2004 年 7 月 6 日

総合科目　数理・自然 6「統計学入門」講義レジュメ　(今野) —　標本分布

用語：母集団と標本

- 母集団 — 知りたいと思う集団の全体（その集団の属性値）
- 母集団分布 — 母集団の属性値が従う分布 (モデル化！)
- 標本 — 母集団から分析のために選び出される要素
- 統計的推測 — 母集団からその一部を選び出す；それをもとに分析する；母集団の分布について推測する
  - 選挙の出口調査では，実際に調べた人の投票行為の集計が目的ではなく，全投票者での選挙結果の予測が目的である．
  - ある病気の新薬開発の臨床試験では，試験に参加した患者への投与結果からその病気の患者全体での効果を推測するのが目的である．
  - ある製造工程から作られた機械部品のうちの何個かを検査して不良品率を調べるとき，真の興味は製造工程そのものにおける不良品率である．
- ランダムネスを利用する
  - 標本抽出 — 標本を母集団から選び出すこと
  - 被験者集団をランダムにわける．
- 標本抽出法
  - 単純ランダムサンプリング (単純無作為抽出)（注：非復元抽出・復元抽出）— 母集団の固体が標本として選ばれる確率が同じである．乱数表を利用.
  - 層別抽出法—母集団を階層ないでは均一とみられる階層にわけ，各階層から標本を取り出す方法.
    - 層別無作為抽出法—特に各層から標本を無作為抽出する方法．また，各層からの標本数は階層の大きさに比例する比例抽出法がよく利用される．
  - ２段抽出法—母集団をいくつかの階層にわけ．まずこれら階層のいくつかをランダムに抽出し，つぎに選ばれた階層からランダムに標本をとる方法.
- 標本 $X_1, X_2, \ldots, X_n$ はこの母集団分布に従う確率変数とする．n のことを標本の大きさという．

母数と統計量

- 母平均と母分散 — 母集団分布の平均と分散 (母集団分布を特定する代表的な母数)
- 統計量 — 標本の情報 (実際にわかる情報) を要約し，母集団分布の推測に利用するもの．標本分布 — 統計量の分布
- 統計量の例 — 標本平均，標本分散，標本メデアン，最小値，最大値など

例 1

値	3	6	9	12	15
確率	1 / 5	1 / 5	1 / 5	1 / 5	1 / 5

(*)

標本	標本の値	$\bar{x}$（標本平均)	m(標本メデアン)
1	$3, 6, 9$	6	6
2	$3, 6, 12$	7	6
3	$3, 6, 15$	8	6
4	$3, 9, 12$	8	9
5	$3, 9, 15$	9	9
6	$3, 12, 15$	10	12
7	$6, 9, 12$	9	9
8	$6, 9, 15$	10	9
9	$6, 12, 15$	11	12
10	$9, 12, 15$	12	12

$\bar{x}$ の値	6	7	8	9	10	11	12
確率	0.1	0.1	0.2	0.2	0.2	0.1	0.1

m の値	6	9	12
確率	0.3	0.4	0.3

宿題 1

(1) (*) の分布の平均と分散を求めよ．
(2) $\bar{x}$ と m の平均を求めよ．

例 2 K 内閣を支持する人には数字の 1，そうでない人には数字の 0 を対応させると，母集団は数字の 1 と 0 の集まり（有権者分）となる．このうち，数字の 1 の割合 p が K 内閣の支持率であり，この p に興味がある．

この場合，母集団分布は離散分布であたって，その確率関数 $f(x)$ は

$$f(x) = p^x(1-p)^{(1-x)}, \qquad x = 0,\ 1,$$

である．するとこの母集団分布の母平均は p，母分散は $p(1-p)$ となる．

いま，n 個の標本を無作為抽出する．この標本を $X_1, X_2, \ldots, X_n$ とする．たとえば，標本から計算される平均（標本平均）$\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ は統計量である．標本平均の分布（標本平均の標本分布）は

$$g(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \qquad y = 0, 1, \ldots, n$$

