Hand out

確率論 (*Probability*) 第 4 週

due	1 mai 2014
cur	1 mai 2014
ver	1
rev	0

w.yamamoto

1 標本空間の表現の一般化

多くの事象は、整数もしくは実数のスカラーもしくはベクトルで記述できる。以下では真偽の二値論理に関して、真を 1、偽を 0 とする。また $I(\cdot)$ は識別関数といい、中身が真の時に 1 を、偽の時に 0 を返す関数と定義する。

コイン投げ ($X \in \{\text{表}, \text{裏}\}$) のような 2 状態しかない現象は

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} I(X = \text{表}) \\ I(X = \text{裏}) \end{pmatrix} \in \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \tag{1}$$

と 2 次元のベクトルで表現できる。この集合は天気 ($X \in \{\text{晴}, \text{曇}, \text{雨}\}$) のような 3 状態の現象は

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} I(X = \text{晴}) \\ I(X = \text{曇}) \\ I(X = \text{雨}) \end{pmatrix} \in \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} \tag{2}$$

と 3 次元のベクトルで表現できる。多状態の現象も同様に、状態数の次元の整数ベクトルで表現できる。実は、それぞれ「表でなければ裏」「晴れでも曇りでもなければ雨」という排反な関係を用いれば、

$$\tilde{X} = I(X = \text{表}) \in \{1, 0\} \tag{3}$$

あるいは

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} I(X = \text{晴}) \\ I(X = \text{曇}) \end{pmatrix} \in \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\} \tag{4}$$

とベクトルの次元を簡単に 1 つ減らすことができる。またこれらは情報の符号化に他ならず、可変長でよければもっと効率のよい符号化もあり得る。

どれかの事象が真の時に他のすべての事象は偽となるような排反な分類に基づく現象の捉え方を、名義尺度あるいは分類尺度、と言う。名義尺度や分類尺度をもつ確率変数には、上のような離散値を要素にもつベクトル表現が用いられる。

天気の間に「天気の良さ」という意味を加わると、晴が一番良い天気、曇が次に良い天気、雨は悪い天気、という順序が定まる。このように分類の間に順序が定まる時、この捉え方を順序尺度という。満足度アンケートにおける「とてもおいしい」「まあまあおいしい」「どちらとも言えない」「あまりおいしいとは思わない」「ぜんぜんおいしくない」などの段階評価も、順序尺度の例となる。ただしこの講義では、順序尺度は紹介に留める。このような順序のある分類を $\{5, 4, 3, 2, 1\}$ の 5 点評価に換算することもあるが、これも順序尺度のままであり、引き算に意味はない。

さいころ投げは、出る目が $\{1, 2, 3, 4, 5, 6\}$ の 6 種類の数値の現象である。このときはそのまま

$$X = k \in \{1, 2, 3, 4, 5, 6\} \tag{5}$$

と離散集合で表す。これは数値に意味はあるが、飛び飛びの値しか取れないので、離散尺度という。値の差に意味があるので、間隔尺度とも言われる。成功回数、来客数など、特定の事象の発生回数なども非負の整数全体の離散集合で表せる。

身長、体重、気温、気圧、湿度、あるいは実験の測定結果など、連続量はそのまま

$$X \in (-\infty, \infty) = \Re \tag{6}$$

と 1 次元の数直線上のスカラー量で表現する。体重を、kg を単位に測定する場合、その取りうる範囲は $(0, 500)$ 程度で十分なはずだが

$$(0, 500) \subset \Re \tag{7}$$

より、数直線全体で表すことが多い。これを連続尺度と言う。
連続尺度であれ離散尺度であれ、複数の現象の組み合わせは

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \in (-\infty, \infty) \otimes (-\infty, \infty) \otimes \cdots \otimes (-\infty, \infty) = \mathcal{R}^p \tag{8}$$

とやはり p 次元のユークリッド空間内の点として表現できる。

2 集合

個々の問題の標本空間は典型的な集合
自然数の集合

$$\mathcal{N} = \{1, 2, 3, \ldots\} \tag{9}$$

非負の整数の集合

$$\mathcal{Z}^+ = \mathcal{N} \cup \{0\} \tag{10}$$

整数の集合

$$\mathcal{Z} = \mathcal{N} \cup \{0\} \cup \{-x; x \in \mathcal{N}\} \tag{11}$$

実数の集合、数直線、1 次元ユークリッド空間

$$\mathcal{R} = \{x; -\infty < x < \infty\} \tag{12}$$

非負の実数の集合

$$\mathcal{R}^+ = \{x; 0 \leq x < \infty\} \tag{13}$$

次元を上げるときは、空間の直積を用いる。
p 次元ユークリッド空間

$$\mathcal{R}^p = \mathcal{R} \otimes \mathcal{R} \otimes \cdots \otimes \mathcal{R} = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; x_1, \ldots, x_p \in \mathcal{R} \right\} \tag{14}$$

p 次元整数ベクトル空間

$$\mathcal{Z}^p = \mathcal{Z} \otimes \mathcal{Z} \otimes \cdots \otimes \mathcal{Z} = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; x_1, \ldots, x_p \in \mathcal{Z} \right\} \tag{15}$$

p 次元空間のシンプレックス

$$\mathcal{S}^p(k) = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; x_1 + \cdots + x_p = k, x_1, \ldots, x_p \in \mathcal{R}^+ \right\} \tag{16}$$

p 次元整数ベクトル空間のシンプレックス

$$\mathcal{D}^p(k) = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; x_1 + \cdots + x_p = k, x_1, \ldots, x_p \in \mathcal{Z}^+ \right\} \tag{17}$$

