Theoretical Statistics

伴 尚哉 *
61715382
慶應義塾大学理工学部数理科学科 白石研究室

2020 年 4 月 15 日

1 Probability and Measure

多くの統計的理論は、確率・測度論をよく理解していなくとも理解できるであろう。この本も、確率・測度論を細かいところまで扱ってはいない。しかし、先の分野のいくつかの基礎知識はとても役に立つのである。統計学に出てくる文献の多くに測度論は使われていて、基礎知識のない人々には利用しにくいのだ。また、測度論の表記法のおかげで、我々は離散的な確率変数と連続的な確率変数を同じ事項として扱えるのだ。加えて、離散的でも連続的でもないといった興味深い確率変数の動きを追うのに必要なことも、測度論の表記法のおかげで我々は扱えるのである。つまり、測度論は、多くのことを正しく表記するのに必要なのである。結局、本書は、証明の中で測度論の細かな部分は省かれているものの、確率論に博識な人が行間を埋めるくらいには丁寧に書かれてはいるのだ。

この章では、確率・測度論を紹介していて、また、いくつかのよく使われる事実については証明なしで記載されている。

1.1 測度論

集合 $\mathcal{X}$ 上の測度 μ とは、$\mathcal{X}$ に含まれる部分集合 A に対しての非負値 $\mu(A)$ のことである。以下に 2 つ例を挙げる。

例 1.1 . $\mathcal{X}$ が可算集合と仮定する。このとき、$\mathcal{X}$ 上の数え上げ測度を以下で定義することができる。

$$\mu(A) = \#A = A\text{の元の個数} \tag{1}$$

*naoyayg30@keio.jp

例 1.2 . $\mathcal{X} = \mathbb{R}^n$ とする。$\mu(A)$ を

$$\mu(A) = \int_{\mathrm{A}} \cdots \int dx_1 \cdots dx_n. \tag{2}$$

と定めたとき、n=1,2,3 の下では $\mu(A)$ は各々長さ、面積、体積と呼ばれる。また一般的に、上で定義した μ は $\mathbb{R}^n$ 上でのルベーグ測度と呼ばれる。実際、いくつかの集合では、式 (2) で定義した $\mu(A)$ の値を求める方法が明らかではない。そこで以下我々が示すように、測度論の定理が積分学の根本的な問題に根強くかかわってくるのである。

先ほどの例で挙げられた測度は、ある意味それぞれ異なっている。数え上げ測度は、個々の点の個数を測ってた。つまり、$\mu(\{x\}) = 1\ (\forall x \in \mathcal{X})$ である。一方、ルベーグ測度では、$\mu(\{x\}) = 1$ となるような点はない。一般的に、もし $\mu(\{x\}) > 0$ ならば、x は $\mu(\{x\}) > 0$ な元と呼ばれる。

普通、$\mathcal{X}$ のすべての部分集合 A にたいして、測度を割り当てるのは不可能である。その代わり、測度 μ に対する領域 -"$\sigma -$ 加法族" を用意する。

定義 1.3. A を部分集合、$\mathcal{A}$ を集合 $\mathcal{X}$ の集合族と仮定する。以下が成り立つとき、$\mathcal{A}$ を $\sigma -$ 加法族 と呼ぶ。

(a) $\mathcal{X} \in \mathcal{A}\, and\, \phi \in \mathcal{A}.$

(b) $A \in \mathcal{A} \Rightarrow \mathcal{A}^c = \mathcal{X} - A \in \mathcal{A}$

(c) $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

以下で定義することは、集合関数 μ が測度と呼ばれるためには必ず満たさなければならない基本的項目である。例 1.1,1.2 でもこれらの項目は直観的に満たすことがわかる。

定義 1.4. 以下を満たすとき、$\sigma -$ 加法族 $\mathcal{A}$ 上の関数 μ が測度であるという。

(a) $\forall A \in \mathcal{A}, 0 \le \mu(A) \le \infty$

(b) $A_1, A_2, \cdots \in \mathcal{A},\ A_i \cap A_j = \phi\ (\forall i \ne j)\ \Rightarrow\ \mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

ここで、上で定義したことについての興味深くまた有用な結果として以下のことが言える。

可測集合 $B_n\ (n \ge 1)$ が、単調増加且つ $B = \bigcup_{n=1}^{\infty} B_n$ を仮定する。

$$\mu(B) = \lim_{n \to \infty} \mu(B_n) \tag{3}$$

これは、測度が連続といった特徴だと捉えられる。

$\mathcal{A}$ が集合 $\mathcal{X}$ の $\sigma-$ 加法族 のとき、$(\mathcal{X}, \mathcal{A})$ を可測空間と呼ぶ。また、μ を $\mathcal{A}$ 上の測度としたとき、$(\mathcal{X}, \mathcal{A}, \mu)$ を測度空間と呼ぶ。

