3}$$

- 分散

$$V = m_2 - \mu^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{a^2 - 2ab + b^2}{12} = \frac{(a-b)^2}{12}$$

- 標準偏差

$$\sigma = \sqrt{V} = \frac{a-b}{2\sqrt{3}}$$

- 積率母関数

$$m_t(X) = \int_{-\infty}^{\infty} e^{tx} f(x) dx = \frac{1}{b-a} \int_a^b e^{tx} dx = \frac{1}{b-a} \left[ \frac{1}{t} e^{tx} \right]_a^b = \frac{e^{tb} - e^{ta}}{(b-a)t}$$

#### 4.1.2 指数分布 (Exponential distribution)

確率密度関数  $p(x)$  が

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

である分布を (パラメータ  $\lambda > 0$  の) 指数分布という.

● 対応するモデル

- (i) 電球の寿命などのさまざまな事物の存続期間
- (ii) 一定の時間間隔における出来事回数がポアソン分布であるとき、つぎつぎに起こる出来事の時間間隔（詳しくは次のガンマ分布のところの説明する。）

指数分布の性質  $X$  の分布が指数分布であるとき

$$(4.1) \quad P(X \geq t+h | X \geq t) = P(X \geq h)$$

が成り立つ。つまり  $X$  を電球の寿命とすると時刻  $t$  まで電球が切れていない時、時刻  $t+h$  まで電球が切れない確率は  $t$  に無関係である。また、(4.1) を満たし、 $P(X \leq h) = h\lambda + o(h)$ ,  $h \rightarrow 0$  が成り立てば  $X$  の分布はパラメータ  $\lambda$  の指数分布である。

証明 前半は演習とする。後半を証明する。

$$\begin{aligned} P(X \geq t+h) &= P(X \geq t+h | X \geq t) \times P(X \geq t) \\ \text{寿命が } t+h \text{ 以上} &\quad \text{寿命が } t \text{ 以上である} \quad \text{寿命が } t \text{ 以上} \\ &\quad \text{条件の下で } t+h \text{ 以上} \\ &= P(X \geq h)P(X \geq t) \end{aligned}$$

をえる。両辺を微分すると

$$\begin{aligned} \frac{d}{dt}P(X \geq t) &= \lim_{h \rightarrow 0} \frac{P(X \geq t+h) - P(X \geq t)}{h} \\ &= P(X \geq t) \lim_{h \rightarrow 0} \frac{P(X \geq h) - 1}{h} \\ &\quad \downarrow \\ &\quad -\lambda \\ &= -\lambda P(X \geq t) \end{aligned}$$

となる。したがって  $P(X \geq t) = ce^{-\lambda t}$  が導かれるが、 $t=0$  のとき  $P(X \geq 0) = 1$  であることに注意すると  $c=1$  をえるので

$$P(X \geq t) = e^{-\lambda t} = \int_t^{\infty} p(x)dx$$

両辺を  $t$  で微分すれば  $-\lambda e^{-\lambda t} = -p(t)$  つまり  $p(t) = \lambda e^{-\lambda t}$  をえる。

指数分布の平均、2次モーメント、分散、標準偏差を計算するためには次の部分積分の公式を用いる。（ $t$  の範囲について注意するように。）

部分積分の公式

$$\int_a^b g(x)h(x)dx = [G(x)h(x)]_a^b - \int_a^b G(x)h'(x)dx$$

ここで  $G(x) = \int_0^x g(x)dx$ .

● 平均

$$\mu = \int_0^{\infty} p(x)x dx = [-xe^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^{\infty} = \frac{1}{\lambda}$$

- 2次モーメント

$$m_2 = \int_0^{\infty} p(x)x^2 dx = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx = [-x^2 e^{-\lambda x}]_0^{\infty} + 2 \int_0^{\infty} x e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

- 分散

$$V = m_2 - \mu^2 = \frac{1}{\lambda^2}$$

- 標準偏差

$$\sigma = \sqrt{V} = \frac{1}{\lambda}$$

**演習** 指数分布の積率母関数を計算し、その結果から平均、2次モーメント、分散を求めよ。

#### 4.1.3 ガンマ分布 ( $\Gamma$ -distribution)

確率密度関数  $p(x)$

$$p(x) = \begin{cases} \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

を持つ分布を (パラメータ  $\lambda > 0, r > 0$  の) ガンマ分布という。

ここで  $\Gamma(r)$  は、ガンマ関数

$$\Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x} dx = \int_0^{\infty} \lambda (\lambda x)^{r-1} e^{-\lambda x} dx$$

であり、

$$\begin{aligned} \Gamma(n+1) &= n! && : n \text{ は } 0 \text{ 以上の整数} \\ \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi} && : \left(-\frac{1}{2}\right)! = \sqrt{\pi} \\ \Gamma(r+1) &= r\Gamma(r) && : r > 0 \end{aligned}$$

という性質が知られている。

注) 指数分布はガンマ分布の  $r = 1$  の場合に対応する。

- 対応するモデル ( $r$  個の電球の寿命の和)  $i$  番目の電球の寿命を  $X_i$  とする。 ( $X_i$  の分布は指数分布) このとき  $Y = \sum_{i=1}^r X_i$  の分布はガンマ分布となる。

ガンマ分布の平均、分散、積率母関数の計算。

まず積率母関数を計算してから、平均、分散を計算する。

- 積率母関数  $t < \lambda$  のとき

$$\begin{aligned} m_t(Y) &= \int_0^{\infty} p(x)e^{tx} dx = \int_0^{\infty} \frac{\lambda^r}{\Gamma(r)} e^{tx} x^{r-1} e^{-\lambda x} dx \\ &= \left(\frac{\lambda}{\lambda-t}\right)^r \int_0^{\infty} \frac{(\lambda-t)^r}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} dx = \left(\frac{\lambda}{\lambda-t}\right)^r \end{aligned}$$

- 平均

$$\begin{aligned}\mu &= \left. \frac{\partial}{\partial t} m_t(Y) \right|_{t=0} = \left. \frac{\partial}{\partial t} \left( \frac{\lambda}{\lambda - t} \right)^r \right|_{t=0} \\ &= \lambda^r \left. \frac{\partial}{\partial t} (\lambda - t)^{-r} \right|_{t=0} = r\lambda^r (\lambda - t)^{-r-1} \Big|_{t=0} = \frac{r}{\lambda}\end{aligned}$$

- 2次モーメント

$$\begin{aligned}m_2 &= \left. \frac{\partial^2}{\partial t^2} m_t(Y) \right|_{t=0} = \left. \frac{\partial}{\partial t} \{ r\lambda^r (\lambda - t)^{-r-1} \} \right|_{t=0} \\ &= r\lambda^r (r+1) \lambda^{-r-2} = \frac{r(r+1)}{\lambda^2}\end{aligned}$$

- 分散

$$V(Y) = m_2 - \mu^2 = \frac{r(r+1)}{\lambda^2} - \left( \frac{r}{\lambda} \right)^2 = \frac{r}{\lambda^2}$$

ガンマ分布と指数分布の関係を表にしてみる.

	ガンマ分布	指数分布
積率母関数 : $m_t(Y)$	$\left( \frac{\lambda}{\lambda - t} \right)^r$	$\frac{\lambda}{\lambda - t}$
平均 : $\mu$	$\frac{r}{\lambda}$	$\frac{1}{\lambda}$
分散 : $m_2$	$\frac{r}{\lambda^2}$	$\frac{1}{\lambda^2}$

ポアソン分布との関係  $Z$  をパラメータ  $\lambda > 0$  のポアソン分布を持つ確率変数とする. つまり

$$P(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

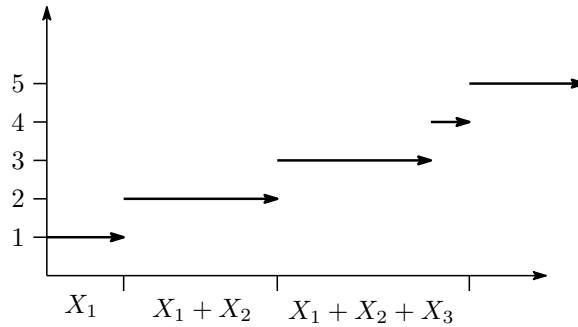
が成り立つ. また,  $X_i, i = 1, 2, \dots, k$  を互いに独立であり各々が同じパラメータ  $\lambda$  の指数分布を持つ確率変数とする. すでに述べたように  $X_1 + X_2 + \dots + X_k = Y$  はパラメータ  $(\lambda, k)$  のガンマ分布をもつ.  $Y$  が 1 以下である確率を計算してみると

$$\begin{aligned}P(X_1 + X_2 + \dots + X_k \leq 1) &= P(Y \leq 1) \\ &= \int_0^1 \frac{\lambda(\lambda x)^{k-1}}{\Gamma(k)} e^{-\lambda x} dx = \frac{\lambda^k}{\Gamma(k)} \int_0^1 x^{k-1} e^{-\lambda x} dx \\ &= -\frac{\lambda^{k-1}}{\Gamma(k)} e^{-\lambda} + \frac{\lambda^{k-1}}{\Gamma(k-1)} \int_0^1 x^{k-2} e^{-\lambda x} dx \\ &= 1 - \sum_{j=1}^k \frac{\lambda^{j-1}}{\Gamma(j)} e^{-\lambda} = \sum_{j=k}^{\infty} \frac{\lambda^j}{\Gamma(j+1)} e^{-\lambda}\end{aligned}$$

