母集団と標本

1　標本分布

ある母集団（適当な確率分布にしたがうとする）から無作為抽出 (ランダム・サンプリング）によって抽出された標本からつくられ統計量，たとえば平均，分散，標準偏差など（これらは特に標本平均とか，標本分散とよばれる）について，元の母集団分布との関係を調べることがこの章の目的である。

そのために，２つの確率変数を加えるとどういう新しい分布になるか，などいくつかの確率変数から作られる統計量の分布を求めることを考える。このような分布を求めるにはいくつかの方法があるが,

(i) 直接に分布関数の関係式から計算する。下記に記した、和、定数倍、最大・最小、２乗の各分布の公式をつかう。

(ii) 分布の変換（特性関数，確率母関数とよばれる）を利用する。確率密度関数 f_X を変換した、特性関数 $\phi_X(t) = E[\exp(itX)]$, (ただし $i = \sqrt{-1}$ は虚数単位で複素関数論の知識が必要), 逆変換で確率密度関数を求める。確率母関数 $m_X(t) = E[\exp(tX)]$(収束かどうか調べる必要あり) では変数 t の微分で平均 $EX = m'_X(0)$ 、分散 $V[X] = m''_X(0) - (m'_X(0))^2$ が計算できる。

などがある。

以下では確率変数の独立性を仮定し，直接計算で求める方法を公式としてまとめてみます。

(a) ２つの和の分布：f_X, f_Y から f_{X+Y} を求めること。

$$f_{X+Y}(z) = \int f_X(x) f_Y(z-x) dx = \int f_X(z-y) f_Y(y) dy = f_X * f_Y(z)$$

とくに $Y = y$ と定数に退化しているならば、$f_{X+y}(z) = f_X(z-y)$ となる。

(b) 定数倍の分布：X から $Y = cX$, $c > 0$ を求めること。

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right)$$

(c) 最大・最小の分布: X, Y から $Z_M = \max\{X, Y\}, Z_m = \min\{X, Y\}$，を求めること。

$$F_M(z) = F_X(z)\, F_Y(z)$$
$$f_M(z) = \frac{F_M(z)}{dz} = f_X(z)\, F_Y(z) + F_X(z)\, f_Y(z)$$

ここで $F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t) dt$ など。また

$Z_m = \min\{X, Y\}$ について

$$F_m(z) = 1 - (1 - F_X(z))(1 - F_Y(z)) = F_X(z) + F_Y(z) - F_X(z)\, F_Y(z),$$
$$f_m(z) = f_X(z)(1 - F_Y(z)) + (1 - F_X(z)) f_Y(z)$$

これらの関係式を考えると、集合の独立な事象関係；$P(A \cap B) = P(A)\, P(B)$ および $P(A \cup B) = 1 - P(\overline{A \cup B}) = 1 - P(\overline{A} \cap \overline{B}) = 1 - P(\overline{A}) P(\overline{B}) = 1 - (1 - P(A))(1 - P(B)) = P(A) + P(B) - P(A)\, P(B)$ と形が近い。

(d) 2乗の分布：X から $Y = X^2$ を求めること。

$$f_Y(y) = \frac{1}{2\sqrt{y}}\{f_X(\sqrt{y}) + f_X(-\sqrt{y})\}, \quad y > 0$$

もし X が負の値をとらないばあい， $X > 0$ であるならば,

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}), \quad y > 0$$

(e) 商の分布：X, Y から X/Y を求めること。

$$f_{X/Y}(z) = \int_{-\infty}^{\infty} y f_X(yz) f_Y(y) dy$$

なぜなら、(b) より定数に $1/y$ を適用して、 $f_{X/y}(z) = y\, f_X(yz)$ であるから, この右辺の y をランダム化「$Y = y; f_Y(y)$ と重みをつけて平均」すると上式が得られる。

