経済統計学講義ノート No.6

標本分布

蛭川雅之

2022 年 10 月 25 日

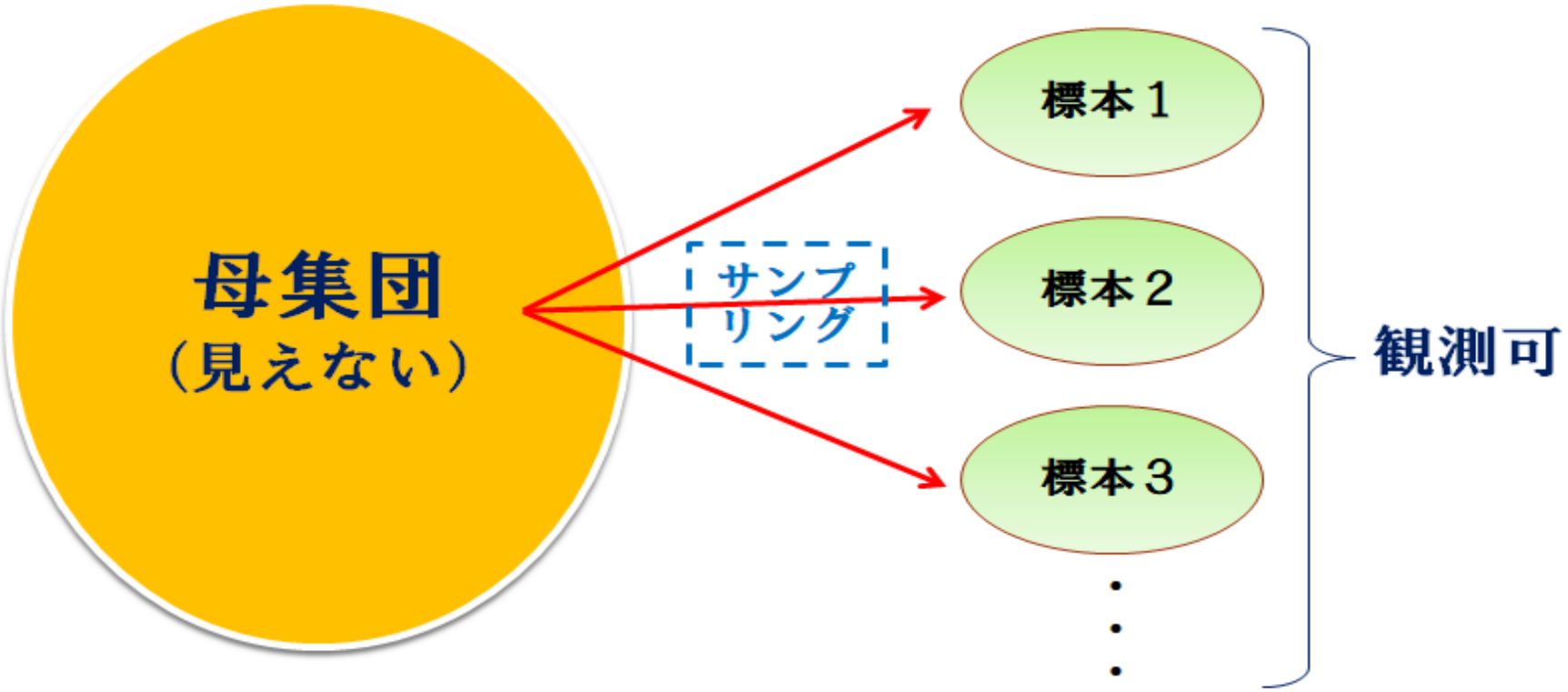
目　次

1　母集団と標本

1.1　記述統計から推測統計へ

- 記述統計では、全てのデータが与えられたものとの前提で、そのデータの整理・加工を行う。
- 推測統計では、このデータを元々属していたより大きな集団（母集団）から選び出された一つの標本とみなす。
  - 標本は母集団に関する何らかの結論を導き出すための手掛かりとなる。
  - 標本は母集団の“良い縮図”でなければならない。
    - 標本調査 = 母集団（鍋の中のカレー）の調査（味見）

■母集団と標本との関係



1.2　幾つかの疑問

1. 母集団の要素は有限か無限か？
   - 有限母集団：袋の中の豆、ある選挙区の有権者数...
     - 推測統計を行う際、有限母集団修正と呼ばれる補正が必要になる。
   - 無限母集団：ある機械から生産される製品全体、企業の日々の売り上げ記録全体、有史以来の胃癌患者全体...
2. （有限母集団に対して）なぜ全数調査（センサス）を行わないのか？
   - 全数調査（例：国勢調査、経済センサス）は費用面での負担が大きく、無回答等の処理も面倒である。
   - 標本調査の例：家計調査、選挙の出口調査、視聴率調査、世論調査、工業製品の抜き取り検査...