となる. したがって

$$\begin{aligned}P(X_1 + X_2 + \dots + X_n \leq 1 < X_1 + X_2 + \dots + X_{n+1}) \\ = P(X_1 + X_2 + \dots + X_n \leq 1) - P(X_1 + X_2 + \dots + X_{n+1} \leq 1) = P(Z = n)\end{aligned}$$

が成り立つ. 電球の寿命のモデルでこの関係を説明してみると, 「時刻 1 までに切れる電球の個数の分布はポアソン分布である」ということになる.



#### 4.1.4 ベータ分布 (Beta distribution)

$\alpha > 0, \beta > 0$  に対して,

$$p(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} & x \in S = (0, 1) \\ 0 & (\text{その他}) \end{cases}$$

と定義したとき,  $p(x)$  を確率密度とする分布をベータ分布という.  $B(\alpha, \beta)$  は積分して 1 にするための規格化定数であって,

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

で定義される. これをベータ関数とよび, ガンマ関数  $\Gamma(x)$  と

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

という関係をもつ. とくに  $\alpha = 1, \beta = 1$  のときには, ベータ分布が一様分布と一致することがわかる. ベータ関数の特徴は  $\alpha, \beta$  の値によっていろいろな形をとることである. 確率分布のおおよその形がわかっているその形に関数をあてはめたいときに, ベータ関数が用いられる. 確率密度関数の形より平均, 分散は簡単に計算できる.

- 平均  $\mu = \frac{\alpha}{\alpha + \beta}$ .
- 分散  $V = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$ .

#### 4.1.5 正規分布 (Normal distribution)

$\sigma > 0, -\infty < \mu < \infty$  とする. 確率密度関数  $p(x)$  が

$$p(x) = p_{\mu\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in S = (-\infty, \infty)$$

で定まる分布をパラメータ  $\mu, \sigma$  の正規分布という. ( $N(\mu, \sigma^2)$  と書く)  $\mu = 0, \sigma = 1$  のときの正規分布を標準正規分布 (規準正規分布) という. このとき確率密度関数は,

$$p(x) = p_{01}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad x \in S = (-\infty, \infty)$$

となる.

まず  $p(x)$  が確率密度関数であること, つまり

$$\int_{-\infty}^{\infty} p(x)dx = 1,$$

を示しておく.

証明 変数変換  $y = \frac{x-\mu}{\sigma}$  を用いると

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{y^2}{2}\right\} \sigma dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \end{aligned}$$

となるので,

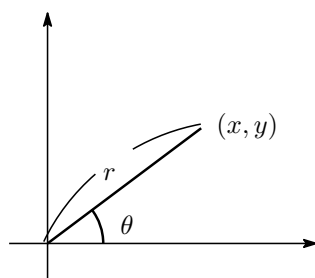
$$\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$$

を示せばよい.

$$\begin{aligned} \left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right)^2 &= \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \times \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2+y^2}{2}\right\} dx dy \\ &= \int_0^{2\pi} \int_0^{\infty} \exp\left\{-\frac{r^2}{2}\right\} r dr d\theta \quad \leftarrow \text{極座標 ( )} \\ &= 2\pi \int_0^{\infty} r e^{-\frac{r^2}{2}} dr \\ &= 2\pi [-e^{-\frac{r^2}{2}}]_0^{\infty} \\ &= 2\pi \end{aligned}$$

となり, 極座標への変数変換を用いると計算できる.

極座標変換とヤコビアン



$$\left. \begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned} \right\} x^2 + y^2 = r^2$$

$$dx dy = r dr d\theta \quad ( )$$

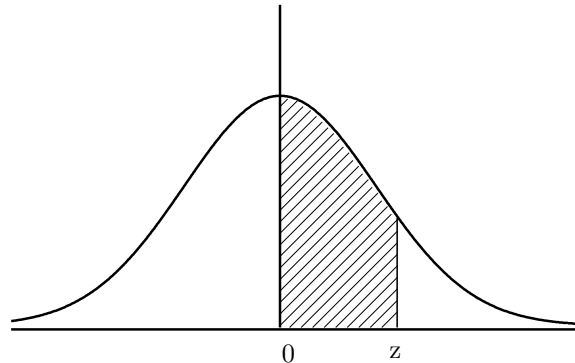
$$\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r \cos^2 \theta + r \sin^2 \theta = r$$

区間  $[a, b]$  の確率

$$P_{\mu\sigma}([a, b]) = \int_a^b p_{\mu\sigma}(x) dx$$

は一般の  $a, b$  で積分は計算できない. そのため標準正規分布表を参考にする.

標準正規分布表は  $F(z) = \int_0^{|z|} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx$  の値をまとめている表である.



例)

$$F(1) = 0.3413, \quad F(2) = 0.4772, \quad F(3) = 0.4987.$$

$F(x)$  は次の性質を持つ.

$$F(z) = F(-z), \quad F(\infty) + F(-\infty) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = 1$$

このことから  $F(\infty) = \frac{1}{2}$  がわかる. そして標準正規分布に対して区間  $[a, b]$  の確率は以下の性質を用いて表わすことができる.

$(0 \leq a < b)$  のとき

$$\begin{aligned} P_{01}([a, b]) &= \int_a^b p_{01}(x) dx = \int_0^b p_{01}(x) dx - \int_0^a p_{01}(x) dx \\ &= F(b) - F(a) \end{aligned}$$

$(a < b \leq 0)$  のとき

$$\begin{aligned} P_{01}([a, b]) &= \int_a^b p_{01}(x) dx = \int_{-b}^{-a} p_{01}(x) dx \\ &= F(-a) - F(-b) \\ &= F(a) - F(b) \end{aligned}$$

$(a \leq 0 \leq b)$  のとき

$$P_{01}([a, b]) = \int_a^b p_{01}(x) dx = F(a) + F(b)$$

【2】例)

$$P_{01}([1, 2]) = F(2) - F(1) = 0.4772 - 0.3413 = 0.1359$$

$$P_{01}([-1, 2]) = F(-1) + F(2) = 0.4772 + 0.3413 = 0.8185$$

$$P_{01}([1, \infty]) = F(\infty) - F(1) = 0.5 - 0.3413 = 0.1587$$

【3】一般の  $\mu$  と  $\sigma$  の場合は標準化（規準化）を用いて確率を計算する.

【4】定理（標準化, 規準化）

$X$  の分布がパラメータ  $\mu, \sigma$  の正規分布  $N(\mu, \sigma^2)$  であるとき,

$Y = \frac{X - \mu}{\sigma}$  の分布は標準正規分布  $N(0, 1)$  である.

証明)

$y = \frac{x - \mu}{\sigma}$  として変数変換を行うと,

$$\begin{aligned} P(a \leq Y \leq b) &= P(a\sigma + \mu \leq X \leq b\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{a\sigma + \mu}^{b\sigma + \mu} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left\{-\frac{y^2}{2}\right\} dy \end{aligned}$$

例)

$$\int_{\mu}^{\mu + k\sigma} p_{\mu\sigma}(x) dx = F(k), \quad k = 1, 2, \dots,$$

平均, 分散, 標準偏差の計算は積分公式

$$(1) \int_{-\infty}^{\infty} y \exp\left\{-\frac{y^2}{2}\right\} dy = 0$$

$$(2) \int_{-\infty}^{\infty} y^2 \exp\left\{-\frac{y^2}{2}\right\} dy = \int_{-\infty}^{\infty} \exp\left\{-\frac{y^2}{2}\right\} dy = \sqrt{2\pi}$$

を用いて示される. (1) は  $y \exp\left\{-\frac{y^2}{2}\right\}$  が奇関数であり

$$\int_{-\infty}^0 y \exp\left\{-\frac{y^2}{2}\right\} dy = - \int_0^{\infty} y \exp\left\{-\frac{y^2}{2}\right\} dy$$

が成り立つことより得られ, (2) は  $g(y) = y \exp\left\{-\frac{y^2}{2}\right\}$ ,  $h(y) = y$  として部分積分の公式を用いると,

$$\int_{-\infty}^{\infty} y^2 \exp\left\{-\frac{y^2}{2}\right\} dy = \left[-y \exp\left\{-\frac{y^2}{2}\right\}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\left\{-\frac{y^2}{2}\right\} dy$$

となることより得られる.

