2. 確率変数とその分布

標本空間上の関数としての確率変数を考える。分布関数と生存関数の意味を述べる。離散分布について、ベルヌイ試行から導かれる二項分布、負の二項分布、幾何分布、偶然事象観測や標本抽出での超幾何分布とポアソン分布を説明し、二項分布との関係を導く。連続分布では、一様分布、正規分布について考える。確率変数や分布を要約する平均、分散、標準偏差を定義し、具体的な分布に対して計算する。確率変数間の関連を表現する共分散、相関係数および分散共分散行列について説明する。

2.1. 分布関数

標本空間Ωの根元事象に実数を対応させることは、現象を数量的に取り扱う上で重要である。例えば、ある電気製品の寿命を分析する場合は故障するまでの時間 t を観測する。また、ある実験を独立に n 回反復する場合は成功に 1、失敗に 0 を対応させ、成功回数 n_1 を記録する。このような関数

$$X:\Omega\to R$$

を確率変数(random variable)という。確率変数は大文字の $X,Y,Z,...$等を慣例で用い、その値は $x,y,z,...$で示される。

[例 2.1] サイコロを 1 個投げる試行では、$\Omega=\{i$ の目, $i=1,2,3,4,5,6\}$, P(i の目が出る) = 1/6 (i=1,2,3,4,5,6)である。この試行で

$$X(i\text{ の目})=i \qquad (i=1,2,3,4,5,6)$$

とすれば、この関数は確率変数である。

[例 2.2] ある母集団から無作為に一人を選び、その身長 X、体重 Y、あるいは血中成分 Z の濃度を測定するとき、これらの属性(attribute)は確率的に観測される量で、変量(variate)とよばれ確率変数である。

定義 2.1 確率変数 X に対して

$$F(x) = \mathrm{P}(X < x) \tag{2.1}$$

を X の分布関数(distribution function)という。 □

[例 2.3] 上の定義で、X をあるガンウイルスに現在感染している人（キャリア）がガンを発症するまでの時間とすれば、分布関数(2.1)は時間 x までに発症する確率を意味している。次の関数は時間 x までに発症していない確率であり、生存関数(survival function)という。

$$\begin{aligned} S(x) &= 1-F(x) \\ &= \mathrm{P}(X \geqq x) \end{aligned}$$ □

上の定義から分布関数は単調非減少、すなわち

$$x<y \text{ ならば } F(x)\leqq F(y)$$

である。また、

$$F(-\infty)\equiv \lim_{x\to-\infty} F(x)=0, \qquad F(+\infty)\equiv \lim_{x\to+\infty} F(x)=1 \tag{2.2}$$

が示される（証明略）。

問 2.1. (2.2)を証明せよ。

問 2.2. 例 2.1 で確率変数 X の分布関数を求め、そのグラフを描け。

2.2. 離散分布

確率変数 X の値域が実数 R のたかだか可算部分集合である場合、X を離散確率変数といい、このときの分布を離散分布(discrete distribution)という。標本空間を $\Omega=\{\omega_1,\omega_2,...,\omega_n,...\}$ とし、

$$X(\omega_n)=x_n \qquad (n=1,2,...)$$

であるとき、その確率分布 $p(x_n)\equiv \mathrm{P}(X(\omega_n)=x_n)$ $(n=1,2,...)$は

$$p(x_n)\geqq 0, \qquad \sum_{n=1}^{\infty} p(x_n)=1 \tag{2.3}$$

を満たす。従って、分布関数は

$$\begin{aligned} F(x) &= \mathrm{P}(X<x) \\ &= \sum_{x_n<x} p(x_n) \end{aligned}$$

である。上の和は $x_n<x$ を満たす全ての x_n に関する和を意味する。確率変数 X の平均 $\mathrm{E}(X)$ を

$$\mathrm{E}(X) = \sum_{n=1}^{\infty} x_n p(x_n) \tag{2.4}$$

で定義する。平均は期待値(expectation)とも呼ばれる。また、分布の散布度を分散(variance)として、次式で定義する。

$$\mathrm{Var}(X) = \sum_{n=1}^{\infty} (x_n-\mathrm{E}(X))^2 p(x_n) \tag{2.5}$$

分散は偏差 $X-\mathrm{E}(X)$の平方平均である。(2.5)式は

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{n=1}^{\infty} x_n^2 p(x_n) - \mathrm{E}(X)^2 \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \end{aligned}$$