3. どのように標本を選び出す（＝サンプリング（抽出）を行う）のか？
   - 有意抽出：調査者が標本を意図的に選び出す。
   - 無作為抽出（ランダム・サンプリング）：母集団を構成する各要素が抽出される可能性を同一にする。
4. 観測値の個数 n のサンプリングを複数回行う場合、それぞれの標本から異なる代表値（＝推定結果）が得られるが、どれを信用すべきか？
   - それぞれの標本は確率変数であるから、ばらつきが生じるのは当然であり、どの推定結果を信用してもよい。
   - 推測統計を行う際、ばらつきも考慮に入れる（⇒ 標本分布）。
   - 無作為抽出による標本は相互に独立である。

結論 1 本講義で扱う推測統計は、無限母集団から無作為抽出された標本を前提とする。

2 標本分布

2.1 標本平均の分布

定理 2 平均 μ と分散 σ^2 を持つ母集団から大きさ n の標本 $X_1, \ldots, X_n$ を無作為に抽出するとき、標本平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

の分布の平均と分散はそれぞれ

$$E\left(\bar{X}\right) = \mu, Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

である。

- 標本平均 $\bar{X}$ は母集団の平均 μ の不偏推定量である。

2.2 大数の法則

定理 3 平均 μ と分散 σ^2 を持つ母集団から大きさ n の標本 $X_1, \ldots, X_n$ を無作為に抽出するとき、観測値の個数 n が大きくなるにつれて、標本平均 $\bar{X}$ が母集団の平均 μ から離れている確率はゼロにいくらでも近づく。

- 定理 3 を書き換えると、任意の $\varepsilon\,(>0)$ に対し、n を十分大きくすると（即ち、$n \to \infty$ のとき）

$$\Pr\left(\left|\bar{X} - \mu\right| > \varepsilon\right) \to 0$$

となる。

  - $n \to \infty$ のとき $Var\left(\bar{X}\right) \to 0$ となり、$\bar{X}$ のばらつきは限りなく小さくなる。
  - その結果、$\bar{X}$ は自身の分布の平均 $E\left(\bar{X}\right) = \mu$（＝母集団の平均）周りにより集中する。

- 定理 3 はしばしば

$$\bar{X} \xrightarrow{p} \mu : [\bar{X}\text{は}\mu\text{に確率収束する}]$$
$$\operatorname{plim} \bar{X} = \mu : [\bar{X}\text{の確率極限は}\mu\text{である}]$$

とも表現される。

  - 標本平均 $\bar{X}$ は母集団の平均 μ の一致推定量である。

2.3 中心極限定理

定理 4 平均 μ と分散 σ^2 を持つ母集団から大きさ n の標本 $X_1, \ldots, X_n$ を無作為に抽出するとき、観測値の個数 n が大きくなるにつれて、標本平均 $\bar{X}$ を標準化したものの確率分布は標準正規分布 $N(0,1)$ にいくらでも近づく。

- 定理 4 を書き換えると、$n \to \infty$ のとき

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1)$$

となる。

  - “$\xrightarrow{d}$”は分布収束を表す数学記号である。

- 定理 4 は、平均 μ と分散 σ^2 を持つ任意の分布から大きさ n の標本を無作為に抽出して得られる標本平均 $\bar{X}$ の分布は、n が十分大きいとき正規分布 $N\left(\mu, \sigma^2/n\right)$ で近似できることを示す。

例 5 母集団に成功確率 p のベルヌーイ分布を仮定し、そこから大きさ n の標本を無作為に抽出する。中心極限定理により、標本平均 $\bar{X}$（即ち、標本における成功比率 $\hat{p}$）について

$$\frac{\bar{X} - E\left(\bar{X}\right)}{\sqrt{Var\left(\bar{X}\right)}} = \frac{\hat{p} - E\left(\hat{p}\right)}{\sqrt{Var\left(\hat{p}\right)}} = \frac{\hat{p} - p}{\sqrt{p\left(1-p\right)/n}} \xrightarrow{d} N\left(0,1\right)$$