- 平均

$y = \frac{x - \mu}{\sigma}$  で変数変換を行うと ( $dy = \frac{1}{\sigma} dx, x = \mu + \sigma y$ )

$$\begin{aligned} \int_{-\infty}^{\infty} x p_{\mu\sigma}(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma y) \exp\left\{-\frac{y^2}{2}\right\} dy \\ &= \mu \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{y^2}{2}\right\} dy + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y \exp\left\{-\frac{y^2}{2}\right\} dy \\ &= \mu \end{aligned}$$

を得る. 最後の等号を導くために積分公式 (1) を用いた. 従って,

$$\boxed{N(\mu, \sigma^2) \text{ の平均} = \mu}$$

• 分散

$y = \frac{x - \mu}{\sigma}$  で変数変換を行うと,

$$\begin{aligned} \int_{-\infty}^{\infty} (x - \mu)^2 p_{\mu\sigma}(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 y^2 \exp\left\{-\frac{y^2}{2}\right\} dy \\ &= \sigma^2 \end{aligned}$$

を得る. 最後の等式を導くために積分公式 (2) を用いた. 従って,

$N(\mu, \sigma^2)$ の分散	: $\sigma^2$
$N(\mu, \sigma^2)$ の標準偏差	: $\sigma$

• 積率母関数

$$\begin{aligned} \int_{-\infty}^{\infty} e^{tx} p_{\mu\sigma}(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left\{tx - \frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ & \hspace{15em} ( ) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}\right\} \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} dx \\ &= \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}\right\} dx \\ & \hspace{15em} \downarrow \\ & \hspace{15em} 1 \\ &= \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} \end{aligned}$$

$$\begin{aligned} ( ) \quad tx - \frac{(x-\mu)^2}{2\sigma^2} &= -\frac{1}{2\sigma^2} \{(x-\mu)^2 - 2\sigma^2 tx\} \\ &= -\frac{1}{2\sigma^2} (x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx) \\ &= -\frac{1}{2\sigma^2} \{(x-\mu-\sigma^2 t)^2 - 2\mu\sigma^2 t - \sigma^4 t^2\} \\ &= -\frac{1}{2\sigma^2} (x-\mu-\sigma^2 t)^2 + \frac{\sigma^2 t^2}{2} + \mu t \end{aligned}$$

**演習** インフルエンザの予防注射をした人のうち 5% は注射に対してアレルギー反応を示すという. 注射した 200 人中 8% 以上の人アレルギー反応を示す確率を正規分布で近似して求めよ.

4.1.6 その他の確率分布

【1】コーシー分布 (Cauchy distribution)  $\alpha > 0, \lambda \in (-\infty, \infty)$  に対して,

$$p(x) = \begin{cases} \frac{\alpha}{\pi\{\alpha^2 + (x - \lambda)^2\}} & x \in S = (-\infty, \infty) \\ 0 & (\text{その他}) \end{cases}$$

と定義したとき,  $p(x)$  を確率密度とする分布をコーシー分布という. コーシー分布の密度関数の形は正規分布のそれと似ているが, 詳しく調べるとまったく異なっている. 最も大きく異なることは, 平均も分散も存在しないことである.

【2】対数正規分布 (log-normal distribution)  $\mu \in (-\infty, \infty), \sigma > 0$  に対して,

$$p(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} & x \in S = (0, \infty) \\ 0 & (\text{その他}) \end{cases}$$

と定義したとき,  $p(x)$  を確率密度とする分布を対数正規分布という. その平均と分散は

$$\text{平均 } \mu = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$\text{分散 } V = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2),$$

である.

ランダムに世帯を選びその年間所得  $X$  を調べると, 低い方は一定限度があるが高い方には明確な限度がない. このような場合は, 対数をとると有効である. この例では  $\log X$  の分布が正規分布に近いことが知られている.  $\log X$  が正規分布に従うときもとの  $X$  は対数正規分布に従う.

【3】ワイブル分布 (Weibul distribution)  $a, b > 0$  に対して,

$$p(x) = \begin{cases} \frac{bx^{b-1}}{a^b} \exp\left\{-\left(\frac{x}{a}\right)^b\right\} & x \in S = (0, \infty) \\ 0 & (\text{その他}) \end{cases}$$

と定義したとき,  $p(x)$  を確率密度とする分布をワイブル分布という. 一般に耐用年数や寿命は確率変数だが, 故障が偶発故障なら瞬間故障率は一定になり, 確率変数は指数分布に従う. もし劣化が進行し故障率が増加するときは IFR (Increasing Failure Rate) といい, 指数分布に従わない. またいわゆる「初期故障」の時期には故障率の減少が起こる. これを DFR (Decreasing Failure Rate) といい, このときも指数分布に従わない. これらの現象での分布はワイブル分布に近いことが知られている.

ワイブル分布の平均と分散は

$$\text{平均 } \mu = a\Gamma\left(1 + \frac{1}{b}\right),$$

$$\text{分散 } V = a^2\left\{\Gamma\left(2 + \frac{1}{b}\right) - \Gamma\left(1 + \frac{1}{b}\right)^2\right\},$$

である. 母数  $a, b$  はそれぞれ尺度母数, 形状母数とよばれ  $b$  の値をかえると分布の形が変化する.  $b$  が大きいときのワイブル分布は正規分布に近づく. 正規分布に似ているが厳密には正規分布ではない場合の精密なあてはめにも用いられる.

### 演習

【1】 $X$  が平均 12, 標準偏差 2 の正規分布に従うとき, 標準正規分布表を用いて次の確率を求めよ. (a)  $X > 14$ , (b)  $X > 11$ , (c)  $X < 10$ , (d)  $X < 10.5$ , (e)  $10 < X < 13$ .

【2】男子学生の身長  $X$  は平均 69 インチ, 標準偏差 3 インチの正規分布に従うと仮定して, 標準正規分布表より次の確率を求めよ. (a)  $X < 66$  インチ, (b)  $65$  インチ  $< X < 71$  インチ.

- 【3】高校のある体育教師が、個々の生徒の体育実技の成績評価は全生徒の成績を考慮して相対的に行うと宣言した。この教師は全生徒の 20% に A をつけたいとする。過去の経験から、走り高跳びの平均は 4 フィート 10 インチで標準偏差は 4 インチであったとすると、生徒が A をもらうにはどのくらいの高さをとばねばならないか。
- 【4】ある大学の学生の I.Q. が平均 115, 標準偏差 8 の正規分布に従うとして, I.Q. が (a) 130 以上, (b) 100 未満, (c) 105 から 125 の間, の学生の百分率をそれぞれ求めよ。
- 【5】基礎英語の試験の得点分布は平均 130 点, 標準偏差 20 点の正規分布にほぼ近い形をしていた。100 点以上を合格とするとき, この試験で不合格になる学生の百分率はいくらか。

### 章末演習問題

- 【1】 $X$  が  $(1, 2)$  の範囲の上で一様分布に従う場合,  $P[X > z + \mu_X] = \frac{1}{4}$  となる  $z$  を求めよ。
- 【2】 $X$  が平均 2, 分散 1 の正規分布とすると,  $P[|X - 2| < 1]$  を求めよ。
- 【3】 $X$  が平均 2 の指数分布をもつとき, 条件付確率  $P[X < 1 | X < 2]$  を求めよ。
- 【4】 $X$  がパラメータ  $\lambda$  の指数分布をもつとする。  $P[X \leq 1] = P[X > 1]$  であるとき, 分散  $V[X]$  を求めよ。
- 【5】 $X$  が平均 1, 分散  $4/3$  の一様分布に従う連続型確率変数であるとき,  $P[X < 0]$  を求めよ。
- 【6】 $X$  が  $\exp(e^t - 1)$  なる積率母関数をもつ確率変数であるとき,  $E[X]$  を求めよ。
- 【7】2 項分布である確率変数  $X$  について

$$P[X \geq k] = \sum_{j=k}^n \binom{n}{j} p^j q^{n-j} = \frac{1}{B(k, n-k+1)} \int_0^p u^{k-1} (1-u)^{n-k} du$$

であることを示せ。すなわち,  $X$  が,  $n, p$  をパラメータとする 2 項分布であり,  $Y$  が  $k, n-k+1$  をパラメータとするベータ分布であれば,  $F_Y(p) = 1 - F_X(k-1)$  となることを示せ。ここで  $F_Y(y) = P(Y \leq y)$ ,  $F_X(x) = P(X \leq x)$  である。

## 5 2次元確率分布

### 5.1 同時確率分布と周辺確率分布

二つの確率変数  $X, Y$  があるとし, 2次元のベクトル  $(X, Y)$  を考える. 2変数を同時に考えるのは, それらの間に互いに関係があると考えているからである. とりあえずは,  $X, Y$  は離散型とする.  $X = x$  であり同時に  $Y = y$  である確率

$$P(X = x, Y = y) = p(x, y)$$

を, 2次元確率変数  $(X, Y)$  の同時離散型確率密度関数 (joint probability density function) という.  $p(x, y)$  は, 1次元のときと同じく

$$p(x, y) \geq 0 \quad \text{かつ} \quad \sum_x \sum_y p(x, y) = 1$$

を満たさなければならない.

