基礎統計

furaga

7　多次元の確率分布

7.1　同時確率分布と周辺確率分布

- 同時確率分布
  二つの離散型の確率変数 X,Y をベクトル (X,Y)[*1]と表したときの、X=x かつ Y=y である確率。

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

  と定義する。また、事象 A が起こる確率 P(A) は

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

  で求められる。
- 同時確率密度関数
  上の同時確率分布の積分バージョン。確率変数 X,Y が連続型のときのもの。

$$P((X, Y) \in A) = \int \int_{A} f(x, y) dx dy$$

- 周辺確率分布
  離散型確率変数 X、Y 単独の確率分布。要する (?) に、条件が「X=x」(または「Y=y」) だけしかなくて、Y（X）の値は別に何でもいいよーってときの確率。２つ上の同時確率分布から、

$$g(x) = P(X = x, Y = \text{すべての実数}) = \sum_{y} f(x, y)$$

$$h(y) = P(X = \text{すべての実数}, Y = y) = \sum_{x} f(x, y)$$

- 周辺確率分布関数
  上の周辺確率分布の積分バージョン。X,Y が連続型のとき。

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

[*1] つまり、X、Y はそれぞれ xy 座標系の x 座標、y 座標の値

- 共分散

  確率変数 X、Y が関連しながらばらつく程度を表したもの。X,Y の関係の方向[*2]を

$$\begin{aligned} Cov(X,Y) &= E\{(X-\mu_X)(Y-\mu_Y)\} \qquad (\mu_X = E(X), \mu_Y = E(Y) \\ &= E(XY) - E(X)E(Y) \qquad (\text{実際の計算ではこちらを使う}) \end{aligned}$$

- 相関係数

  共分散を X,Y の標準偏差で割って、X,Y の関係の強さの程度を判断できるようにしたもの。

$$\begin{aligned} \rho_{XY} &= \frac{Cov(X,Y)}{D(X)D(Y)} \\ &= \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} \qquad (V(X), V(Y) : \text{それぞれ } X, Y \text{ の分散}) \end{aligned}$$

  なお、$-1 \leq \rho \leq 1$。

  $0 \leq \rho$ なら、X, Y と同じ大小の向きに変化し、$\rho \leq 0$ ならその逆になる傾向がある。ここでいう傾向とは平均的・確率的なもので、ρ の絶対値が大きいほど、その傾向は確定的になる。

  特に $\rho = \pm 1$ のとき Y=aX+b という一次式が成立する。(ただし $\rho = 1$ なら $a > 0$、$\rho = -1$ なら $a < 0$)

- 無相関

  $\rho_{XY} = 0$ であるとき、「X と Y は無相関である」という。独立とは似て非なるもの。

- 独立

  任意の x,y について、条件

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

  が成り立つとき、X,Y は互いに独立であるという。

  独立ならば無相関だが、無相関でも独立だとは限らない (独立⇒無相関)

7.2 条件付確率分布と独立な確率変数

- 条件付確率

  あらかじめ、何かしらの条件が与えられた後に、ある事象が起こる確率。

- 条件付確率密度関数

  Y=y という条件が与えられたときの X の条件付確率密度関数を

$$g(x|y) = \frac{f(x,y)}{h(y)}$$

[*2] 正の相関関係があるか、負の相関関係があるか、など

X=x という条件が与えられたときの Y の条件付確率密度関数を

$$h(y|x) = \frac{f(x,y)}{g(x)}$$

と定義する。

- 条件付期待値 ・ 条件付分散

  条件付確率における、期待値と分散。

  Y=y と与えられたとき、X の条件付期待値、条件付分散は、

  X, Y が離散型のとき、

$$E(X|y) = \mu_{X|Y} = \sum_x x \cdot g(x|y) \qquad (1)$$

$$V(X|y) = \sum_x (x - \mu_{X|Y})^2 g(x|y) \qquad (2)$$

  連続型のとき、

$$E(X|y) = \mu_{X|Y} = \int_{-\infty}^{\infty} x \cdot g(x|y)dx$$

$$V(X|y) = \int_{-\infty}^{\infty} (x - \mu_{X|Y})^2 g(x|y)dx$$

- 独立[*3]

  任意の x,y について、条件

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

  が成り立つとき、X,Y は互いに独立であるという。

  このとき、

$$g(x|y) \equiv {}^{*4}g(x),\ h(y|x) \equiv h(y)$$

  が成り立つ。(条件付確率の式 (1)(2) に $f(x) = g(x)h(y)$ を代入すれば出てきます)