のように変形できる。分散の正の平方根は標準偏差(standard deviation)である。平均、分散および標準偏差はギリシャ文字を用いて、それぞれ μ、σ^2, σ で示すのが慣例である。

(i) ベルヌイ分布

ある試行を一回行い、事象 A（成功）の出現に注目し、$\mathrm{P}(A)=p$（$q=1-p$）とする。いま、$X=1$ (A が出現)、$X=0$ ($\overline{A}$ が出現)とするとき、この分布をベルヌイ分布という。

$$\mathrm{E}(X)=1\times p+0\times q=p,\ \mathrm{Var}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2=pq$$

を得る。

(ii) 二項分布

一回の試行で起こる事象 A の出現に注目し、P(A)=p とする。この試行（ベルヌイ試行）を独立に n 回反復するときの A の出現回数を X で示すと、

$$\mathrm{P}(X=k) = {}_n\mathrm{C}_k p^k q^{n-k} \qquad (k = 0,1,2,...,n) \tag{2.6}$$

で与えられる。ただし、$q=1-p$ である。この分布を二項分布(binomial distribution)といい、記号 $B_N(n\,,\,p)$で表現する。例えばある実験における結果 A=「成功」に注目する場合や、実験動物を使った実験や無作為に抽出された患者に対する臨床試験結果を A=「効果あり」と「なし」で記録する場合の分布は二項分布である。二項分布の平均は

$$\mathrm{E}(X) = \sum_{k=0}^{n} k \times {}_nC_k p^k q^{n-k} = np\sum_{k=1}^{n} {}_{n-1}C_{k-1} p^{k-1} q^{n-k}$$

$$= np\sum_{k=0}^{n-1} {}_{n-1}C_{k-1} p^{k} q^{n-k-1} = np$$

である。次に分散を求める。

$$\mathrm{E}(X(X\text{-}1)) = \sum_{k=0}^{n-1} k(k-1)\,{}_n\mathrm{C}_k p^k q^{n-k}$$

$$= n(n-1)p^2 \sum_{k=2}^{n} {}_{n\text{-}2}\mathrm{C}_{k\text{-}2}\, p^{k-2} q^{n-k} = n(n-1)p^2$$

このことから

$$\mathrm{E}(X^2) = \mathrm{E}(X(X-1)) + \mathrm{E}(X) = n(n-1)p^2 + np$$

が得られる。従って

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = npq$$

を得る。

問 2.3　(2.6)で $n = 5$, $p = 1/3$ として、分布関数のグラフを描け。

(iii) 超幾何分布

N 人のクラスの中に男子が M 人、女子が N-M 人いる。この中から n 人を無作為に抽出するとき、その中に男子が k 人含まれる確率は

$$p(k) = {}_M\mathrm{C}_k\, {}_{N-M}\mathrm{C}_{n-k} / {}_N\mathrm{C}_n \tag{2.7}$$

となる。但し、$n \leqq N$, $k \leqq M$, $n\text{-}k \leqq N\text{-}M$ である。この分布を超幾何分布という。この分布で N と M が非常に大きい場合を考える。いま、$p = M/N$ とすれば

$$p(k) = {}_nC_k p^k (1-p)^{n-k} \tag{2.8}$$

で近似される。例えば、ある地域の人口 N 人が十分大きく、その中で特定のウイルスのキャリアが M 人（未知）いるとされる場合の調査を考える。この場合はこの地域の住民から無作為に n 人を抽出し、その中に含まれるキャリアの数 k 人を観測する事になる。このときキャリアの比率 $p = M/N$ に興味があり、標本の分布は二項分布 $\mathrm{B}_N(n,p)$として良い。ま

た、比率 p は k/n で推定する。

[例2.4] ヒトT細胞白血病ウイルスI型のキャリアは日本ではおよそ1%と推定されている。大分県の献血データ(1999 年)では k=1027, n=69388 が調査され、キャリア比率の推定値 k/n ＝0.0148 を得る。□