2次元の確率変数の場合, 事象も2次元空間の部分集合である. 事象  $A$  の確率  $P(A)$  は

$$P((X, Y) \in A) = \sum_{(x, y) \in A} p(x, y)$$

で与えられる.

$X, Y$  が連続型の確率変数の時には  $p(x, y)$  は2次元の確率密度関数で, 同時確率密度関数 (joint probability density function) と呼ばれ

$$p(x, y) \geq 0 \quad \text{かつ} \quad \int \int_S p(x, y) dx dy = 1$$

を満たす. ここで標本空間  $S$  は2次元ユークリッド空間 (平面) の全範囲のことである.  $p(x, y)$  によって, 事象  $A$  の確率は, 積分で

$$P((X, Y) \in A) = \int \int_A p(x, y) dx dy$$

と定義される. とくに,  $A$  が区間  $[a, b] \times [c, d]$  ならば次のようになる.

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d p(x, y) dx dy$$

である.

同時確率分布から,  $X, Y$  単独の確率分布が

$$g(x) = \sum_y p(x, y), \quad h(y) = \sum_x p(x, y)$$

で求められる. 周辺にあるから, それぞれ  $X, Y$  の周辺確率分布 (marginal probability distribution) と呼ばれる. 連続型の場合も,  $X, Y$  の単独確率密度関数は同時確率密度関数から

$$g(x) = \int_{-\infty}^{\infty} p(x, y) dy, \quad h(y) = \int_{-\infty}^{\infty} p(x, y) dx$$

で与えられる. これらを周辺確率密度関数 (marginal probability density function) という. この場合も周辺確率密度関数が与える確率分布を周辺確率分布という.

(例) 二つのさいころの目  $X_1, X_2$  の最大 (大きい方), 最小 (小さい方) をそれぞれ  $X = \max(X_1, X_2), Y = \min(X_1, X_2)$  としよう. ただし2数が等しいときのために,  $\max(x, y) = \min(x, y) = x$  と約束する. このとき,  $X = x$  であり同時に  $Y = y$  となる確率  $p(x, y)$  は下のように縦横の表になる. たとえば  $(X, Y) = (4, 3)$  は  $(X_1, X_2) = (3, 4), (4, 3)$  に対応するからその確率は  $\frac{1}{36} + \frac{1}{36} = \frac{2}{36}$  となる.

	X	1	2	3	4	5	6	$h(y)$
Y								
$p(x, y)$	1	1/36	2/36	2/36	2/36	2/36	2/36	11/36
	2	0	1/36	2/36	2/36	2/36	2/36	9/36
	3	0	0	1/36	2/36	2/36	2/36	7/36
	4	0	0	0	1/36	2/36	2/36	5/36
	5	0	0	0	0	1/36	2/36	3/36
	6	0	0	0	0	0	1/36	1/36
$g(x)$		1/36	3/36	5/36	7/36	9/36	11/36	1

### 2個のさいころの大きい方と小さい方の確率分布

周辺確率分布は同時確率分布から導かれる。逆はそうではない。表で  $P(X = 3) = 5/36, P(Y = 2) = 9/36$  から  $P(X = 3, Y = 2) = 1/18$  を直接に導くことはできない。このように、 $g(x), h(x)$  から、 $p(x, y)$  を求めることはできない。なぜなら、同じ  $g(x), h(y)$  を与える  $p(x, y)$  は無限にあるからである。一通りの  $p(x, y)$  が与えられたときはじめて  $X, Y$  の関係が定まる。

## 5.2 共分散と相関係数

2変数  $X, Y$  の間に関連があれば、一方の変化は他方に及ぶと考えられるから、ばらつきの指標としても分散には、単純な加法が成立しないことは想像できる。定義に基づいて計算してみると、公式  $(a+b)^2 = a^2 + 2ab + b^2$  より

$$\begin{aligned}
 V(X+Y) &= E[(X+Y - \mu_X - \mu_Y)^2] \\
 &= E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] \\
 &= V(X) + V(Y) + 2Cov(X, Y)
 \end{aligned}$$

となる。ただし、

$$Cov(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\} \quad (\mu_X = E(X), \mu_Y = E(Y))$$

である。 $Cov(X, Y)$  は  $X, Y$  の共分散 (covariance) と呼ばれ、 $X$  と  $Y$  が、それぞれの平均  $\mu_X, \mu_Y$  から互いに関連しながら、ばらつく程度を表す。分散の加法性  $V(X+Y) = V(X) + V(Y)$  が必ずしも成立しないのは、 $X+Y$  のばらつきには、 $X, Y$  単独のばらつきのほかに相互関連によるばらつき  $Cov(X, Y)$  が存在し、それを入れてはじめて等号が成立するからである。 $X - \mu_X, Y - \mu_Y$  の正負の符号の全体 (平均) 的傾向から、 $Cov(X, Y) > 0$  なら、 $X, Y$  は大小が同傾向、 $< 0$  なら反対傾向の関係となる。

株式投資を例にして考えてみる。A 石油の株価を  $X$ , B 石油の株価  $Y$  とおく。同一業種の株価はエネルギー危機など共通の経済的要因によって同傾向に連動するから、 $Cov(X, Y) > 0$  となり、単独の分散 (ばらつき) の和  $V(X) + V(Y)$  以上にばらつくからである。つまり、変動のリスクが連動の分だけ大きくなる。

共分散  $Cov(X, Y)$  は  $X, Y$  の関係の方向を表すが、その強さの程度を判断する基準がない。そこで、この値を標準偏差で割って調整し確率変数  $X, Y$  の相関係数 (correlation coefficient) を

$$\rho_{XY} = Cov(X, Y) / \sqrt{V(X)} \sqrt{V(Y)}$$

と定義する。

注意  $\rho_{XY}$  は必ず  $-1 \leq \rho_{XY} \leq 1$  の範囲に入る. この性質は次のようにして示すことができる.  $t$  の 2 次式

$$\begin{aligned} Q(t) &= V(tX + Y) \\ &= E[(tX + Y - E(tX + Y))^2] \\ &= t^2V(X) + 2tCov(X, Y) + V(Y) \end{aligned}$$

を考えると, これは負にならない 2 次式であるから判別式  $\leq 0$  でなければならない. このことから  $(Cov(X, Y))^2 \leq V(X) \cdot V(Y)$  がただちに導かれる.

以下,  $\rho_{XY}$  を簡単に  $\rho$  と表すことにする.  $\rho$  の定義から  $\rho > 0$  なら  $X, Y$  は同じ大小の向きに変化する傾向があり,  $\rho < 0$  なら逆である. ここでいう傾向は平均的, 確率的傾向であるが,  $|\rho|$  が大きくなると確定的な関係に近づく. もっとも極端な場合は,  $\rho = \pm 1$  であり, このときは  $X, Y$  の間には厳密に次の 1 次式の関係が成り立つ.

$$Y = aX + b \quad (\text{ただし, } \rho = 1 \text{ なら } a > 0, \rho = -1 \text{ なら } a < 0)$$

逆に  $\rho = 0$  (つまり  $Cov(X, Y) = 0$ ) の場合は,  $X, Y$  はどちらの関係を持つともいえない. この場合,  $X, Y$  は無相関 (uncorrelated) であるという. 無相関とは「関連がない」ということの 1 つの表現である. 独立であれば無相関である. しかし無相関であっても独立であるとはかぎらない.