解析学としては高度な計算を必要とするが，正規分布から導かれる分布として，推定や検定には必ずといっていいほど用いられる：

1. 2項分布：$\binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \cdots, n.$ コインを n 枚投げて，さらに m 枚投げる。これらは2項分布である。合計したもの $n + m$ もやはり2項分布である。つまり2項分布について

$$X \sim B(n, p),\, Y \sim B(m, p) \quad \Rightarrow \quad X + Y \sim B(n + m, p)$$

2. 正規分布 $N(\mu, \sigma^2)$: μ は平均、σ^2 は分散。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$$

積率母関数は $M_X(t) = \exp\left(\mu t + \dfrac{\sigma^2 t^2}{2}\right)$, 特性関数は $\phi_X(t) = \exp\left(\mu\, it - \dfrac{\sigma^2\, t^2}{2}\right)$
とくに $N(0, 1)$ を標準正規分布という。

$$X \sim N(\mu_1, \sigma_1^2),\, Y \sim N(\mu_2, \sigma_2^2) \quad \Rightarrow \quad X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

正規分布の標本平均:

$$X_k \sim N(\mu, \sigma^2),\, k = 1, 2, \cdots, n \quad \Rightarrow \quad \overline{X}_n = \frac{1}{n}\sum_i X_i \sim N(\mu, \sigma^2/n)$$

3. カイ 2 乗分布 χ^2_ν：自由度 $\nu = 1, 2, \cdots$

$$f(x; \nu) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, x > 0, \quad f(x; \nu) = 0, x < 0$$

- $X_i \sim N(0, 1), i = 1, 2, \cdots, \nu$ が独立ならば、ν 個の和 $Z = \sum_{i=1}^{\nu} X_i \sim \chi^2_\nu$
- 再生性：$X \sim \chi^2_m, Y \sim \chi^2_n$ が独立ならば、$X + Y \sim \chi^2_{m+n}$

注意；χ(カイ) と x(エックス) の違い。

4. スチューデントの t 分布，t_ν あるいは $t(\nu)$:

$$f(x; \nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} (1 + x^2/\nu)^{-(\nu+1)/2}, -\infty < x < \infty$$

- 独立な同一分布の確率変数に対し、

$$X_i \sim N(\mu, \sigma^2), i = 1, 2, \cdots, n \quad \Rightarrow \quad T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

ただし標本平均 $\overline{X}_n = \dfrac{1}{n}\sum_i X_i$, 標本不偏分散 $S_n^2 = \dfrac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2 = \dfrac{1}{\binom{n}{2}}\sum_{i<j}(X_i - X_j)^2/2$ とする。

- $Z = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ との対応に注意。スチューデントの t 分布になる変数は，正規確

率変数の関数として記述することもできる．$X \sim N(0,1)$ と $Z \sim \chi^2(\nu)$ が独立した確率変数であるとき、$\dfrac{X}{\sqrt{Z/\nu}} \sim t_\nu$ となる。スチューデントの t 分布は縦軸について対称であり，正規変数とその標準偏差に対する割合を特徴付けたものといえる。自由度=1 ならば，t 分布はコーシー分布（平均が存在しない！）と同じである。

5. スチューデントの t 分布: 密度関数は

$$f(x) = \frac{1}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (-\infty < x < \infty)$$

☆標準正規分布をカイ２乗分布の平方根で割ったもの。データの正規化変換（平均０、分散１に変換するもの）した統計量の分布☆

6. フィッシャーの F 分布 (フィッシャー-スネデカー分布 (Fisher-Snedecor distribution)): $F(m,n)$ あるいは F_n^m
F 分布はカイ２乗分布の比率に対する統計量分布, つまり 2 つの独立したカイ２乗分布をそれぞれの自由度で割ったときの比率の分布である。