問 2.4. 近似式(2.8)を証明せよ。

問 2.5. 人や動物を用いた実験では、種々の条件や費用面から多くの実験単位（被験者や動物）を用いることが出来ない場合がある。12 単位を無作為に 6 単位ずつ 2 群に分け、一方をコントロール、他方を処置群とした。この実験での効果は次の表で示される。処置群で効果ありが 3 単位出現したことは特に顕著な結果と言えるか考える。そのためには、処置効果がないと仮定して、周辺度数を固定したときの処置群での効果あり単位数 X が 3 以上となる確率を計算する（フィッシャーの正確検定)。この分布が超幾何分布であり、この確率を計算せよ。

表 2.1. 実験結果

	効果あり	なし	計
処置	3	3	6
コントロール	1	5	6
計	4	8	12

2.3. 連続分布

人の身長や体重、電気製品の故障までの使用時間（寿命）等は連続値（計量値）として観測される。このような確率変数を連続確率変数という。X を連続確率変数とするとき、分布関数 $F(x)$は

$$F(x) = \mathrm{P}(X < x) = \int_{-\infty}^{x} f(t)dt \tag{2.9}$$

で示される。この場合、

$$\frac{dF(x)}{dx} = f(x) \tag{2.10}$$

であり、これを密度関数(density function)という。確率変数の平均を

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{2.11}$$

で定義する。また、散布度を示す分散は

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} \big(x - E(X)\big)^2 f(x)dx \tag{2.13}$$

で定義し、その正の平方根 $\sqrt{\mathrm{Var}(X)}$ は標準偏差である。離散分布の場合と同様に

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

が成立する。平均、分散および標準偏差はそれぞれギリシャ文字 μ, σ^2, σ を用いて表す場合が多い。

(i) 一様分布

区間[0,1]から無作為に 1 個の実数 X を取り出すとき、この分布は次の密度関数を持つ。

$$f(x) = \begin{cases} 1 & (0 \le x \le 1) \\ 0 & (\text{その他の場合}) \end{cases} \tag{2.14}$$

である。この分布を区間[0,1]上の一様分布(uniform distribution)という。分布関数は

$$F(x) = \begin{cases} 0 & (x < 0) \\ x & (0 \le x \le 1) \\ 1 & (x > 1) \end{cases} \tag{2.15}$$

である。この場合は E(X) = 1/2 である。一様分布は乱数を生成する場合に利用される（第 1 章)。この分布を利用して、他の分布の乱数を発生させる実験に後の節で触れる。一般に区間[a,b]上の一様分布が考えられる。

[例 2.5]　半径 1 の円で図のように自由に動く針を動かす。止まったときの角度 X(孤度）を測定すれば、X は区間[0,2π)上の一様分布に従う

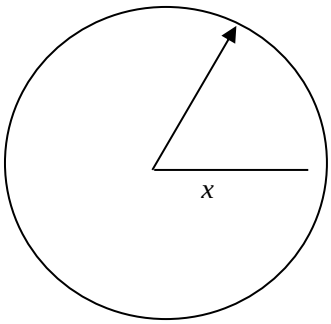

図2.1. 円盤上の針と偏角X

区間[0,1]上の一様分布の分散は次のように計算できる。

$$\mathrm{Var}(X) = \int_0^1 \left(x - \frac{1}{2} \right)^2 dx = \frac{1}{12}$$

このとき、P(μ−σ≦X≦μ+σ)と Pr(μ−2σ≦X≦μ+2σ)を求める。

$$\mu-\sigma \le X \le \mu+\sigma \quad \Rightarrow \quad \frac{1}{2} - \frac{\sqrt{3}}{6} \le X \le \frac{1}{2} + \frac{\sqrt{3}}{6}$$

このことから、

$$\mathrm{P}(\mu-\sigma \le X \le \mu+\sigma) = \frac{\sqrt{3}}{3} \approx 0.577$$

一方

$$\mu-2\sigma \leqq X \leqq \mu+2\sigma \quad \Rightarrow \quad -0.08 \approx \frac{1}{2}-\frac{\sqrt{3}}{3} \leq X \leq \frac{1}{2}+\frac{\sqrt{3}}{3} \approx 1.08$$

であり、上の範囲は区間[0,1]を含んでいる。従って、

$$P(\mu-2\sigma \leqq X \leqq \mu+2\sigma) = 1$$

となる。以上から、この分布は平均に関して約$\pm 2\sigma$に範囲に、全てのデータが存在することが分かる。

問 2.6.　X が$[a,b]$上の一様分布に従うとき $Y = (X - a)/(b - a)$の分布を求めよ。

問 2.7.　X が$[a,b]$上の一様分布に従うとき、X の平均と分散を求めよ。

(ii) 正規分布

確率変数 X の密度関数が

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.16}$$

を正規分布(normal distribution)という。ここに、μ は平均と σ^2 は分散である。この分布を $N(\mu, \sigma^2)$で示す。この分布は計量値の分布として最も広く用いられる。分布関数は