$\rho$  の計算に必要な  $X, Y$  の共分散は, 同時確率密度関数を用いて

$$Cov(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) \cdot p(x, y) \quad (\text{離散型})$$

$$Cov(X, Y) = \int \int_S (x - \mu_X)(y - \mu_Y) \cdot p(x, y) dx dy \quad (\text{連続型})$$

と表されるが,

$$(5.1) \quad Cov(X, Y) = E(XY) - \mu_x \mu_y = E(XY) - E(X)E(Y)$$

となる (各自確認) ことに注意すると,

$$E(XY) = \sum_x \sum_y xy \cdot p(x, y), \quad E(X)E(Y) = \sum_x \sum_y x \cdot p(x, y) \sum_x \sum_y y \cdot p(x, y),$$

$$E(XY) = \int \int_S xy \cdot p(x, y) dx dy, \quad E(X)E(Y) = \int \int_S x \cdot p(x, y) dx dy \int \int_S y \cdot p(x, y) dx dy$$

から計算することができる.

	X	1	2	3	4	5	6	$h(y)$
	Y							
$p(x, y)$	1	1/36	2/36	2/36	2/36	2/36	2/36	11/36
	2	0	1/36	2/36	2/36	2/36	2/36	9/36
	3	0	0	1/36	2/36	2/36	2/36	7/36
	4	0	0	0	1/36	2/36	2/36	5/36
	5	0	0	0	0	1/36	2/36	3/36
	6	0	0	0	0	0	1/36	1/36
	$g(x)$	1/36	3/36	5/36	7/36	9/36	11/36	1

2 個のさいころの大きい方と小さい方の確率分布

(例) 表の  $X, Y$  の相関係数を得るために、定義に従い順次計算してみよう。  $X$  の周辺確率分布  $g(x)$  から、その平均、分散

$$\begin{aligned} E(X) &= \frac{1}{36}\{1 \times 1 + 2 \times 3 + 3 \times 5 + 4 \times 7 + 5 \times 9 + 6 \times 11\} = 161/36, \\ E(X^2) &= \frac{1}{36}\{1 \times 1 + 4 \times 3 + 9 \times 5 + 16 \times 7 + 25 \times 9 + 36 \times 11\} = 791/36, \\ V(X) &= 791/36 - (161/36)^2 = 2555/36^2, \end{aligned}$$

$Y$  の周辺確率分布  $h(y)$  から、その平均、分散

$$\begin{aligned} E(Y) &= \frac{1}{36}\{1 \times 11 + 2 \times 9 + 3 \times 7 + 4 \times 5 + 5 \times 3 + 6 \times 1\} = 91/36, \\ E(Y^2) &= \frac{1}{36}\{1 \times 11 + 4 \times 9 + 9 \times 7 + 16 \times 5 + 25 \times 3 + 36 \times 1\} = 301/36, \\ V(Y) &= 301/36 - (91/36)^2 = 2555/36^2, \end{aligned}$$

$(X, Y)$  の同時確率分布  $p(x, y)$  から

$$\begin{aligned} E(XY) &= \sum_{x,y=1}^6 xy \cdot p(x, y) = 441/36, \\ Cov(X, Y) &= 441/36 - (161/36)(91/36) = 1225/36^2 \end{aligned}$$

を得る。いまは  $V(X) = V(Y)$  であるから、相関係数は、結局

$$\rho = (1225/36^2)/(2555/36^2) = 1225/2555 = 0.4795$$

となる。

(例) 同時確率密度関数

$$p(x, y) = \begin{cases} 6(x-y) & 0 \leq y < x \leq 1 \\ 0 & \text{それ以外} \end{cases}$$

をもつ  $X, Y$  の周辺確率分布、平均、分散、および、 $X, Y$  の相関係数を求めよう。

$y$  については  $0 \leq y < x$  以外で被積分関数は 0 であること ( $x$  についても同様) に注意して

$$g(x) = \int_0^x 6(x-y)dy = 3x^2, \quad h(y) = \int_y^1 6(x-y)dx = 3(1-y)^2$$

となる。これらの密度関数は  $1/2$  について線対称であるので分散は同じである。平均は

$$\begin{aligned} E(X) &= \int_0^1 x \cdot 3x^2 dx = 3/4 = 0.75, \\ E(Y) &= \int_0^1 y \cdot 3(1-y)^2 dy = 1/4 = 0.25 \end{aligned}$$

である。また、分散は、 $E(X^2) = \int_0^1 x^2 \cdot 3x^2 dx = 3/5$  から

$$V(X) = 3/5 - (3/4)^2 = 3/80, \quad V(Y) = 3/80$$

したがって、 $\sigma(X) = \sigma(Y) = \sqrt{3/80} = 0.194$  である。

さらに、共分散は次のように計算される。

$$\begin{aligned}
 E(XY) &= \int_0^1 \int_y^1 xy \cdot 6(x-y) dy dx \\
 &= \int_0^1 y \left\{ \int_y^1 (6x^2 - 6xy) dx \right\} dy \\
 &= \int_0^1 y [2x^3 - 3y \cdot x^2]_y^1 dy = \int_0^1 y \{2 - 3y - (2y^3 - 3y^3)\} dy \\
 &= \int_0^1 (y^4 - 3y^2 + 2y) dy = 1/5 \\
 Cov(X, Y) &= 1/5 - (3/4)(1/4) = 1/80
 \end{aligned}$$

したがって、 $X$  と  $Y$  の相関係数は

$$\rho = (1/80) / (\sqrt{3/80} \cdot \sqrt{3/80}) = 1/3 = 0.333$$

である。

**演習**  $[0,1]$  上の一様乱数を 3 個とり、それを小さいほうから  $X_{(1)}, X_{(2)}, X_{(3)}$  とし、 $X = X_{(3)}, Y = X_{(1)}$  としよう。 $X$  は最大値、 $Y$  は最小値であるが、この同時確率分布が  $p(x, y)$  であることを示せ。

この  $p(x, y)$  は一様分布からのレコード値 (極値) の確率分布である。レコード値は記録値ともいい、たとえば、スポーツや気候の日本記録、世界記録などのように、いくつかの数字のうち最大 (小) のものである。上の  $\rho$  の値から、最大値と最小値は正に弱く相関していることがわかる。

### 5.3 2次元正規分布

標準正規分布  $N(0, 1)$  に従う独立な二つの確率変数  $X_1, X_2$  があるとき、 $a, b, c, d$  を定数 (ただし  $ad - bc \neq 0$ ) として、確率変数の変換

$$Y_1 = aX_1 + bX_2, \quad Y_2 = cX_1 + dX_2$$

を行ったとき、

- (a)  $Y_1, Y_2$  の平均  $\mu_1, \mu_2$
- (b)  $Y_1, Y_2$  の分散  $\sigma_1^2, \sigma_2^2$ , 共分散  $\sigma_{12}$ , 相関係数  $\rho$
- (c)  $Y_1, Y_2$  の同時確率密度関数  $p(y_1, y_2)$

を求めてみよう。

(a) 平均は  $E(X_1) = E(X_2) = 0$  から線形性により  $E(Y_1) = E(Y_2) = 0$  となる。よって  $\mu_1 = \mu_2 = 0$ 。なお、 $V(X_1) = E(X_1^2), V(X_2) = E(X_2^2)$  に注意しておこう。

(b) 独立性から、分散の加法性により

$$V(aX_1 + bX_2) = V(aX_1) + V(bX_2) = a^2V(X_1) + b^2V(X_2) = a^2 + b^2$$

さらに独立性から、 $E(X_1X_2) = E(X_1)E(X_2) = 0$  だから、

$$E\{(aX_1 + bX_2)(cX_1 + dX_2)\} = E(acX_1^2 + bdX_2^2) = acE(X_1^2) + bdE(X_2^2) = ac + bd$$

つまり、分散、共分散、相関係数は

$$\sigma_1^2 = a^2 + b^2, \quad \sigma_2^2 = c^2 + d^2, \quad \sigma_{12} = ac + bd, \quad \rho = \sigma_{12} / \sigma_1 \sigma_2$$

(c)  $Y_1, Y_2$  の同時確率密度関数については、まず  $X_1, X_2$  はそれぞれ  $N(0, 1), N(0, 1)$  に従うから、それらの同時確率分布は密度関数の積となって

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\{-(x_1^2 + x_2^2)/2\}$$

である。 $Y_1, Y_2$  の同時確率密度関数  $p(y_1, y_2)$  については変換

$$\phi : y_1 = ax_1 + bx_2, \quad y_2 = cx_1 + dx_2$$

を逆に解いて、逆変換

$$\psi : x_1 = a'y_1 + b'y_2, \quad x_2 = c'y_1 + d'y_2$$

(ただし、 $a' = d/D, \quad b' = -b/D, \quad c' = -c/D, \quad d' = a/D$  ( $D = ad - bc$ ) である.)