であり、$\hat{p}$ の分布は n が十分大きいとき正規分布 $N\left(p, p\left(1-p\right)/n\right)$ で近似できることがわかる。

例 6 例 5 から

$$\frac{n\left\{\bar{X}-E\left(\bar{X}\right)\right\}}{n\sqrt{Var\left(\bar{X}\right)}}=\frac{\sum_{i=1}^{n}X_i-np}{\sqrt{np\left(1-p\right)}}\xrightarrow{d}N\left(0,1\right)$$

であり、$\sum_{i=1}^{n}X_i\sim B\left(n,p\right)$ であることに着目すると、二項分布 $B\left(n,p\right)$ は n が十分大きいとき正規分布 $N\left(np,np\left(1-p\right)\right)$ で近似できることもわかる。

3 正規母集団からの標本分布

3.1 正規分布の再生性

定理 7 独立な正規確率変数 $X_1 \sim N\left(\mu_1, \sigma_1^2\right), \ldots, X_n \sim N\left(\mu_n, \sigma_n^2\right)$ に対し、

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

が成り立つ。

- このように、同一の分布に属する独立な確率変数の和の分布が元の分布と同型になる性質を再生性という。

- 再生性を持つ分布の例：
  - （成功確率が同一の）二項分布
  - ポアソン分布
  - 正規分布
  - （尺度母数が同一の）ガンマ分布（例：χ^2 分布）...

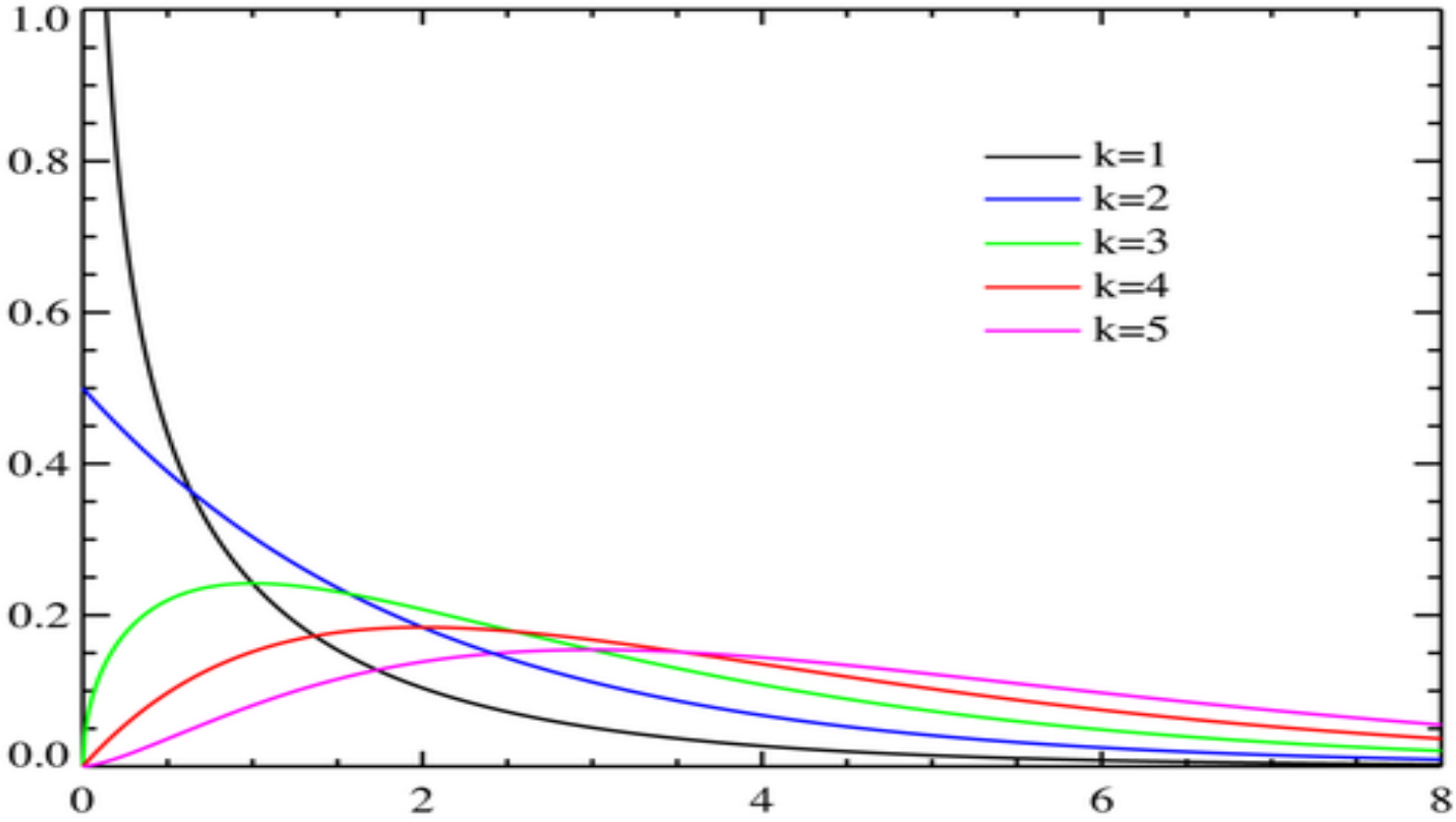
補題 8 正規母集団 $N\left(\mu, \sigma^2\right)$ から大きさ n の標本 $X_1, \ldots, X_n$ を無作為に抽出するとき、標本平均 $\bar{X}$ の確率分布は正規分布 $N\left(\mu, \sigma^2/n\right)$ である。