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \tag{2.17}$$

であり、明確な表現は取れない。確率の計算はコンピュータによる数値計算が行われる。ここで、次の変数変換を考える。

$$Z = \frac{X-\mu}{\sigma} \tag{2.18}$$

これを、X の標準化という。(2.18)の分布は積分(2.16)の変数変換を考えれば良く、その密度関数 $\phi(z)$は

$$\phi(z) = f(\sigma z+\mu)\frac{dx}{dz} = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \tag{2.19}$$

である。これを標準正規分布(standard normal distribution)といい、$N(0,1)$で示す。正規分布表は標準正規分布に関して作成されているので、正規分布表を用いる場合はこの標準化が重要である。標準正規分布の分布関数を$\Phi(z)$で示す。表計算ソフト（エクセル）を用いて、分布関数の値は簡単に求められる。例えば、変数を標準化していれば、$\Phi(z)$は次のように計算できる。セル A1 に特定の数値 1.96 を代入し、分布関数の値を A2 に出力する形式を作るためにはセル A2 に関数

=NORMSDIST(A1)

を設定すれば、自動的に 0.975002 が得られる。このことを、

A2=NORMSDIST(A1)

で表現することにする。A1 に 1 を代入すれば 0.841345 が得られる。このようなプログラムは出来るだけ汎用的に作成した方が便利である。

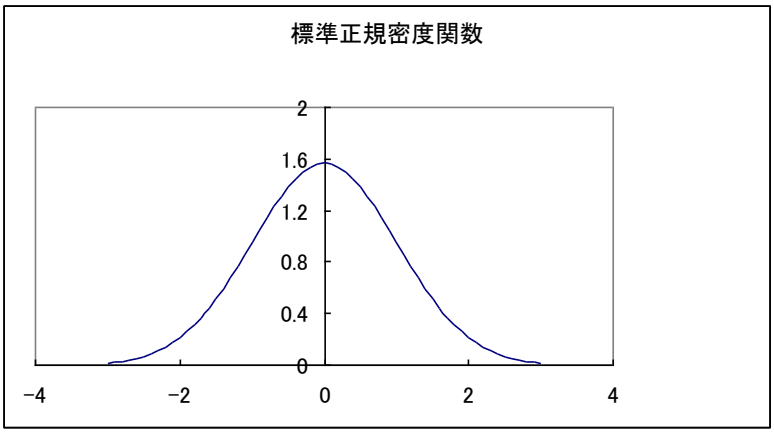

図 2.2. 標準正規分布の密度関数のグラフ

問 2.8. 平均, 分散と確率変数の値を代入し、分布関数の値を出力するプログラムをエクセルのファイルで作成せよ。

正規分布の平均が μ であることを確かめる。

$$E(X)=\int_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx=\int_{-\infty}^{+\infty}(\sigma z+\mu)\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}dz \qquad (\because (2.18, 19)\text{から})$$

$$=\int_{-\infty}^{+\infty}\sigma z\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}dz+\int_{-\infty}^{+\infty}\mu\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}dz=\mu$$

である。正規密度関数のグラフから、母数 σ^2 は分布の分散である。すなわち、σ^2 が大きくなるとき分布はなだらかな一山型になり、σ^2 が小さくなれば $X=\mu$ の周辺に集中する。母数 μ は平均であり、分布から中央値でもある。

問 2.9. 正規分布の分散が σ^2 になることを確かめよ。

問 2.10. 確率変数(2.18) の平均と分散がそれぞれ 0 と 1 であることを示せ。

表計算ソフトを用いれば､正規分布表を作ることが出来る｡標準化した確率変数(2.18) の平均と分散はそれぞれ 0 と 1 である。

表 2.2 は 0.01 毎に $0\le Z\le 3$ の範囲で作成した上側確率

$$Q(z)=1-\Phi(z) \tag{2.20}$$

の表である。A2 から A31 に 0.0 から 3.0 までを 0.1 刻みで代入し、1 行目の B1 から K1 に 0 から 0.09 までを 0.01 刻みで代入する。内側のセルには、次のように関数を代入する。例えば、B2 には A2+B1 を z として、関数(2.20)を代入する。すなわち