また、測度 μ が有限であるとは $\mu(\mathcal{X}) < \infty$ となる時であり、測度 μ が σ 有限であるとは、$\mu(A_i) < \infty\,(\forall i \in \mathbb{N} \wedge \bigcup_{i=1}^{\infty} A_i = \mathcal{X})$ となる $A_1, A_2, \ldots$ が $\mathcal{A}$ に含まれている時である。本書では、測度はすべて σ 有限であると仮定する。

また、測度 μ が確率測度であるとは $\mu(\mathcal{X}) = 1$ を満たすときであり、その時は測度空間のことを確率空間と呼ぶ。確率測度や有限測度では、減少列に対して式 (3) で類推したような事が成立する。

可測集合 $B_i\ (i \in \mathbb{N})$ が単調減少で、$B = \bigcap_{i=1}^{\infty} B_i$ を満たしているとき、

$$\mu(B) = \lim_{n\to\infty} \mu(B_n)$$

となる。

例 (1.1) の時を考える。数え上げ測度は、$\mathcal{X}$ のどんな部分集合に対しても $\mu(A) = \#A$ と定義することができるる。よって、$\sigma-$ 加法族 $\mathcal{A}$ は、$\mathcal{X}$ のすべての部分集合を集めたものになる。この $\sigma-$ 加法族 は $\mathcal{X}$ のべき集合と呼ばれ、$\mathcal{A} = 2^{\mathcal{X}}$ と定義される。

例 (1.2) の時を考える。集合 A のルベーグ測度は、暗黙の了解で、R^n の Borel 集合と呼ばれる σ一加法族 $\mathcal{A}$ の元 A に対して定義される。正式には、$\mathcal{A}$ はすべての矩形を含む最小の $\sigma-$ 加法族 である。矩形の定義は以下である。

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n) = \{x \in \mathbb{R}^n \ :\ a_i < x_i < b_i,\ i = 1, 2, \cdots\}$$

しかし、ボレル集合族ではない集合族はたくさんあるものの一つとして明示的に定義できるものはない。

1.2 積分

この章の目標は、測度 μ に対して性質の良い関数 f の積分を適切に定義することである。積分は、必要に応じて $\int f d\mu$ *or* $\int f(x) d\mu(x)$ といった表記で書かれる。この後の発展のために、数え上げ測度とルベーグ測度についての積分を述べる。

例 1.5 μ が $\mathcal{X}$ 上の数え上げ測度とする。このとき μ に対する関数 f の積分は以下で定義される。

$$\int f d\mu = \sum_{x\in\mathcal{X}} f(x).$$

例 1.6 μ が $\mathbb{R}^n$ 上のルベーグ測度であるとする。このとき μ に対する関数 f の積分は以下で定義される。

$$\int f d\mu = \int \cdots \int f(x_1, \cdots, x_n) dx_1 \cdots dx_n.$$

x をベクトル $(x_1, x_2, \ldots, x_n)'$ と見て、ルベーグ測度に対する上記の積分を $\int \cdots \int f(x)dx \; or \; \int f(x)dx$ と書くのが便利である。

ここで挙げられている現代の積分の定義は、ほとんどの基本的な微積分の授業で与えられている定義だ。積分は、性質の良い関数 f は $\int f d\mu$ が一意的でなければならないという主張のもとに構成されている。積分学の規則のポイントとして、以下の定義で挙げられるような"可測である"というのがある。

定義 1.7 $(\mathcal{X}, \mathcal{A})$ が可測空間であるとし、$\mathcal{X}$ 上の関数 f が実数値のみを取るものとする。このとき、f が可測であるとは、すべてのボレル集合 B に対して、

$$f^{-1}(B) \stackrel{\text{def}}{=} \{x \in \mathcal{X} \; : \; f(x) \in B\} \in \mathcal{A}$$

が成り立つことである。

可測でない関数は多く存在するものの、それらを明確に表記することはできない。連続関数や、区分連続関数は可測である。もっと興味深い例として、$f : \mathbb{R} \to \mathbb{R}$ に対して、$f(x) = 1 \; (x \in (0,1)) \; and \; f(x) = 0 \; (x \notin (0,1))$ とすると、f は可測である。基礎の微積分学で使われるようなリーマン積分は、このような可測関数 f に対して $\int f(x)dx$ を定義できない。ここで紹介されるより一般的な方法では、$\int f(x)dx = 1$ といった整数値を返す。結局、興味のある関数というのは一般的に可測であることを想定されているのだ。

集合 A に対する指示関数 1_A とは以下で定義される関数のことである。

$$1_A(x) = I\{x \in A\} = \begin{cases} 1, \; x \in A; \\ 0, \; x \notin A. \end{cases}$$