3　加法族の一般化

上のように標本空間を数直線あるいはユークリッド空間に統一しようとしたとき、点ではなく区間が確率を定義するための単位事象となる。

$$Pr[X \in [150, 180]] \tag{18}$$

は定義できるが、

$$Pr[X = 174.32353321\ldots] \tag{19}$$

は定義できない。なぜかといえば、「実数の個数」はたとえ区間を $[0, 1] = \{0 \leq X \leq 1\}$ に限定しても、非可算無限の個数だけ存在する。

例えば個々の実数を仮に数え上げられたとして、それぞれに限りなく 0 に近く微小だが非負の確率を付与したとしよう。それらを $p_1, p_2, \ldots, p_n \ldots$ と表す。それらの中の最小値を ϵ と置くと

$$\sum_{i=1}^{\infty} p_i \leq \sum_{i=1}^{\infty} \epsilon = \epsilon \times \infty = \infty \tag{20}$$

より、確率の和が ∞ になってしまう。これは確率の公理のうちの、全確率は 1 という公理に反する。実際には、実数の個数はここで記した ∞ の意味よりももっと多いので、個々の実数に確率を付与することはできない。

実数に関する確率について、点に確率を定義できないなら、と登場するのが区間である。$a < b$ を満たす任意の 2 実数 $a, b \in \Re$ について

$$Pr[a < X < b] = Pr[X \in (a, b)] \tag{21}$$

$$Pr[a \leq X < b] = Pr[X \in [a, b)] \tag{22}$$

$$Pr[a < X \leq b] = Pr[X \in (a, b]] \tag{23}$$

$$Pr[a \leq X \leq b] = Pr[X \in [a, b]] \tag{24}$$

などの区間は、1 次元のスカラー量 X で表せる確率現象の確率を定義するための基礎となる。$(a, b]$ を基本とすることも多い。このような区間自身はもう、互いに排反ではないが、例えば体重を 0.1kg 刻みで測定するなら、

$$\ldots, (40.1, 40.2], (40.2, 40.3], \ldots \tag{25}$$

などの区間は互いに排反となる。

本講義では ω と写像と可測集合と可測性については、言及を避けておく。ただしこれらは確率論を理解する上ではとても重要な概念であるので、興味を持った学生はぜひ参考書を紐解いて、2.1 節を理解することを勧める。

4　累積分布関数 (離散と連続の場合)

$\mathcal{X}$ が要素に自然な順序がある集合とする。そして $\mathcal{X}$ の部分集合のうち、確率を付与したい部分集合、およびそれらの和集合、そしてそれらの補集合、などを集めて作った「部分集合のリスト」をここでも加法族と呼ぶ。

1 次元の問題に対して、区間 $(a, b]$ の確率は

$$Pr[a < X \leq b] = Pr[-\infty < X \leq b] - Pr[-\infty < X \leq a] \tag{26}$$

と表せる。もし $a_1 < b_1 < a_2 < b_2$ として、$(a_1, b_1] \cap (a_2, b_2]$ の確率も同様に

$$Pr[a_1 < X \leq b_1 \text{ または } a_2 < X \leq b_2] = Pr[-\infty < X \leq b_1] - Pr[-\infty < X \leq a_1] + Pr[-\infty < X \leq b_2] - Pr[-\infty < X \leq a_2] \tag{27}$$

と表せる。このように、

$$F_X(x) = Pr[-\infty < X \leq x] = Pr[X \in (-\infty, x]) \tag{28}$$

だけを関数として定めておけば、加法族に関する確率はすべてこの関数で

$$Pr\left[a < X \leq b\right] = F\left(b\right) - F\left(a\right) \tag{29}$$

$$Pr\left[a_1 < X \leq b_1 \text{ または } a_2 < X \leq b_2\right] = F\left(b_1\right) - F\left(a_1\right) + F\left(b_2\right) - F\left(a_2\right) \tag{30}$$

と表現できる。この関数を累積分布関数もしくは分布関数という。省略形は cdf である。

標本空間が離散の場合にも、上で定義したユークリッド 空間内の離散点の集合のため、同様に定義できる。

累積分布関数は次の 3 つの性質を満たす。

1. 非減少性 $x_1 \leq x_2 \Rightarrow F\left(x_1\right) \leq F\left(x_2\right)$
2. 右連続性 $\lim_h \to +0 F\left(x + h\right) = F\left(x\right)$
3. 有界性 $\lim_{x \to -\infty} F\left(x\right) = 0, \lim_{x \to +\infty} F\left(x\right) = 1$

5 確率関数 (離散の場合)

6 確率密度関数 (連続の場合)