B2 =1−NORMSDIST(A2+B1)

とする。後で、この関数を全てのセルにコピーするために

B2 =1−NORMSDIST($A2+$B1)

とすると便利で、この関数を全てのセルにコピーするだけで表 2.4 を得る。この表から

$$P(\mu-\sigma \leqq X \leqq \mu+\sigma)=1-0.1587\times 2=0.6826$$
$$P(\mu-2\sigma \leqq X \leqq \mu+2\sigma)=1-0.0228\times 2=0.9545$$

を得る。

[例 2.6]　女子学生の平均身長 Xcm は N(159.5,25)とする。このとき、P(X > 165)を求める。確率変数の標準化は

$$Z=\frac{X-159.5}{5}$$

であるから、求める確率は

$$P(X>165) = P(Z>1.10)= Q(1.10)$$

になる。表 2.2 を用いる場合は、第 1 列の 1.1 と第 1 行の 0.00 に対するセルの数値から 0.1256 を得る。エクセルで直接計算する場合は 1−NORMSDIST(1.10)で計算できる。

問 2.11. 例 2.6 で P(150 < X < 162)を求めよ。

問 2.12. 正規分布の密度関数(2.16)の変曲点を求めよ。

大学入試等で用いられる偏差値は学力偏差値と呼ばれ広く用いられるが、その定義は次のように与えられる。試験の成績 X の平均と分散をそれぞれ μ, σ^2 とするとき、

$$S=10\frac{X-\mu}{\sigma}+50$$

である。すなはち、偏差値の平均は 50 で、分散は 100 である。

問 2.12. 試験の成績が正規分布しているものとして、偏差値 75 以上の人は受験生の何%か。

正規分布に関する取り扱いや解析法は理論的に十分確立されていて、通常の統計解析では連続データに正規分布を仮定して行う場合が最も多い。実際の現象ではデータ分布の右裾が長い、非対称分布である場合もある（図 2.3)。

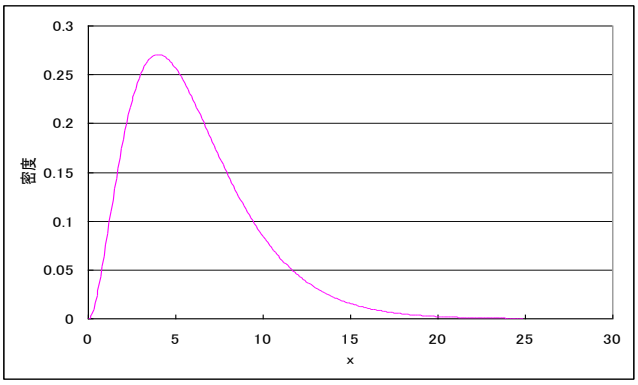

図 2.3. 右裾が重い分布

表 2.2. 正規分布表 Q(z)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

2.4. 多次元分布

複数の確率変数を同時に考えることは、確率変数間の関係を考える上で重要である。

定義 2.2. 確率変数 $X_1,X_2,...,X_p$ に対して、ベクトル $X = (X_1,X_2,...,X_p)'$ を p 次元確率変数、または p 次元確率ベクトル(random vector)という。ここに$(X_1,X_2,...,X_p)'$ における記号' は転置であり、p 次元の列ベクトルを意味する。

一次元の場合と同様に、分布関数や密度関数はその拡張として定義される。

定義 2.3. 確率ベクトル $X = (X_1,X_2,...,X_p)'$に対して、その平均を

$$E(X)= (E(X_1),E(X_2),...,E(X_p))'$$

で定義し、これを平均ベクトルという。

[例 2.7] 一回の試行結果について排反事象 $A_1,A_2,...,A_I$ に注目する場合を考える。例えば、ある実験結果を複数のカテゴリー(category)で評価する場合に相当し、とくに $I = 2$ であれば、(i)になる。この試行を独立に n 回反復し、事象 A_i が起こる回数を X_i とすれば、$(X_1,X_2,...,X_I)$ の分布は

$$P\{(X_1,X_2,...,X_I) = (k_1,k_2,...,k_I)\} = \frac{n!}{k_1!k_2!...k_I!} p_1^{k_1} p_2^{k_2} ... p_I^{k_I} \tag{2.21}$$