3.2 χ^2 分布

定義 9 標準正規分布 $N(0,1)$ に従う相互に独立な n 個の確率変数の 2 乗和の分布を自由度 n の χ^2 分布という。

- 自由度 n の χ^2 分布を $\chi^2(n)$ などと表記する。
- $\chi^2(n)$ に従う確率変数の期待値は n、分散は $2n$ である。

■χ^2 分布の確率密度関数の形状



■正規母集団からの標本の 2 乗和の分布

- 正規母集団 $N\left(\mu, \sigma^2\right)$ から大きさ n の標本 $X_1, \ldots, X_n$ を無作為に抽出するとき、それらを標準化したものの 2 乗和 W は自由度 n の χ^2 分布 $\chi^2(n)$ に従う。

$$W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

- W における母集団の平均 μ を標本平均 $\bar{X}$ に置き換えた 2 乗和 V は自由度 $(n-1)$ の χ^2 分布 $\chi^2(n-1)$ に従う。

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

■自由度とは？

- 変数が自由に動き回ることのできる空間の次元を指す。
- V を計算する際、n 個の変数 $X_1, \ldots, X_n$ の間に関係式

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

が存在するため、自由度が観測値の個数 n から 1 下がる。

3.3 t 分布

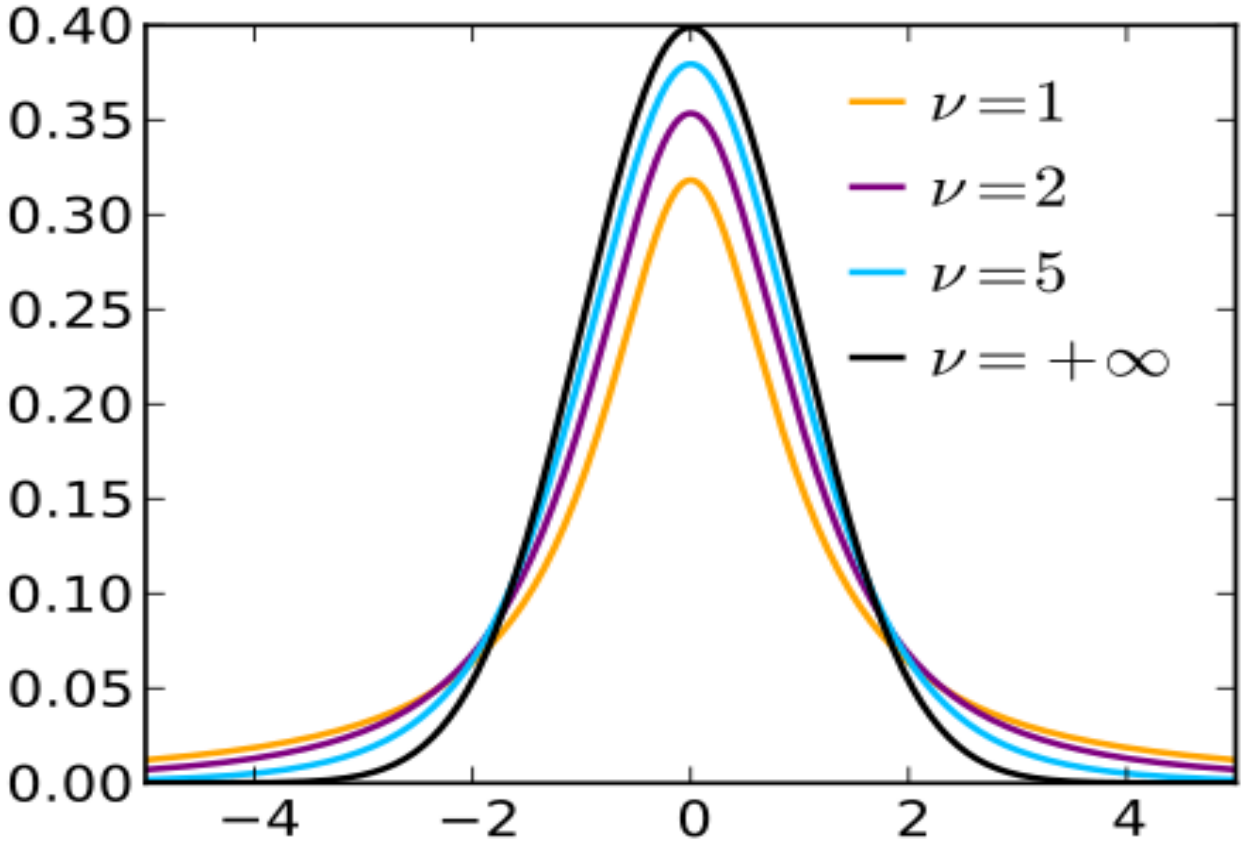
定義 10 標準正規分布 $N(0,1)$ に従う確率変数 Z、およびこれと独立に自由度 n の χ^2 分布 $\chi^2(n)$ に従う確率変数 W が与えられた場合の

$$T = \frac{Z}{\sqrt{W/n}}$$

の分布を自由度 n の t 分布という。

- 自由度 n の t 分布を $t(n)$ などと表記する。
- $t(n)$ に従う確率変数の期待値は 0（ただし、$n \geq 2$）、分散は $n/(n-2)$（ただし、$n \geq 3$）である。