を、 $f(x_1, x_2)$  に代入すればよい。これで  $p(y_1, y_2)$  の主要部分が得られる。つまり、

$$f(x_1, x_2) = f(a'y_1 + b'y_2, c'y_1 + d'y_2)$$

密度関数は面積当たりのものであるから、変換  $\psi$  による面積の伸縮率をも考えるべきである。4点  $(y_1, y_2) = (0, 0), (0, 1), (1, 0), (1, 1)$  のつくる面積 1 の単位正方形が  $(x_1, x_2)$  の空間に移された平行四辺形で、どれだけの面積をもつかを調べれば、平面の幾何学から、簡単に  $|a'd' - b'c'|$  とわかる。実際  $|a'd' - b'c'| = 1/|D|$  である。

したがって、同時確率密度関数  $p(y_1, y_2)$  は

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(y_1, y_2) p(y_1, y_2) dy_1 dy_2 &= E[G(Y_1, Y_2)] \\ &= E[G(ax_1 + bx_2, cx_1 + dx_2)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(ax_1 + bx_2, cx_1 + dx_2) f(x_1, x_2) dx_1 dx_2 \\ &= \frac{1}{|D|} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(y_1, y_2) f(a'y_1 + b'y_2, c'y_1 + d'y_2) dy_1 dy_2 \end{aligned}$$

となる。したがって

$$p(y_1, y_2) = f(a'y_1 + b'y_2, c'y_1 + d'y_2) \cdot (1/|D|)$$

で求められる。この  $f$  の中を計算すると、式を考えて

$$(a'y_1 + b'y_2)^2 + (c'y_1 + d'y_2)^2 = Ay_1^2 + Cy_1y_2 + By_2^2$$

ただし

$$\begin{aligned} A &= \frac{c^2 + d^2}{D^2} = \frac{1}{\sigma_1^2(1 - \rho^2)}, \\ B &= \frac{a^2 + b^2}{D^2} = \frac{1}{\sigma_2^2(1 - \rho^2)}, \\ C &= -\frac{2(ac + bd)}{D} = -\frac{2\rho}{\sigma_1\sigma_2(1 - \rho^2)} \end{aligned}$$

である。ここで、恒等式  $(ac + bd)^2 + (ad - bc)^2 = (a^2 + b^2)(c^2 + d^2)$  から

$$D^2 = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1 - \rho^2)$$

となることを用いた。

結果として、 $(Y_1, Y_2)$  の同時確率密度関数は

$$p(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)} \left(\frac{y_1^2}{\sigma_1^2} - \frac{2\rho y_1 y_2}{\sigma_1\sigma_2} + \frac{y_2^2}{\sigma_2^2}\right)\right\}$$

となる。これを平均をそれぞれ  $0, 0$ , 分散をそれぞれ  $\sigma_1^2, \sigma_2^2$ , 共分散を  $\sigma_{12}$  とする 2 次元 (2 変量) 正規分布 (bivariate normal distribution) といい,  $N((0, 0), (\sigma_1^2, \sigma_2^2, \sigma_{12}))$  と表す。なお, 一般に平均が  $\mu_1, \mu_2$  であるときは  $y_1$  が  $y_1 - \mu_1$  に  $y_2$  が  $y_2 - \mu_2$  になる。これを平均  $\mu_1, \mu_2$ , 分散  $\sigma_1^2, \sigma_2^2$ , 共分散  $\sigma_{12}$  の 2 次元正規分布といい,  $N((\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2, \sigma_{12}))$  で表す。なお,  $(y_1, y_2)$  の指数部分は, 楕円の式であることに注意したい。

さらに一般の (次元が 3 以上である) 多次元 (多変量) 正規分布 (multivariate normal distribution) も定義することができる。

### 演習

- 【1】(独立と無相関) 二つのつぼ A, B の中に 3 個のボールを投げ入れる。つぼ A の中に入ったボールの数を  $X$ , ボールが入っているつぼの数を  $Y$  とするとき,  $X, Y$  の同時確率分布を求めて,  $X$  と  $Y$  は無相関であるが, 独立ではないことを示せ。(  $X$  の分布はパラメータ  $(n, p) = (3, \frac{1}{2})$  の 2 項分布であることに注意。)
- 【2】(相関係数の線形不変性)  $U = aX + b, V = cY + d$  (ただし  $ac > 0$ ) のとき,  $\rho_{UV} = \rho_{XY}$  が成り立つことを証明せよ。
- 【3】(2 次元正規確率変数の生成)  $X, Y$  は独立で, とともに標準正規分布  $N(0, 1)$  に従う確率変数とする。
- 定数  $c$  を適当に選んで  $X, cX + Y$  の相関係数が 0.5 となるようにせよ。
  - 同じく, 一般に  $\rho$  となるようにせよ。
  - $X, Y$  から, 与えられた 2 次元正規分布  $N((0, 0), (\sigma_1^2, \sigma_2^2, \rho))$  に従う 2 次元確率変数  $U, V$  を作れ。

## 6 大数の法則と中心極限定理

### 6.1 大数の法則

真の値への集中 公正なコインを 10 回投げることを考えてみる. このように, 1 回の実験で 2 種類の結果 (この場合, 表か裏) のいずれかが生じ, しかもそのような事象が生起する確率が常に一定 (この場合 1/2) であるような試行をベルヌーイ試行と呼ぶ. 「成功」を表とし, 表が出た回数の割合について考えてみよう.  $i$  回目のコイン投げで表が出た場合 1, 裏が出た場合 0 をとる確率変数  $X_i$  を考える. 10 回のコイン投げで, 表の出た回数 (頻度) は, 和

$$S_{10}(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_{10}(\omega)$$

である. 表が出た回数の割合  $\hat{p} = S_{10}(\omega)/10$  は観測された成功率であって,  $\hat{p} = 0, 0.1, 0.2, \dots$  となる. 一般に,  $n$  をコイン投げの回数とすると,  $S_n(\omega)/n$  は相対頻度である.  $S_n$  は確率変数で,  $n = 10, p = 0.5$  の 2 項分布  $B_i(10, 0.5)$ , すなわち

$$f_{10}(x) = \binom{10}{x} (1/2)^{10}, \quad x = 0, 1, 2, \dots, 10$$

に従い, その期待値, 分散は  $E(S_n) = np = 5$ ,  $V(S_n) = np(1-p) = 2.5$  だから, 割合  $S_n/n$  の期待値, 分散は

$$E(S_n/n) = p = 0.5, \quad V(S_n/n) = p(1-p)/n = 0.025$$

である. ここで  $p = 0.5$  は真の成功率となっている. 成功の割合が  $x/10$  となる割合は  $f_{10}(x)$  で, 実際に計算すると次のようになる.

観測された 成功率 $x/10$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
確率 $f_{10}(x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

期待値である真の成功率 0.5 およびその周辺の発生確率が高いが, 表の割合が 0.2 以下, および 0.8 以上であるような確率もまた 11.0% 近くある.

ここで, コイン投げの回数を  $n = 10$  から増やして,  $S/n$  の期待値  $E(S/n) = p$ , およびその周辺が発生する確率がどのように変わっていくかを調べてみよう. 期待値  $p = 0.5$  の周辺としては,  $0.5 \pm 0.1$  の範囲をとり, 0.4 から 0.6 までとする. コイン投げの回数は, 10 回から 20 回, 30 回, 40 回, 50 回, 100 回と増やしていくと, この確率は

$$\begin{aligned} P(0.4 \leq S_{10}/10 \leq 0.6) &= \sum_{x=4}^6 f_{10}(x) = 0.65625 \\ P(0.4 \leq S_{20}/20 \leq 0.6) &= \sum_{x=8}^{12} f_{20}(x) = 0.73682 \\ P(0.4 \leq S_{30}/30 \leq 0.6) &= \sum_{x=12}^{18} f_{30}(x) = 0.79951 \\ P(0.4 \leq S_{40}/40 \leq 0.6) &= \sum_{x=16}^{24} f_{40}(x) = 0.84614 \\ P(0.4 \leq S_{50}/50 \leq 0.6) &= \sum_{x=20}^{30} f_{50}(x) = 0.88108 \\ P(0.4 \leq S_{100}/100 \leq 0.6) &= \sum_{x=40}^{60} f_{100}(x) = 0.96780 \end{aligned}$$

などとなる. これから,  $n$  を増やしていくと確率は上がり,  $n = 100$  では, 表の観測された成功率  $\hat{p} = S_n/n$  が 0.4 から 0.6 までの確率は 96% を超え, ほとんどの値が真の成功率  $p = 0.5$  の周囲に集中する.