で与えられる。ただし、

$$\sum_{i=1}^{I} p_i = 1, \quad \sum_{i=1}^{I} k_i = n$$

である。この分布を多項分布(multinomial distribution)という。 □

問 2.14. 上の多項分布$(X_1,X_2,...,X_I)$で$\left(X_1, \sum_{i=2}^{I} X_i\right)$は二項分布に従うことを示せ。

また、平均ベクトルを求めよ。

上の例のように確率ベクトル$(X_1,X_2,...,X_I)$の分布を確率変数 $X_1,X_2,...,X_I$ の同時分布(joint distribution)という。また、問 2.24 の場合は各確率変数 X_i の分布を考えているので、この分布を周辺分布(marginal distribution)いう。確率ベクトル $X = (X_1,X_2)'$で、確率変数 X_1, X_2 はそれぞれカテゴリー$\{R_1,R_2,\ldots,R_I\}$, $\{C_1,C_2,\ldots,C_J\}$をもつ場合に、同時分布 π_{ij} と周辺分布 π_{i+},π_{+j} は表 2.7 で示される。ここに、

$$\pi_{ij} = P\left(X = R_i, Y = C_j\right), \quad \pi_{i+} = P\left(X = R_i\right), \pi_{+j} = P\left(Y = C_j\right)$$

とする。同時分布は確率変数間の連関情報をもち、周辺分布は確率変数それぞれの個別の分布である。

表 2.3. $I \times J$ 分割表分布

$X \backslash Y$	C_1	C_2	…	C_J	合計
R_1	π_{11}	π_{12}	…	π_{1J}	π_{1+}
R_2	π_{21}	π_{22}	…	π_{2J}	π_{2+}
…					
R_I	π_{I1}	π_{I2}	…	π_{IJ}	π_{I+}
合計	π_{+1}	π_{+2}	…	π_{+J}	$\pi_{++}=1$

連続多変量分布を簡単のための２変量確率ベクトル $X = (X_1,X_2)$'で説明する。X の同時分布関数を $F(x_1,x_2)$とすれば

$$F(x_1,x_2)= P(X_1<x_1,X_2<x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(t_1,t_2)\mathrm{d}t_1\mathrm{d}t_2$$

で示される。この場合、$f(x_1,x_2)$は $X = (X_1,X_2)$'の同時密度関数である。

[例 2.8] $X = (X_1,X_2)$の同時密度関数を

$$f(x_1,x_2) = 1 \ ((x_1,x_2)\in[0,1]\times[0,1]), \quad 0 \quad (\text{その他の場合})$$

とすれば、同時分布関数は

$$F(x_1,x_2)=\begin{cases} x_1x_2 \ ((x_1,x_2)\in[0,1]\times[0,1]), \\ x_1 \ ((x_1,x_2)\in[0,1]\times(1,\infty), \\ x_2 \ ((x_1,x_2)\in(1,\infty)\times[0,1]), \\ 1 \ ((x_1,x_2)\in(1,\infty)\times(1,\infty)), \\ 0 \text{（その他）} \end{cases}$$

である。

問 2.15. 例 2.8 で $Y = X_1+2X_2$ の分布を求めよ。

二変量正規分布については、2.7 節で触れる。

2.5 条件付き確率分布

p 次元確率ベクトル $X= (X_1,X_2,...,X_p)$' における変数 X_k を２つのグループ $X^{(1)}, X^{(2)}$に分割するとき、$X^{(2)}= x^{(2)}$を与えたときの $X^{(1)}$の分布を条件付き分布という。簡単のために、X と Y の 2 つの確率変数を考える。X と Y が離散的でそれぞれ$\{1,2,...,I\}$と$\{1,2,...,J\}$の値をとるとき、$X = i$ を与えたときの Y の条件付き分布は

$$\mathrm{P}(Y=j|X=i)= \mathrm{P}(X=i,Y=j)/\mathrm{P}(X=i) \quad (j= 1,2,...,J). \tag{2.22}$$