• $X \sim \chi^2(m), Y \sim \chi^2(n)$ で独立ならば $\dfrac{X/m}{Y/n} \sim F(m,n)$。
これは仮定検定で 2 つの母集団の分散を比較するときに広く使われる．密度関数は

$$f(x) = \frac{m^{m/2}n^{n/2}}{B(m/2,n/2)}\frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}, \quad x > 0$$

ここでベータ関数 $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ とする。

もとの分布が独立同一分布にしたがうばあい，（無作為抽出したとき）これらの標本データから作られる統計量の代表的なものにつぎが知られていて、母数の推定や仮説の検定にはよく用いられる。以上のようには、正規母集団からの標本抽出で表れる分布で推定、検定に必須の分布である。

2　標本平均に関する挙動

標本データを計算式でまとめたものが統計量である。典型的なものが，標本平均や標本分散などであった。母集団からランダム・サンプリングをすると，独立，同一分布である確率変数が得られる。このとき，これら統計量がどのような確率分布に従うかを調べた。和，定数倍，２乗，$\max, \min$ などの演算で新しい分布が求められる。正規分布，スチューデントの $t-$ 分布，カイ２乗分布，フィッシャの F 分布などが代表的な例である。

ここでは標本平均を考えよう。データから得られた値は，データの個数 (大きさという）が増えれば，変動が安定してくると予想される。これを解析していくことにする。

チェビシェフの定理 確率変数 X は 平均 μ, 分散 σ^2 をもつとすれば，任意の $c > 0$ に対し，

$$P(|X-\mu| \geqq c\sigma) \leqq \frac{1}{c^2}$$

が成り立つ。

この不等式は確率変数の値 X と平均とのずれを分散 σ^2 をつかい，評価している。

この定理は一般的な確率変数で成り立つ非常に強力であり，つぎの命題を証明することができる。

大数の弱法則 確率変数列 $X_1, X_2, \cdots$ が互いに独立で同じ分布にしたがいこれらの平均を μ, 分散を σ^2 とすれば，任意の $\epsilon > 0$ で

$$P(|\overline{X_n} - \mu| > \epsilon) \to 0 \quad (n \to \infty)$$

ここで $\overline{X_n} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$。つまりどんなに小さな ϵ をとってもずれの確率がゼロになる，ほとんど起こらない。標本の大きさが大きければ，ずれることが起こりにくくなるということ。平均のまわりに集中をしてくる。標本平均と母集団の平均との関係式である。

中心極限定理（有名なガウスによる）確率変数列 $X_1, X_2, \cdots$ が互いに独立で同じ分布にしたがいこれらの平均を μ, 分散を σ^2 とする（ランダム・サンプリング)。任意の a, b に対し

$$P\left(a \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq b\right) = P\left(a \leq \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \leq b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2}}\, dx \quad (n \to \infty)$$

標本データの大きさが大きくなると，標本平均を標準化したデータの分布は，正規分布に近づくことを主張している。もとのデータが離散型，たとえば，コイン投げの結果であっても，極限は正規分布になる。まさにラプラスが 2 項分布の極限として発見した正規分布が，ガウスによって，どんな分布であっても適当な条件で，正規分布に近づくことが示された。

この極限の関係式は，2 項分布を計算するばあい，正規分布で近似計算できることも表している。もし 2 項分布 $B(n,p)$ のばあいには

$$P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2}}\, dx \quad (n \to \infty)$$

近似を良くするために，つぎの半数補正をおこなう。

X が 2 項分布 $B(n,p)$ のとき，a, b を整数として

$$P(a \leq X \leq b) = P\left(a - \frac{1}{2} \leq X \leq b + \frac{1}{2}\right) = P\left(\overline{a} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \overline{b}\right) = \int_{\overline{a}}^{\overline{b}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2}}\, dx$$

ここで整数値の半分，$1/2$ を補正する：$\overline{a} = \dfrac{a - 1/2 - np}{\sqrt{np(1-p)}}, \quad \overline{b} = \dfrac{b + 1/2 - np}{\sqrt{np(1-p)}}$。