実際, 上の結果を式の形で表現すると

$$P(|S_n/n - 0.5| \leq 0.1) \rightarrow 1 \quad (n \rightarrow \infty)$$

ということである。一般に、 $\varepsilon$  がどのように小さい (正の) 数であっても

$$P(|S_n/n - 0.5| \leq \varepsilon) \rightarrow 1 \quad (n \rightarrow \infty)$$

となることが保証されるが、これが大数の法則 (law of large numbers) と呼ばれるものの一つの形である。

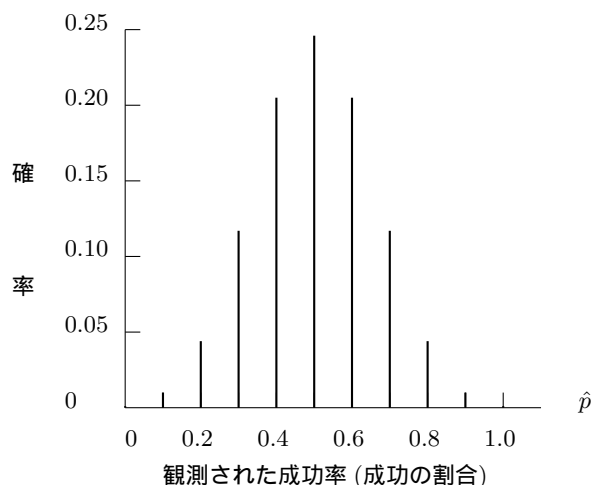


図 6.1 真の成功率  $p = 0.5$  のときの、観測された成功率  $\hat{p} = S_n/n$  ( $n = 10$  の場合)

コインを投げたときの理論上の成功率 (真の成功率) は  $p = 0.5$  であるが、実際に投げて観測したときの成功率 (成功の割合) は、正しく  $\hat{p} = 0.5$  ではない。ただし、期待値だからそうなる確率は比較的大きい。実際、 $n = 10$  ならば、 $0.5$  から外れる確率も小さくはない。 $n$  が大きくなると、事実上  $\hat{p} \doteq 0.5$  だけとなってゆくことを主張するのが、大数の法則である。なお、この分布の形が微妙に正規分布に似ていることは、中心極限定理を暗示している。

#### 統計学上の意義

ベルヌーイ試行における成功の回数が二項分布に従うことや、試行回数  $n$  を増やしていったとき、現実に観測された出現率 (標本での出現率) が、もとの集団 (後に「母集団」といわれる) における出現率 (母出現率)  $p$  に近づくこと、すなわち「大数の法則」を示したのは、ヤコブ・ベルヌーイ (Jacobus Bernoulli, 1654-1705) であったが、この大数の法則は統計学の歴史上、画期的な意味を持っていた。つまり、大数の法則は、十分な大きさの標本を調べれば、母集団の様々な特性をかなり正確に知ることができるという認識につながり、統計的推測の理論を生み出すことになった。

大数の法則は、一般的に、大標本では、観察された標本平均を母集団の真の平均 (母平均) とみなしてよいという常識を、数学的に厳密に証明したものに他ならない。応用例は現実に数多く見られる。

二項分布からもう少し一般化すると、 $\varepsilon$  を任意の正の定数、もとの確率分布の平均を  $\mu$ 、その分布から  $n$  個とられた観測値の平均を  $\bar{X}_n$  としたとき、

$$P(|\bar{X}_n - \mu| < \varepsilon) \rightarrow 1 \quad (n \rightarrow \infty)$$

となるといいかえられる。

## 6.2 中心極限定理

和の正規性 確率論の大定理である大数の法則は、統計学に応用されるものとしては、標本の大きさ  $n$  が十分大ならば、標本平均  $\bar{X} = (X_1 + \dots + X_n)/n$  の確率分布は母集団確率分布の平均 (母平均)  $\mu$  の近くに集中していることを保証している。例は、社会調査法であった。

いま一つの中心極限定理 (central limit theorem) は、大数の法則よりくわしい大定理であり、ごく大まかにいえば、母集団分布が何であっても、和  $X_1 + \dots + X_n$  の確率分布の形は、 $n$  が大なるときには、大略正規分布と考えるとよいということである。図 6.2-図 6.5 で見るように、母集団分布の平均、分散 (母平均、母分散) を  $\mu, \sigma^2$  とすると、母集団分布が何であっても、標本の大きさ  $n$  が大なるときは、大略

$$\begin{aligned} S_n = X_1 + X_2 + \dots + X_n & \text{ は } N(n\mu, n\sigma^2) \text{ に,} \\ \bar{X} = (X_1 + X_2 + \dots + X_n)/n & \text{ は } N(\mu, \sigma^2/n) \text{ に} \end{aligned}$$

従うと考えるとよい。

とくに  $\bar{X}$  については集中を保証する大数の法則よりくわしい。正規分布の形をとりながら集中する ( $\sigma^2/n \rightarrow 0$ ) のことを示しているからである。

母集団分布に正規分布を仮定せず、それが何であっても、和  $X_1 + X_2 + \dots + X_n$  が正規分布に従うという事実は、ある意味で驚くべき結果である。

中心極限定理を一応厳密に表すと、 $n \rightarrow \infty$  のとき

$$P(a \leq (X_1 + X_2 + \dots + X_n - n\mu)/\sqrt{n}\sigma \leq b) \rightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

が成り立つということである。いいかえれば  $n$  が大きければ

$$P(a \leq (X_1 + X_2 + \dots + X_n - n\mu)/\sqrt{n}\sigma \leq b) \doteq \Phi_{01}(b) - \Phi_{01}(a)$$

としてよい。ここで  $\Phi_{01}$  は標準正規分布の累積分布関数である。つまり

$$\Phi_{01}(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

である。なお、左辺は標準化関数の形で

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right) \doteq \Phi_{01}(b) - \Phi_{01}(a)$$

の形にしておいてもよい。

さいころを例に中心極限定理を実感として把握してみよう。さいころの目の出方の確率分布は離散型の一様分布であり、 $\mu = 7/2, \sigma^2 = 35/12$  である。この確率分布は、正規分布とは相当に違っている。この母集団からランダムに  $n = 2$  の標本  $X_1, X_2$  を取り出したときの標本平均  $(X_1 + X_2)/2$  とは、つまり、さいころを 2 回振ったときに出る目の平均値である。さいころを 2 回振ったときの目の出方は  $6 \times 6 = 36$  通りであるが、平均値にはいくつか同じものが出てくるので、整理すると次のようになる。

平均値	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
確率	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

これをグラフにしたものが図 6.2 である。1 回のときは峰 (モード) がないが、2 回のときは、それができており、中心極限定理の様子が既にあらわれている。中心極限定理は  $n$  を大きくすれば、和や標本平均の分布は正規分布に近づいていくというものだった。和の回数  $n$  はこの場合にはさいころを振る回数に等しいから、これを  $n = 2$  から 3, 4, 5 と増やしていったとき、出る目の和や標本平均の分布が正規分布に近づいていくことが、中心極限定理の述べている内容である。

さいころを 2 回振ったときと同様の方法で、3, 4, 5 回振ったとき出る目の平均値それぞれについて確率を論理的に計算し、図 6.2 と同様にグラフにしたものが図 6.3-図 6.5 である、明らかに分布は釣鐘型の正規分布に近づいていくことがみてとれよう。