以下に、積分の基本的な性質を述べる。

1. どんな $A \in \mathcal{A}$ についても、$\int 1_A d\mu = \mu(A)$ である。
2. もし、$f \cdot g$ 共に非負可測関数であるなら、正の定数 $a \cdot b$ にたいして、

$$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu. \tag{4}$$

が成り立つ。

3. $f_1 \leq f_2 \leq \cdots$ となる f_i が $\forall i \in \mathbf{N}$ で非負可測関数かつ、$f(x) = \lim_{n\to\infty} f_n(x)$ であるとすると

$$\int f d\mu = \lim_{n\to\infty} \int f_n(x) d\mu.$$

が成り立つ。

上記の特徴の一つ目は $\int f d\mu$ と測度 μ を連携させている。二つ目は、線形性についてを述べている。三つめは、極限を取るときに使えるものである。

最初の二つの特徴を使うと、$a_1, \ldots, a_m$ を正の定数として、$A_1, \ldots, A_m$ を $\mathcal{A}$ の集合としたとき以下が成り立つ。

$$\int \left(\sum_{i=1}^{m} a_i 1_{A_i} \right) d\mu = \sum_{i=1}^{m} a_i \mu(A_i)$$

が成り立つ。このような形の関数を単関数と呼ぶ。本書図 1.1 では、$1_{(1/2,\pi)} + 2 * 1_{(1,2)}$ と定義される単関数のグラフが示されている。以下の結果は非負可測関数なら、単関数で近似できることを表している。

定理 1.8 関数 f が非負可測であるとすると、$f_1 \leq f_2 \leq \ldots$ 且つ $f = \lim_{n\to\infty} f_n$ を満たす単関数 f_n が存在する。

極限に関わる積分の特徴と関係のあるこの定理のおかげで、原理上我々は多くの非負可測関数 f の積分をすることができるのだ。また、特徴的な結果として、上記を満たす単関数を別々に取ってきても、積分の結果は一致する。ここで、一般の可測関数についても積分するために関数 f の正部分と負部分を以下のように定義する。

$$f^+(x) = max\{f(x), 0\}\ f^-(x) = max\{-f(x), 0\} \tag{5}$$

このとき、f^+ と f^- はどちらも非負可測であり、$f = f^+ - f^-$ と書ける。f の積分は正部分の積分と負部分の積分、相違となる。この相違は、正部分と負部分の積分がどちらも正の無限に発散する時、あいまいなものとなる。なので、正部分の積分と負部分の積分のどちらかが発散しない時、f の積分を以下のように定義する。

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

$\int f d\mu = \infty$ とならない条件下で、上記の式は積分の線形性の式 (4) から出されている。また、$|f| = f^+ + f^-$ より、以上の定義によって $\int f d\mu$ が有限値となる同値条件は以下である。

$$\int f^+ d\mu + \int f^- d\mu = \int |f| d\mu < \infty$$

また、以上の条件を満たすとき、f は積分可能と呼ばれる。

1.3 事象・確率・確率変数

P を可測空間 $(\varepsilon, \mathcal{B})$ 上の確率測度とする。つまり、$(\varepsilon, \mathcal{B}, P)$ は確率空間となる。$\mathcal{B}$ の集合 B を事象と呼び、ε の元 e を結果と呼ぶ。そして、$P(B)$ を B の確率と呼ぶ。

可測関数 $X : \varepsilon \to \mathbb{R}$ を確率変数と呼ぶ。ボレル集合 A に対して、確率測度 P_X は以下のように定義され、X の分布と呼ばれる。。

$$P_X(A) = P(\{e \in \varepsilon : X(e) \in A\}) \stackrel{\text{def}}{=} P(X \in A)$$

また、

$$X \sim Q$$

と書けば、X が分布 Q を持つ、つまり $P_X = Q$ ということを表している。X の累積分布関数は $x \in \mathbb{R}$ に対して以下のように定義される。

$$F_X(x) = P(X \leq x) = P(\{e \in \varepsilon : X(e) \leq x\}) = P_X((-\infty, x])$$

1.4 零集合

μ を $(\mathcal{X}, \mathcal{A})$ 上の測度とする。N が零集合である定義は以下のとおりである。

$$\mu(N) = 0$$

もし、主張が $x \in \mathcal{X} - N$ $(\mu(N) = 0)$ で成立するなら、その主張はほとんどいたるところで (μ a.e. *or* (a.e.)) 成立するという。例えば、$f = 0$ (a.e. μ) というのは $\mu(\{x \in \mathcal{X} : f(x) \neq 0\}) = 0$ と同値である。

μ が確率測度の時、上記には別の表現がある。主張が $x \in B$ のみで成立すると仮定する。そのとき主張が (a.e. μ) で成立するということと、$\mu(B^c) = 0$ ということと $\mu(B) = 1$ は同値である。これが、「この主張は確率 1 で成立する」ということを表している。