- 積の期待値

  X,Y が独立のとき、積 XY の期待値 E(XY) について、

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

  が成り立つ。

- 独立と無相関の関係

  上の式から、X, Y が独立のとき、

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0 \quad \text{より、}$$

$$\rho_{XY} = \frac{Cov(X,Y)}{D(X)D(Y)} = 0$$

  よって、独立　⇒　無相関。逆は一般に成り立たない。

[*3] 大事そうなので、あえて二回書きました。

[*4] 両辺は同値ですよって意味

- 独立のときのモーメント母関数 X,Y が独立ならモーメント母関数について、

$$M_{X+Y}(t) = X_X(t)M_Y(t)$$

7.3 独立な確率変数の和

- 期待値、分散の加法性確率変数 $X_1,\ X_2,\ \cdots,\ X_n$ について、
  期待値は独立の如何にかかわらず、つねに

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

  が成立する。一方分散は $X_1,\ X_2,\ \cdots,\ X_n$ が独立の時に限り、

$$V(X_1 \pm X_2 \pm \cdots \pm X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$$

  特に $X_1,\ X_2,\ \cdots,\ X_n$ が同一の確率分布に従うとき、その期待値 ・ 分散を μ、σ^2 とすれば、

$$E(X_1 + X_2 + \cdots + X_n) = n\mu, \qquad V(X_1 + X_2 + \cdots + X_n)n\sigma^2$$

  であり、標準偏差は、$D(X_1 + X_2 + \cdots + X_n) = \sqrt{n}\sigma^2$。したがって、**標準偏差は $\sqrt{n}$ に比例**する。
- 相加平均 $X_1,\ X_2,\ \cdots,\ X_n$ の平均 $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ の期待値 ・ 分散はそれぞれ、

$$E(\bar{X}) = \mu, \qquad V(\bar{X}) = \frac{\sigma^2}{n}$$

  したがって、変数の数 n が大きくなるにつれて、分散は小さくなり、0 に収束する。(→大数の法則)
- たたみこみ
  独立な二つの確率変数 X,Y(それぞれの確率分布を g(x),h(y) とする) において、$X + Y(= z)$ の確率分布 k(z) は、

$$k(z) = \int_{-\infty}^{\infty} g(x)h(z - x)$$

- 再生性確率変数 X,Y が同一種類の確率分布にしたがっているとき、
  g,h のたたみこみの結果、ふたたび同一種類の確率分布 k がえられるとき、その確率分布は再生的であるという。再生的な確率分布の例 : 二項分布 ・ ポアソン分布 ・ 正規分布など

8 大数の法則と中心極限定理

- 大数の法則
  試行回数をを多くすると、観測結果が真の値に近づく、という法則。
  → 大標本では、観測された標本平均を真の平均値（母平均）とみなしてよい。

- 中心極限定理
  母集団分布が何であれ、確率変数の和の確率分布は、n が大きくなるにつれて正規分布に近づく。
  つまり、母集団分布の平均、分散（母平均、母分散）を μ, σ^2 とすると、確率変数の和
  $S_n = X_1 + X_2 + \cdots + X_n$ は、$N(n\mu, n\sigma^2)$ に従い、ゆえに、

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \quad \text{は、} \quad N(\mu, \frac{\sigma^2}{n}) \quad \text{に従う}$$

  → n を十分大きくすると、標本平均 $\bar{X}$ は母平均 μ に限りなく近づく。 → 大数の法則

  これを利用して、二項分布を正規分布で近似したり、正規分布にしたがう乱数をつくったりできる。(授業じゃ取り上げられなったようなので蛇足かも)

- コンピュータ・シミュレーション
  コンピュータでシミュレーションすること。
  しばしば乱数が使われる。Excel なら「= RAND()」、C 言語とかなら「srand((unsigned)time(NULL));」などと書けば乱数を発生させられる。

- ベルヌーイ試行
  一回の実験で 2 種類の事象のいずれかが生じ、しかもそのような事象が常に一定の確率で起こるような試行のこと。
  例 : コインを投げて裏表を見る試行[*5]

[*5] たまにコインが立つこともあるけど無視しましょう。

9 標本分布

- 母集団
  自分が持っている標本のデータから知りたいと思う集団全体のこと。
  (例) 日本人の意識調査を行う場合　→　日本人全体が母集団。

- 標本
  母集団から分析のために選び出された要素 ・ 属性値のこと。

- 標本抽出
  母集団から標本を選び出すこと。

- 統計的推論
  母集団について何か知りたくても、現実には不可能なことがある。例えば、
  1. 母集団が非常に多く (無限大の場合もある) の要素からなる場合。
  2. 全体の調査が意味を持たなかったり、予算上の問題から全数の調査が無理な場合。
  3. 将来に起こるため、現在は測定が不可能な要素を含む場合。

  そんなとき、
  1. 母集団からその一部を選び出し（標本抽出をして)、
  2. 標本をを分析して、
  3. 母集団について推測する。

  ということが行われる。これを統計的推論という。

- 母集団分布
  母集団の確率分布。得られた標本は、母集団分布に従う確率変数だと考える。統計的推論の最終目標は、これらの標本をほげほげして母集団分布を求めること。

- 標本の大きさ
  標本の数。同一の母集団分布 $f(x)$ に従う独立な確率変数の数とも。

- パラメトリックの場合
  いくつかの定数さえわかれば、母集団分布についてすべて知ることができる場合。
  つまり母集団分布が、既知の確率分布（正規分布 ・ ポアソン分布 ・ 一様分布などなど）であるとわかっている場合。