で与えられる。上式で

$$\mathrm{P}(Y=j|X=i)= \mathrm{P}(Y=j),$$

すなわち、

$$\mathrm{P}(X=i,Y=j) = \mathrm{P}(X=i)\mathrm{P}(Y=j)$$

のとき、X と Y は独立という。I=J=2 の場合に X と Y の連関を考える。相対リスク(relative

risk)は

$$RR = P(Y=2|X=2)/P(Y=2|X=1)$$

で定義され、X=2 の場合の条件付確率と X=1 の場合の確率の比の値である。また、オッズ(odds)は勝負事の勝ち目を意味し、X=1 と X=2 のときのオッズは

$$\frac{P(Y=2\,|\,X=1)}{P(Y=1\,|\,X=1)},\quad \frac{P(Y=2\,|\,X=2)}{P(Y=1\,|\,X=2)}$$

であり、Y=2 で"勝ち"、Y=1 で"負け"を表し、X=1 と X=2 は条件と考えれば理解が容易である。条件 X=2 と X=1 に対するオッズ比は

$$OR = \frac{P(Y=2\,|\,X=2)/P(Y=1\,|\,X=2)}{P(Y=2\,|\,X=1)/P(Y=1\,|\,X=1)}$$

で定義される。この連関測度ではリスクや"勝ち""負け"に合わせて、RR>1 (OR> 1) のとき正の連関、RR<1 (OR<1) のとき、負の連関である。また、RR= 1 (OR= 1)は X と Y が独立であることを意味する。

問 2.16.　RR=1 (OR= 1)のとき X と Y が独立であることを証明せよ。

X と Y が連続のときは、$X = x$ を与えたときの Y の条件付き分布を、次の密度関数で定義する。

$$f(y|x) = f(x,y)/g(x) \tag{2.23}$$

ここに、 $f(x,y)$は X と Y の同時密度関数、g(x)は X の周辺密度関数である。もし、

$$f(x,y)= g(x)h(y)$$

ならば、X と Y は独立という。ここに$h(y)$は Y の周辺密度関数である。X と Y の同時分布関数を $F(x,y)$、X と Y の周辺分布関数を、それぞれ G(x)と H(y)で示すとき、X と Y が独立であることは、

$$F(x,y)=G(x)H(y)$$

と同値である。

[例 2.9] ある成人男性の患者群では喫煙の有無 X と慢性気管支炎 Y との同時分布が次の表のようになっていると仮定する。

		慢性気管支炎 Y		
		無し(0)	有り(1)	計
X	無し(0)	0.57	0.03	0.60
	有り(1)	0.32	0.08	0.40
計		0.89	0.11	1.00

この場合、喫煙の有無Xを与えたときの、慢性気管支炎Yの条件付分布を考える。条件付分布は次のように与えられる。

	慢性気管支炎Y		
	無し(0)	有り(1)	計
X 無し(0)	0.95	0.05	1.0
有り(1)	0.8	0.2	1.0

この表から喫煙の慢性気管支炎に与える影響が分かる。２元分割表で相対危険は

$$P(Y=1|X=1)/P(Y=1|X=0) = 0.2/0.05 = 4$$

である。危険は喫煙者が非喫煙者に比べて4倍に上がることが分かる。

問2.17. 上の例をオッズ比で議論せよ。

2.6. 期待値

ここでは期待値の性質をまとめる。確率変数 X_n (i=1,2,...,n)の関数 $G(X_1,X_2,...,X_n)$も関数であり、その期待値 $E(G(X_1,X_2,...,X_n))$は定義される。

定理 2.2. 確率変数 X_n (i=1,2,...,n)の和および積の期待値に関しては次の命題が成立する。

(i) 実数 α と確率変数 X に対して

$$E(\alpha X)= \alpha E(X).$$

(ii) 確率変数 X と Y について

$$E(X + Y) = E(X) + E(Y).$$

(iii) 実数 α_i と確率変数 X_i (i=1,2,...,n)に対して

$$E\left(\sum_{i=1}^{n}\alpha_i X_i\right)=\sum_{i=1}^{n}\alpha_i E(X_i)$$

(iv) 確率変数 X と Y が独立ならば

$$E(XY)=E(X)E(Y).$$

(v) 確率変数 X と Y が独立ならば

$$Var(X+Y)=Var(X)+Var(Y).$$

[例 2.10] X_i (i = 1,2,...,n)を成功確率 p のベルヌイ試行とすれば、$Y = \sum_{i=1}^{n} X_i$ は n を固定するとき、Y は二項分布 $B_N(n,p)$に従う。$E(X_i)= p$ であるから、上の定理(ii)を用いて