零集合上の関数は積分に影響を与えない。この考えより、認めることの容易な、積分についての重要な事実を以下にいくつか述べる。

1. $f = 0$ (a.e. μ) $\to$ $\int t d\mu = 0$
2. $f \geq 0 \ \wedge \ \int f d\mu = 0 \ \to \ f = 0$ (a.e. μ)
3. $f = g$ (a.e. μ) $\to$ $\int f d\mu = \int g d\mu$(ただし、どちらかの関数の積分が存在するとする)
4. $\int 1_{(c,x)} f d\mu = 0 \ (\forall x > c) \ \to \ f(x) = 0$ (a.e. $x > c$)(ただし、定数 c は $-\infty$ もとれる)

上記二番目の結果より、もし f と g が可積分で $f > g$ ならば、$\mu \equiv 0$ でない限り $\int f d\mu > \int g d\mu$ となる。

1.5 確率密度

確率密度は、統計学の中でとても基本的な役回りである。多くの状況下で、確率変数 X を特定化するのに最も便利な方法が確率密度を与えることである。また、確率密度はベイズ推定量や最尤推定量を算出するに使われる尤度関数を与える。以下の定義で示される通り、測度に関する確率密度は測度がほかの測度に関して絶対連続であるとき存在する。

定義 1.9. P と μ を $\mathcal{X}$ の $\sigma-$ 加法族 $\mathcal{A}$ 上の測度とする。このとき、$\mu(A)=0$ ならば $P(A)=0$ となるなら $P \ll \mu$ と書かれ、 P は μ に関して絶対連続とよばれる。

定理 1.10 有限測度 P が σ 有限測度な μ に対して絶対連続であるとする。このとき非負可測関数 f が存在し、以下のようになる。これを Radon - Nykodym の定理という。

$$P(A)=\int_A f d\mu \stackrel{\text{def}}{=} \int f 1_A d\mu$$

この定理内の関数 f のことを P のμ に関する Radon - Nykodym 導関数、もしくは μ に関する P の確率密度と呼ばれ、そして以下のように書かれる。

$$f=\frac{dP}{d\mu}$$

前章の積分と零集合の三つ目の結果より、確率密度 f は一意ではないが、f_0, f_1 がどちらも確率密度であった場合 $f_0=f_1\ (a.e.\ \mu)$ となる。もし、$X \sim P_X$ で P_X が μ で絶対連続であり確率密度 $p=dP_X/d\mu$ であるとき、一般的に X は μ にたいして確率密度 p を持つと呼ばれる。

例 1.11 (絶対連続な確率変数) もし、確率変数 X が $\mathbb{R}$ 上のルベーグ測度に対して確率密度 p を持っていたとする。このとき X やその分布関数 P_X は、確率密度 p に対して絶対連続だと呼ばれる。Radon - Nykodym の定理より、

$$F_X(x)=P(X\leq x)=P_X((-\infty,x])=\int_{-\infty}^{x} p(u)d\mu$$

となる。微積分学の基本的な定理を使うと、p は累積分布関数 F_X から $p(x)=F_X^{'}(x)$ と導出できる。

例 1.12 （離散確率変数）$\mathcal{X}_0$ を $\mathbb{R}$ の可算な部分集合とする。このとき、測度 μ は以下で定義される。

$$\mu(B)=\#(\mathcal{X}_0 \cap B)$$

このとき、上記の測度はボレル集合 B に対する $\mathcal{X}_0$ 上の数え上げ測度と呼ばれる。例 1.5 で見られるように、

$$\int f d\mu = \sum_{x \in \mathcal{X}_0} f_(x)$$

となる。

もし X が確率変数で $P(X \in \mathcal{X}_0) = P_X(\mathcal{X}_0) = 1$ であると仮定する。このとき X は離散的確率変数と呼ばれる。もし N が μ の零集合であるとき $\mu(N) = 0$ である。測度 μ の定義から、$\#(N \cap \mathcal{X}_0) = 0$ である。つまり、$N \cap \mathcal{X}_0 = \phi$ であり、$N \subset \mathcal{X}_0^c$ となる。このとき、$P_X(N) = P(X \in N) \leq P(X \in \mathcal{X}_0^c) = 1 - P(X \in \mathcal{X}_0) = 0$ となる。つまり、P_X にたいしても N は零集合となる。そして、これは P_X は μ にたいして絶対連続であることを示している。測度 μ に対する P_X の確率密度 p は以下を満たす。

$$P(X \in A) = P_X(A) = \int_A p d\mu = \sum_{x \in \mathcal{X}_0} p(x) 1_A(x)$$

特に、$A = \{y\}$ $(y \in \mathcal{X}_0)$ のとき、$X \in A$ で $X = y$ であるのと同値であるため、$P(X = y) = \sum_{x \in \mathcal{X}_0} p(x) 1_{\{y\}}(x) = p(y)$ が成り立つ。このとき、確率密度 p は X の確率関数と呼ばれる。$\mathcal{X}_0^c$ が零集合であるため、確率密度 $p(y)$ は任意の $y \notin \mathcal{X}_0$ に対して定義できる。慣習では、$y \notin \mathcal{X}_0$ に対して $p(y) = 0$ と取るため、任意の y で $p(y) = P(Y = y)$ となるのだ。