- 母数
  パラメトリックの場合の、求めるべき定数（パラメータ）のこと。
  (例)
  1. 正規分布 $N(\mu, \sigma^2)$ における、平均 (期待値)μ と、分散 σ^2。

2. ポアソン分布 P(λ) における λ。

- ノン ・ パラメトリック
  パラメトリックじゃない場合。もとい、いくつかのパラメータだけでは母集団分布を決定できない場合。この場合、平均 ・ メディアン ・ モード ・ 分散 ・ レンジ ・ 歪度 ・ 尖度などを調べて、母集団分布の形状を考えていく。

- 復元抽出と非復元抽出
  母集団から標本を抽出する際、一度抽出した要素を再び母集団に戻すかどうかという話。もとに戻す抽出方法を復元抽出、戻さない方法を非復元抽出という。
  前者と後者では、組合せの数などが微妙に違ってくるため得られる数値も変わる。しかし、取りだす母集団の要素の数 N が標本の大きさ n と比べて十分大きいなら、どっちの方法でもほとんど差がない。したがって、現実には、手間のかからない非復元抽出がよく行われる。

- 単純ランダム ・ サンプリング
  母集団から標本を選び出す方法のひとつ。要素数 N の母集団から、n 個の標本を抽出するとき、母集団の各要素が標本として選ばれる確率が等しく n/N になるように選ぶ方法。乱数がしばしば使われる。
- 母平均
  母集団分布 $f(x)$ の平均。

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{あるいは} \qquad \mu = \sum_{x} x f(x)$$

  とかける。母数の一つ。

- 母分散
  母集団分布の分散。

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \qquad \text{あるいは} \qquad \sigma^2 = \sum_{x} (x-\mu)^2$$

  とかける。母数の一つ。

- 統計量
  標本を要約し、母集団の母数のいろいろな推測に使われる数値のこと。
  (例)　標本の平均、分散、標準偏差、メディアン、最小値、最大値、相関係数などなど。

- 統計分布
  統計量の確率分布。

- 標本平均
  標本の平均　$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$　のこと。
  母平均 μ を知るのが現実的に難しい場合、標本平均を代用する。その心は、標本平均の期待値 $E(\bar{X}) = \mu$

であり、標本の数が大きくなるほど、$\bar{X}$ は μ に確率収縮するから（大数の法則)。統計量の一つ。

- 標本分散
  標本の分散。
  $$s^2 = \frac{1}{n-1}(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - \bar{X})$$
  注意すべきは、分母が n ではなく n-1 であること。その心は、計算すると、s^2 の期待値　$E(s^2) = \sigma^2$ だから。
  母分散 σ^2 の不偏推定量、または不偏分散という。統計量の一つ。

- 偏りのある標本分散
  $$S^2 = \frac{1}{n}(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - \bar{X})$$
  のこと。このとき、期待値は $\mathrm{E}(S^2)$ は、
  $$E(S^2) = \frac{n-1}{n}\sigma^2$$
  となり、実際の母分散の値より少し小さい値が出る（σ^2 の過小評価が起こる)。統計量の一つ。

- 標本和の標本分布
  パラメトリックの場合で、母集団分布が再生性を持つ場合、標本和の確率分布は結構簡単に求まる。
  (例)
  1. 二項母集団母集団分布が母数 (片方の事象が起こる確率)p のベルヌーイ分布なら、標本分布は二項分布 Bi(1,p) に、和 $X_1 + X_2 + \cdots + X_n$ は二項分布 Bi(n,p) に従う。
     (例)　製品に含まれる不良品の数 ・ 社会調査法におけるある事項に関する賛否

  2. ポアソン母集団母集団分布がポアソン分布 $P_0(\lambda)$ のとき、$X_1 + X_2 + \cdots + X_n$ はポアソン分布 $P_0(n\lambda)$ に従う。
     (例)　交通事故死亡者数

  3. 正規母集団母集団分布が正規母集団 $N(\mu, \sigma^2)$ のとき、$X_1 + X_2 + \cdots + X_n$ は正規分布 $N(n\mu, n\sigma^2)$ に従う。($\bar{X}$ は正規分布 $N(\mu, \sigma^2/n)$ に従う)
     (例)　測定誤差

- 漸近的正規性
  標本平均の分布は、n が十分の大きければ正規分布で近似できる（中心極限定理）

- 有限母集団修正以上はすべて母集団の大きさが無限大の無限母集団についての話だったが、母集団の大きさ N があまり大きくないときや、n/N が大きい場合、以上の内容をそのまま適用するのは無理がある。

そこで、母集団の大きさが有限であることを考慮して、修正を行う必要がある。
有限母集団における、標本平均の期待値 ・ 分散をそれぞれ $E(\bar{X})$、$V(\bar{X})$ とすると、

$$E(\bar{X}) = \mu \qquad V(\bar{X}) = C_N \frac{\sigma^2}{n} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

このとき、　$C_N(= \frac{N-n}{N-1})$　を有限母集団修正という。