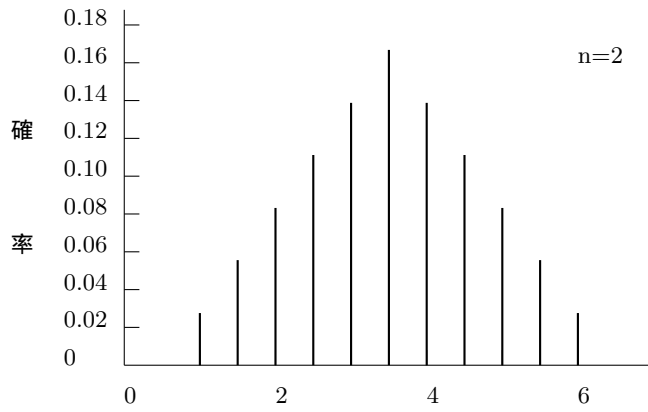


図 6.2

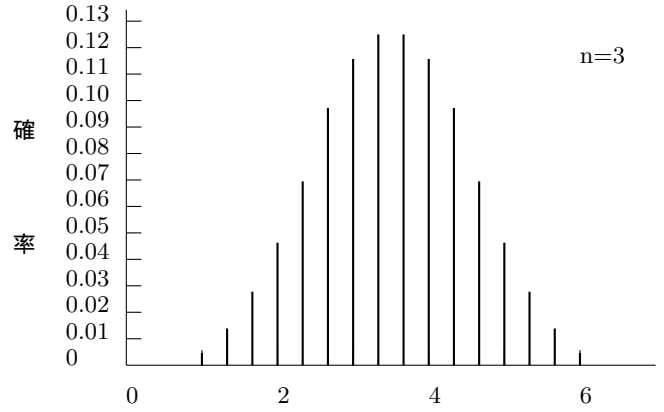


図 6.3

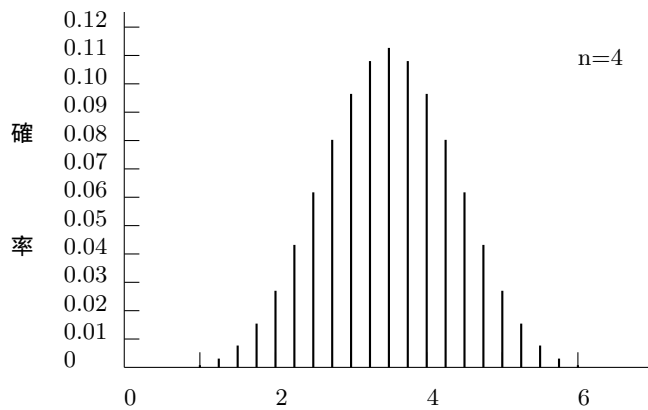


図 6.4

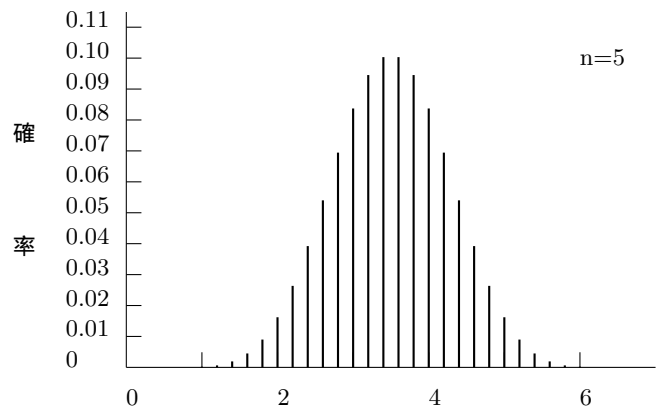


図 6.5

中心極限定理は確率分布の収束 (確率論では法則収束という) についてであるから, その証明にはモーメント母関数を用いる. まず  $V(X_1 + X_2 + \dots + X_n) = n\sigma^2 \propto n$  であって, このままでは確率分布の存在範囲が左右の限界を超えてしまうから, 標準化変数

$$(X_1 + X_2 + \dots + X_n - n\mu)/\sqrt{n}\sigma = (Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}$$

の確率分布が,  $n$  が大きいとき正規分布に近づくことを示そう. ここで

$$Y_1 = (X_1 - \mu)/\sigma, \quad Y_2 = (X_2 - \mu)/\sigma, \quad \dots, \quad Y_n = (X_n - \mu)/\sigma$$

は, おのおの,  $X_1, X_2, \dots, X_n$  の標準化変数であって,

$$E(Y_1) = E(Y_2) = \dots = E(Y_n) = 0, \quad V(Y_1) = V(Y_2) = \dots = V(Y_n) = 1$$

となっている.  $X_1, X_2, \dots, X_n$  の確率分布はみな同一, よって  $Y_1, Y_2, \dots, Y_n$  についてもそうであるから, その一つを  $Y$  とおいてモーメント母関数を作る. 1 次のモーメント (期待値) が  $E(Y) = 0.2$ , 2 次のモーメントが  $E(Y^2) = V(Y) = 1$  に注意して 5.3 節でみたように

$$M_Y(t) = 1 + t^2/2 + \dots$$

$Y_1 + Y_2 + \dots + Y_n \equiv T$  のモーメント母関数は  $Y_1, Y_2, \dots, Y_n$  のそれらの積で

$$M_T(t) = \{M_Y(t)\}^n = (1 + t^2/2 + \dots)^n$$

最終的に  $(Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}$  のモーメント母関数は、 $t$  を  $t/\sqrt{n}$  でおきかえ

$$M_t(t/\sqrt{n}) = \{M_Y(t/\sqrt{n})\}^n = \{1 + t^2/(2n) + \dots\}^n \rightarrow \exp(t^2/2)$$

$\exp(t^2/2)$  は標準正規分布のモーメント母関数であるから、中心極限定理は証明された。

さいころのように日常親しんでいるものからも、中心極限定理を使うと正規分布が生じる。  $n = 4, 5$  程度でも、一見しただけでは見分けがつかないほど、正規分布に近くなる。さいころ自体を「母集団」、出た目を「標本」とすると、統計学の理論として理解しやすい。

### 6.3 中心極限定理の応用

二項分布の正規分布による近似

二項分布  $B_i(n, p)$  は

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots, n$$

で表される。したがってこの式を用いれば、いかなる  $n, p$  においても  $f(x)$  を求めることが原理的には可能である。実際には、二項係数  $\binom{n}{x}$  の値が大きくなるためにそれほど簡単ではない。

既に述べたように、中心極限定理によると、 $n$  が大きいときに  $B_i(n, p)$  は正規分布に近づくので、それで近似することができる。なぜなら、二項分布における成功の回数  $S_n$  は、それぞれが二項分布  $B_i(1, p)$  に従う確率変数  $X_1, X_2, \dots, X_n$  の和  $S_n = X_1 + X_2 + \dots + X_n$  となるからである。  $E(S) = np, V(S) = np(1-p)$  より、中心極限定理により標準化変数

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$$

の分布は  $n$  が大きければ、標準正規分布  $N(0, 1)$  によって近似することができる。  $n$  回の試行のうち、成功の回数が  $k$  回以上  $k'$  回以下である確率は、 $n$  が大きければ、標準正規分布の累積分布関数  $\Phi_{01}$  により

$$\begin{aligned} P(k \leq S \leq k') &= P\left(\frac{k - np}{\sqrt{np(1-p)}} \leq Z_n \leq \frac{k' - np}{\sqrt{np(1-p)}}\right) \\ &\doteq \Phi_{01}\left(\frac{k' - np}{\sqrt{np(1-p)}}\right) - \Phi_{01}\left(\frac{k - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

で求めることができる。

(例) 中心かのゆらぎ 40,000 回コインを投げて 20,400 回以上、あるいは 19,600 回以下表が出ることは、どの程度の確率であろうか? 20,000 回だけ表が出ることが予想され、差 400 回は 2% の誤差であるから、直感では、実際上ありふれたことと考えられる。理論上の結論もそうであろうか。各回の表、裏を  $X_i = 1, 0$  として、 $n$  回中の 1 の総回数  $X = X_1 + X_2 + \dots + X_n$  の確率分布を求めればよい、 $n = 40,000$  で、 $X$  は二項分布  $B_i(40,000, 1/2)$  に従うが、

$$\sum_{19600}^{20400} \binom{40000}{x} (1/2)^x (1/2)^{40000-x}$$

を計算することは不可能である。そこで、中心極限定理をつかう。  $X_i$  は二項分布  $B_i(1, 1/2)$  に従うから、  $\mu = E(X_i) = 1/2, \sigma^2 = V(X_i) = (1/2) \cdot (1/2) = 1/4, n\mu = 20,000, \sqrt{n}\sigma = 100$ 、したがって

$$\begin{aligned} &P(19,600 \leq X_1 + X_2 + \dots + X_{40,000} \leq 20,400) \\ &= P(-4 \leq (X_1 + X_2 + \dots + X_{40,000} - 20,000)/100 \leq 4) \\ &= \Phi_{01}(4) - \Phi_{01}(-4) \\ &= 0.9999 \end{aligned}$$

よって、求める確率は  $1/10,000$  程度であり、事実上ありえないこととなる。

さて、二項分布で  $n$  の値がどのくらい大きければ、正規分布による近似を用いてよいか。実用上十分な精度を得るために通常言われている必要条件は、 $np > 5$  かつ  $n(1-p) > 5$  である。したがって、 $p$  が  $1/2$  のときには  $n$  は  $10$  以上であればよいが、 $p$  が  $0$  や  $1$  に近いときには、 $n$  が相当大きくなければならない。