1.6 期待値

今 X が $(\varepsilon, \mathcal{B}, P)$ 上の確率変数だとする。このとき、X の期待値を以下で定義する。

$$E[X] = \int X dP \tag{6}$$

この式はめったに使われない。代わりに、$X \sim P_X$ だとすると以下のようにあらわせる。

$$E[X] = \int x dP_X(x) \tag{7}$$

また、$Y = f(X)$ のとき、

$$E[Y] = E[f(X)] = \int f dP_X \tag{8}$$

と書ける。二つの式の P_X に関する積分は、よく確率密度を使って表せられる。P_X が測度 μ に対して確率密度 p を持つとする。このとき、

$$\int f dP_X = \int f p d\mu \tag{9}$$

と書ける。この特性のおかげで、形式的に $pd\mu$ を dP_X の代わりに用いることができ、導関数表記 $p = dP_X/d\mu$ とかけるのは至極当然のように思える。これらの事実は、変数変換と同じだと捉えられている。これらの結果の証明は、積分を定義したときの方法が基本となっている。式 (8・9) が指示関数を持つときは、証明は簡単である。線形性より、一般の可測関数の時でも、積分が存在すれば正の単関数や極限を使うと示すことができる。絶対連続や離散的確率変数用に専門化した興味深い例を以下に挙げる。

例 1. 13 今、X が絶対連続な確率変数で、確率密度 p を持つとする。このとき以下が成り立つ。

$$E[X] = \int x dP_X(x) = \int x p(x) dx$$
$$E[f(X)] = \int f(x) p(x) dx \tag{10}$$

例 1. 14 もし、X が離散で、可算集合 $\mathcal{X}_0$ に対して $P(X \in \mathcal{X}_0) = 1$ であるとする。また、μ は $\mathcal{X}_0$ 上の数え上げ測度とする。そして、確率関数は $p(x) = P(X = x)$ であるとする。このとき、以下が成り立つ。

$$E[X] = \int x dP_X(x) = \int x p(x) d\mu(x) = \sum_{x \in \mathcal{X}_0} x p(x)$$
$$E[f(X)] = \sum_{x \in \mathcal{X}_0} f(x) p(x) \tag{11}$$

期待値は線形性がある。今 X, Y が確率変数であるとし、a, b を 0 以外の非負定数とする。また、$E[X], E[Y]$ 共に存在し、$\infty - \infty$ とならないと仮定する。このとき

$$E[aX + bY] = aE[X] + bE[Y] \tag{12}$$

が成り立つ。これは積分の線形性の式 (4) より期待値の定義式 (6) から簡単に示される。期待値のもう一つの重要な特性として、X, Y が有限な期待値で、$X < Y$ $(a.e\ P)$ とする。このとき、$E[X] < E[Y]$ となる。また、線形性と 1.4 章の二つ目の結果から、$X \leq Y$ $(a.e.\ P)$ でどちらも有限の期待値を持つとすると $E[X] \leq E[Y]$ となり、$X = Y$ $(a.e.\ P)$ のときのみ、イコールが成り立つ。

有限な確率変数 X の分散を以下で定義する。

$$Var(X) = E[(X - E[X])^2]$$

もし、X が確率密度 p に対して絶対連続であるとする。式 (10) より以下が成り立つ。

$$Var(X) = \int (x - E[X])^2 p(x) dx$$

また、X が確率関数 p にたいして離散的であるとする。式 (11) より以下が成り立つ。

$$Var(X) = \sum_{x \in \mathcal{X}_0} (x - E[X])^2 p(x)$$

式 (12) を使うと、

$$Var(X) = E[(X^2 - 2XE[X] + (E[X])^2] = E[X^2] - (E[X])^2$$

が成り立つ。この結果は計算を省略するのによく使われる。

期待値が存在するならば、有限な期待値を持つ確率変数 X, Y の共分散は以下で定義される。

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{13}$$

また、$Cov(X, X) = Var(X)$ が成り立つ。式 (12) を使えば、

$$\begin{aligned} Cov(X, Y) &= E(XY - XE[Y] - YE[X] + E[X]E[Y]) \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{14}$$

共分散は、二つの確率変数間での線形的な関係であると考えられている。しかし、共分散は測度のスケールに依存される。なので、より自然な測度として相関係数がある。そして以下で定義される。

$$Cor(X, Y) = \frac{Cov(X, Y)}{[Var(X)Var(Y)]^{\frac{1}{2}}}$$

相関係数は常に $[-1, 1]$ の間にある。さらに、もし ± 1 のとき、二つの確率関数は完全な線形関係がある。

1.7 確率変数ベクトル

今、$X_1, \ldots, X_n$ を確率変数とし、確率変数ベクトルと呼ばれる関数 $X : \varepsilon \to \mathbb{R}^n$ を以下で定義する。

$$X(e) = \begin{pmatrix} X_1(e) \\ \vdots \\ X_n(e) \end{pmatrix} \quad e \in \varepsilon$$