$$E(Y)=\sum_{i=1}^{n} E(X_i) = np$$

を得る。また、(v)を用いると

$$Var(Y)=\sum_{i=1}^{n}Var(X_i)=np(1-p)$$

が分かる。

問 2.24 定理 2.2 を用いて超幾何分布の平均を求めよ。

2.7. 相関係数

確率変数 X, Y に対して共分散(covariance)を

$$\mathrm{Cov}(X,Y) = \mathrm{E}\{(X-\mu_1)(Y-\mu_2)\} \tag{2.24}$$

で定義する。ここに μ_1 と μ_2 は X と Y のそれぞれの平均である。共分散は次のように変形できる。

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mu_1\mu_2.$$

共分散で $X=Y$ のとき共分散は分散である。

問 2.25. 確率変数 X,Y,Z の同時分布が 3 項分布に従うとき、その共分散 $\mathrm{Cov}(X,Y)$を求めよ。ただし、X,Y,Z の周辺分布はそれぞれ $\mathrm{B}_N(n,p_1)$, $\mathrm{B}_N(n,p_2)$, $\mathrm{B}_N(n,p_3)$とする。

共分散は変量間の関連性を論じるのに重要である。

定理 2.3. 確率変数 X, Y に対して、次の不等式が成立する。

$$\mathrm{E}(XY)^2 \leqq \mathrm{E}(X^2)\mathrm{E}(Y^2). \tag{2.25}$$

証明 任意の実数 t に対して

$$\mathrm{E}\{(tX+Y)^2\} = t^2\mathrm{E}(X^2)+2t\mathrm{E}(X,Y)+\mathrm{E}(Y^2) \geq 0$$

が成立する。このことから、定理が示される。

この不等式はコーシーの不等式と呼ばれる。不等式(2.25)で、確率変数 X, Y をそれぞれの偏差 $X-\mu_1$ と $Y-\mu_2$ で置き換えるとき、

$$|\mathrm{Cov}(X,Y)| \leqq \sqrt{Var(X)}\sqrt{Var(Y)}\,. \tag{2.26}$$

が得られる。

定義 2.4. 確率変数 X, Y の相関係数(correlation coefficient)を

$$\mathrm{Corr}(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \tag{2.27}$$

で定義する。

相関係数は

$$-1 \leqq \mathrm{Corr}(X,Y) \leqq 1$$

を満たし、$\mathrm{Corr}(X,Y) > 0$ (<0) のとき正 (負) の相関(positive (negative) correlation)、$\mathrm{Corr}(X,Y) = 0$ のとき無相関(non-correlation)、さらに $\mathrm{Corr}(X,Y)=\mp 1$ のとき完全相関という。完全相関のときは

$$a(X-\mu_1) + b(Y-\mu_2) = 0$$

を満たす定数 a と b が存在する。このことは、一方の確率変数が他方の一次関数になっていることを意味する。確率変数の共分散は幾何ベクトルの内積に相当し、相関係数はベクトル間のなす角の余弦に相当する。この意味で相関係数は確率変数間の線形関係の強さを

表現した測度で、主として連続確率変数に用いられる。相関係数については次の定理が得られる。

定理 2.4. 次の(i)から(iv)は同値である。

(i) Corr(X,Y)＝0

(ii) Cov(X,Y)＝0

(iii) E(XY)＝E(X)E(Y).

(iv) Var(X+Y)＝Var(X)+Var(Y).

証明 (i) <=> (ii) (2.21)式より示される。

(ii) <=> (iii): Cov(X,Y)＝E(XY)−E(X)E(Y)より示される。

(ii) <=> (iv): Var(X + Y)＝Var(X)+2Cov(X,Y)+Var(Y)から示すことが出来る。

注意 2. 確率変数 X, Y が独立ならば、無相関である。多変量正規分布(multivariate normal distribution)の場合は、この定理の逆も成立し、独立性と無相関性は同値である。

確率変数 $X_1,X_2,\ldots,X_p$ で X_i と X_j の共分散を σ_{ij} のとする。ここに σ_{ii} は X_i の分散である。これらの共分散を行列に配列したものを分散共分散行列という。通常は Σ で表現する。すなわち、

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{2.28}$$

この行列は対称行列である。分散共分散行列は統計解析で重要となる。