■t 分布の確率密度関数の形状



- $t\,(1)$ はコーシー分布とも呼ばれる。
- $t\,(\infty)$ は標準正規分布 $N\,(0,1)$ である。

■正規母集団からの標本と t 分布との関連

- 正規母集団 $N\left(\mu, \sigma^2\right)$ から無作為に抽出した大きさ n の標本 $X_1, \ldots, X_n$ から得られた標本平均 $\bar{X}$ および標本分散

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

の分布を考える。

- まず、標本平均 $\bar{X}$ について、

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N\left(0, 1\right)$$

が成り立つ。

- 次に、標本分散 $\hat{\sigma}^2$ に対し、

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2 (n-1)$$

である。

- Z と V は独立であることが知られており（証明略）、結果として

$$T = \frac{Z}{\sqrt{V/(n-1)}}$$

は自由度 $(n-1)$ の t 分布 $t(n-1)$ に従うことがわかる。

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\sqrt{n-1}\left(\bar{X} - \mu\right)}{\hat{\sigma}} \sim t(n-1)$$

3.4 F 分布

定義 11 自由度 n_1 の χ^2 分布 $\chi^2(n_1)$ に従う確率変数 W_1 と、これと独立に自由度 n_2 の χ^2 分布 $\chi^2(n_2)$ に従う確率変数 W_2 とが与えられたとき、これらの比

$$F = \frac{W_1/n_1}{W_2/n_2}$$

の分布を自由度 (n_1, n_2) の F 分布という。

- 自由度 (n_1, n_2) の F 分布を $F(n_1, n_2)$ などと表記する。
- 確率変数 $T \sim t(n)$ であるとき、$T^2 \sim F(1, n)$ である。

■F 分布の確率密度関数の形状