この章で述べられている確率変数ベクトルの多くの事実や結果は、自然且つ直接的に確率変数に拡張される。例えば、$\mathbb{R}^n$ のボレル集合 B 上の X の分布 P_X は以下で定義され、$X \sim P_X$ と書かれる。

$$P_X(B) = P(X \in B) \stackrel{\text{def}}{=} P(\{e \in \varepsilon \ : \ X(e) \in B)$$

P_X が $\mathbb{R}^n$ 上のルベーグ測度に対して絶対連続であるとき、この確率変数ベクトル X や P_X は確率密度 p に対して絶対連続であると呼ばれる。この場合、

$$P(X \in B) = \int_{\mathrm{B}} \cdots \int p(x)dx$$

となる。もし確率変数 X が離散であり、何らかの可算集合 $\mathcal{X}_0 \in \mathbb{R}^n$ に対して $P(X \in \mathcal{X}_0) = 1$ であるとする。このとき、$p(x) = P(X = x)$ ならば P_X は $\mathcal{X}_0$ の数え上げ測度に対しての確率密度 p をもち、

$$P(X \in B) = \sum_{x \in \mathcal{X}_0 \cap B} p(x)$$

となる。確率変数ベクトル X の期待値ベクトルは、

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

となる。

今、$T : \mathbb{R}^n \to \mathbb{R}$ が可測関数であるとする。このとき $T(X)$ は確率変数であり、もし、期待値か積分が存在するならば式 (8) より、

$$E[T(X)] = \int T dP_X$$

となる。また、P_X が測度 μ に対して確率密度 p を持つなら、この積分は $\int Tpd\mu$ と表され、数え上げ測度、ルベーグ測度に対してそれぞれ

$$\sum_{x \in \mathcal{X}_0} T(x)p(x) \ or \ \int \cdots \int T(x)p(x)dx$$

となる。

1.8 共分散行列

行列 W_{ij} すべてが確率変数な時、W を確率変数行列と呼ぶ。もし、W が確率変数行列なら $E[W]$ は各期待値の行列となる。つまり以下が成り立つ。

$$(E[W])_{ij} = E[W_{ij}]$$

ここで、v を定数ベクトルとし、A, B, C を定数行列、X を確率変数ベクトル、W を確率変数行列とする。このとき、以下が成り立つ。

$$E[v\ +\ AX] = v + AE[X] \tag{15}$$
$$E[A\ +\ BWC] = A + B(E[W])C \tag{16}$$

これは、期待値の基本的性質から言えることがわかる。というのも $(v + AX)_i = v_i + \sum_j A_{ij}X_j$ であり、$(A + BWC)_{ij} = A_{ij} + \sum_k \sum_l B_{ik}W_{kl}C_{lj}$ であるからだ。

確率変数ベクトル X の共分散は、共分散行列となり、

$$[Cov(X)]_{ij} = Cov(X_i, X_j)$$

もし、$\mu = E[X]$ で $(X-\mu)'$ を $(X-\mu)$ を転置したもの（行ベクトル）とする。このとき以下が成り立つ。

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X-\mu)(X-\mu)']_{ij}$$
$$Cov(X) = E[(X-\mu)(X-\mu)'] \tag{17}$$

同様に、式 (14) か (16) より

$$Cov(X) = E[XX'] - \mu\mu'$$

となる。また、アフィン変換をつかったあとの共分散をみると、v を定数ベクトル、A を定数行列、X を確率変数ベクトルとすると以下が成り立つ。

$$\begin{aligned} Cov(v + AX) &= E[(v + AX - v - A\mu)(v + AX - v - v\mu)'] \\ &= E[A(X-\mu)(X-\mu)'A'] \\ &= AE[(X-\mu)(X-\mu)']A' \\ &= ACov(X)A' \end{aligned}$$

1.9 直積測度と独立性

測度空間 $(\mathcal{X}, \mathcal{A}, \mu)$ と $(\mathcal{Y}, \mathcal{B}, \nu)$ を用意する。このとき、一意の測度 $\mu \times \nu$ が存在し、直積測度と呼ばれる。$(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \vee \mathcal{B})$ に対して、すべての $A \in \mathcal{A}$ とすべての $B \in \mathcal{B}$ に対して

$$(\mu \times \nu)(\mathcal{A} \times \mathcal{B}) = \mu(A)\nu(B)$$

となる。σ 加法族 $\mathcal{A} \vee \mathcal{B}$ は、すべての部分集合 $A \times B$ ($A \in \mathcal{A}$,$B \in \mathcal{B}$) を含む最小の σ 加法族であると定義される。

例 1.15 μ と ν がそれぞれ $\mathbb{R}^n$ と $\mathbb{R}^m$ 上でルベーグ測度であるとき、$\mu \times \nu$ は $\mathbb{R}^{n+m}$ 上のルベーグ測度となる。

例 1.16 μ と ν がそれぞれ $\mathcal{X}_0$ と $\mathcal{Y}_0$ 上の数え上げ測度とするとき、$\mu \times \nu$ は $\mathcal{X}_0 \times \mathcal{Y}_0$ 上の数え上げ測度となる。

以下の定理は、$\mu \times \nu$ 上の積分結果は μ, ν 上の積分をそれぞれ繰り返すことで得られるということを表している。

定理 1.17 (Fubini)。今、$f \geq 0$ であるとすると、

$$\begin{aligned}\int f d(\mu \times \nu) &= \int \left[\int f(x,y) d\nu(y) \right] d\mu(x) \\ &= \int \left[\int f(x,y) d\mu(x) \right] d\nu(y)\end{aligned}$$

条件を $f \geq 0$ から $\int |f| d(\mu \times \nu) < \infty$ としても上記式は成り立つ。

今、$f = 1_S$ とする。例えば、$\mathcal{A}$ や $\mathcal{B}$ 上の直交座標系上の集合ではない S に対しても以上の方法によって、$(\mu \times \nu)(S)$ は算出される。

定義 1.18 (独立性)。二つの確率変数 $X \in \mathbb{R}^n$ と $Y \in \mathbb{R}^m$ が独立であるとは、すべてのボレル集合 A, B に対して

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B) \tag{18}$$

であるときである。

$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ であるとする。このとき $Z \in A \times B$ は $X \in A, Y \in B$ と同値条件であるため式 (18) は X, Y, Z の分布を使って以下のように書ける。

$$P_Z(A \times B) = P_X(A) P_Y(B)$$

つまり、Z の分布は直積測度であり

$$P_Z = P_X \times P_Y$$

となる。Z の確率密度は X, Y の確率密度の直積で表せられる。これはフビニの定理と式 (18) より表せられる。特に、P_X が μ に対して確率密度 p_X をもち、P_Y が ν に対して確率密度 p_Y 持つとする。このとき、

$$\begin{aligned}P(Z \in S) &= \int 1_S d(P_X \times P_Y) \\ &= \int \left[\int 1_S(x,y) dP_X(x) \right] dP_Y(y) \\ &= \int \left[\int 1_S(x,y) p_X(x) d\mu(x) \right] p_Y(y) d\nu(y) \\ &= \int 1_S(x,y) p_X(x) p_Y(y) d(\mu \times \nu)(x,y)\end{aligned}$$

これは、P_Z が確率密度 $p_X(x)p_Y(y)$ を持つことを表している。この本では、μ, ν は一般的に数え上げ測度かルベーグ測度である。ここでの概論では、二つの確率変数が片方絶対連続・片方離散の場合も考えられている。 もし、$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ ならば、P_Z は $X \cdot Y$ の同時確率分布であるといわれ、P_Z の確率密度は $X \cdot Y$ の同時確率確率密度と呼ばれる。もし、$X \cdot Y$ が確率密度 $p_X(x), p_Y(y)$ 持ち、独立ならば、同時確率確率密度は $p_X(x)p_Y(y)$ で書くことができる。

この考え方は、確率変数がいくつかあるときにも拡張できる。Z が、$X_1, \ldots, X_n$ から生成されているとする。このとき、Z の分布や確率密度は、$X_1, \ldots, X_n$ からなる同時確率確率密度・同時確率分布となる。もし、確率変数ベクトル $X_1, \ldots, X_n$ が独立ならば、ボレル集合 $B_1, \ldots, B_n$ に対して

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \times \cdots \times P(X_n \in B_n)$$

となる。また以下の一意な測度 μ に対して以下が成り立つことから、$P_Z = P_{X_1} \times \cdots \times P_{X_n}$ となる。

$$\mu(B_1 \times \cdots \times B_n) = P_{X_1}(B_1) \times \cdots \times P_{X_n}(B_n)$$

以下の命題は、独立な変数の関数は独立になることを示している。

命題 1.19 $X_1, \ldots, X_n$ を独立な確率変数とする。そして、$f_1, \ldots, f_n$ を可測関数とする。このとき、$f_1(X_1) \ldots f_n(x_n)$ は独立である。

もし、X_i が測度 μ_i に対して確率密度 p_{X_i} を持つならば、$X_1, \ldots, X_n$ は $\mu = \mu_1 \times \cdots \times \mu_n$ に対して同時確率確率密度 p を持ち、

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \times \cdots \times p_{X_n}(x_n)$$

となる。もし、X_i $(i = 1, 2, \ldots, n)$ が独立同一に $X_i \sim Q$ となるなら、$X_1, X_2, \ldots, X_n$ は独立同一分布 (i.i.d.) と呼ばれ、確率変数の列は、Q からの乱数と呼ばれる。

1.10 条件付確率分布

X と Y が確率変数ベクトルであるとする。もし、X が観測され、$X = x$ であるとわかったとする。このとき P_Y は、与えられた Y のみの分布関数と見ることはできない。P_Y に $X = x$ が与えられたという情報を加味して修正しなければならない。もし、X が離散なら条件付確率の式を使って分布関数を求めることができる。つまり、条件付確率分布は、$p_X(x) = P(X = x)$ と確率関数が与えられるとき、X が起こりえる値の集合 $\mathcal{X}_0 = \{x : p_X(x) > 0\}$ をとるとボレル集合 B と $x \in \mathcal{X}_0$ に対して以下のように定義できる。

$$Q_x(B) = P(Y \in B | X = x) = \frac{P(Y \in B, X = x)}{P(X = x)} \tag{19}$$

このとき、$x \in \mathcal{X}_0$ に対して Q_x は確率測度となる。また、$X = x$ が与えられた下での Y の分布関数と呼ばれる。

形式的に、条件付確率は確率的 transition-kernel となる。この確率的 transition-kernel の定義は $Q : \mathcal{X} \times \mathcal{B} \to [0, 1]$ が以下の二つの条件を満たすことである。一つ目は $x \in \mathcal{X}$ について $Q_x(\cdot)$ は $\mathcal{B}$ 上で確率測度となることである。二つ目は $B \in \mathcal{B}$ に対して、$Q_x(B)$ は x 上の可測関数となることである。

完備な空間では、$x \notin \mathcal{X}_0$ に対しても条件付分布を定義できる。どうしてそうなるかは大した問題でもない。なぜなら、Q_x を $x \notin \mathcal{X}_0$ に対してある一定の確率測度を取るようにしたものも一つの確率となるからである。

条件付分布は、X が離散でなくても存在する。しかし、定め方は技術が必要で、6 章で定めるものとは違う。しかしこの章で大事なのは、X が離散かそうでないかということである。特に、X, Y が独立で X が離散なら、x に関係なく $Q_x = P_Y$ となる。この事実は、一般に拡張しても成り立ち、興味深くまた利用しやすい式である。

条件付確率に対する積分は条件付期待値を与える。特に $X = x$ が与えられた下での $f(X, Y)$ の条件付き期待値は以下のように定義される。

$$E[f(X, Y)|X = x] = \int f(x, y) dQ_x(y) \tag{20}$$

もし、X, Y どちらも Y に対して離散で、Y は可算集合 $\mathcal{Y}_0$ 上、X は $\mathcal{X}_0$ 上で定義されているとする。このとき、$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ は可算集合 $\mathcal{X}_0 \times \mathcal{Y}_0$ 上で定義され、離散となり、確率関数は $p_Z(z) = P(Z = z) = P(X = x, Y = y)$ $(z = \begin{pmatrix} x \\ y \end{pmatrix})$ となる。また、式 (19) から $Q_x(\mathcal{Y}_0) = 1$ となる。また、Q_x は離散で $x \in \mathcal{X}_0$ に対し確率関数は以下で定義される。

$$q_x(y) = Q_x(y) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} \tag{21}$$

よって、式 (20) から期待値は以下のように求められる。

$$H(x) = E[f(X, Y)|X = x] = \sum_{y \in \mathcal{Y}_0} f(x, y) q_x(y)$$

となる。今後 $E[|f(X,Y)|] < \infty$ を仮定する。式 (21) で $P(X = x, Y = y) = q_x(y)p_X(x)$ と示されていることから、$f(X,Y)$ の期待値は以下のように書ける。

$$\begin{aligned} E[F(X,Y)] &= \sum_{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{X}_0 \times \mathcal{Y}_0} f(x,y)P(X=x,Y=y) \\ &= \sum_{x \in \mathcal{X}_0} \sum_{y \in \mathcal{Y}_0} f(x,y) q_x(y) p_X(x) \\ &= \sum_{x \in \mathcal{X}_0} H(x) p_X(x) \\ &= E[H(X)] \end{aligned}$$

この式は、条件付問題の基本的な式であり、全確率の法則等で呼ばれる。実際、この性質は X が離散でない場合でも条件付確率や期待値を定義するのに使うほど基本的な事項である。また、式 (20) から得ることができる確率変数 $H(X)$ は以下のように表せられる。

$$H(X) = E[f(X,Y)|X]$$

全確率の法則を使うと、以下が成り立つ。

$$E[f(X,Y)] = E[E[f(X,Y)|X]]$$

特に、$f(X,Y) = Y$ となるとき上記の式は

$$E[Y] = E[E(Y|X)]$$

$Y = 1_B$ のとき、つまり B の指示関数となるとき、$E[Y] = P(B)$ となり、$P(B|X) \stackrel{\text{def}}{=} E[1_B|X]$ から上記の式は

$$P(B) = E[P(B|X)]$$

となる。結果、一般の条件付き期待値や条件付分布でもこれらの性質は成り立ち、以下が成り立つ。

$$\begin{aligned} E[Y|X] &= E[E[Y|X,W]|X] \\ P(B|X) &= E[P(B|X,Y)|X] \end{aligned} \tag{22}$$