となる．すなわち，２項分布である．

2004 年 7 月 9 日

総合科目　数理・自然 6「統計学入門」講義レジュメ　(今野) — 9 章　標本分布の続き

標本平均と標本分散

- 代表的な母数
  - 母平均　—　$\mu = \begin{cases} \int xf(x)\,dx & \text{(連続型)} \\ \sum xf(x) & \text{(離散型)} \end{cases}$
  - 母分散　—　$\sigma^2 = \begin{cases} \int (x-\mu)^2 f(x)\,dx & \text{(連続型)} \\ \sum (x-\mu)^2 f(x) & \text{(離散型)} \end{cases}$
- 統計量
  - 標本平均　—　$\bar{X}_n = (1/n)(X_1 + X_2 + \cdots + X_n)$
  - 標本平均の性質
    - $\mathbb{E}[\bar{X}_2] = \frac{1}{n}\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \frac{1}{n}\{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]\} = \mu$
    - $\mathbb{VAR}[\bar{X}_n] = \frac{1}{n^2}\mathbb{VAR}[X_1 + X_2 + \cdots + X_n] = \frac{1}{n^2}\{\mathbb{VAR}[X_1] + \mathbb{VAR}[X_2] + \cdots + \mathbb{VAR}[X_n]\} = \frac{\sigma^2}{n}$
    - n が大きくなれば，$\bar{X}_n \to \mu$ となる．
  - 不偏分散　—　$s^2 = (1/(n-1))\sum (X_i - \bar{X}_n)^2$
  - 不偏分散の性質
    - $\mathbb{E}[s^2] = \sigma^2$
    - n が大きくなれば，$s^2 \to \sigma^2$
- 例 (p. 185)　17 個の標本の値は

$$16,\ 22,\ 21,\ 20,\ 23,\ 21,\ 19,\ 15,\ 13,\ 23,\ 17,\ 20,\ 29,\ 18,\ 22,\ 16,\ 25$$

であった．このデータに基づく標本平均を計算すれば、

$$\begin{aligned} \bar{X}_{17} &= \frac{16 + 22 + \cdots + 16 + 25}{17} \\ &= \frac{340}{17} = 20 \end{aligned}$$

また，不偏分散は

$$\begin{aligned} s^2 &= \frac{(16-20)^2+(22-20)^2+\cdots+(16-20)^2+(25-20)^2}{17-1} \\ &= \frac{(-4)^2+2^2+\cdots+(-4)^2+5^2}{16} = \frac{254}{16} = 15.875 \end{aligned}$$

したがって、標本標準偏差は $s = 3.984344$ となる．

順番にならべかえると

$$13,\ 15,\ 16,\ 16,\ 17,\ 18,\ 19,\ 20,\ \boxed{20},\ 21,\ 21,\ 22,\ 22,\ 23,\ 23,\ 25,\ 29$$

となるので、標本中央値は 20 となる．よって，峰に関して対称なデータと予想される．

実際には，この母集団分布の母平均や母分散はわからないので、標本平均（標本中央値）や不偏分散で代用することになる．

最後に（平均）$\pm r$(標本標準偏差) の間にどの程度の割合でデータが含まれるかを見てみよう．ただし，$r = 1, 2$

$$(20-s,\ 20+s) = (16.01566,\ 23.98434)$$

では

$$10/17 = 0.5882353$$

となる．また，

$$(20-2s,\ 20+2s) = (12.03131,\ 27.96869)$$

では

$$16/17 = 0.9411765$$

となる．

標本平均の標本分布

- ベルヌーイ分布 — $X_i \sim Bi(1, p)$ ならば $n\bar{X}_n$ は二項分布
- ポアソン分布 — $X_i \sim Poisson(\lambda)$ ならば $n\bar{X}_n \sim Poisson(n\lambda)$
- 正規分布 — $X_i \sim N(\mu, \sigma^2)$ ならば $\bar{X}_n \sim N(\mu, \sigma^2/n)$
- 漸近的に標本分布を調べる — 母集団分布に関係なく，

$$P\left(a \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le b\right) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

- 例　みかんの選果作業をするときに，$p = 0.05$ の割合で不良品に分類される．このとき，$10,000$ 個のみかんを選果したときにどの程度のみかんが不良品となるか（出荷できるか）を予想したい．$i\,(i = 1, 2, \ldots, 10,000)$ 番目のみかんに対応する確率変数を

$$X_i = \begin{cases} 1 & \text{不良} \\ 0 & \text{不良でない} \end{cases}$$

とする．すると各 X_i は独立同一に $Bi(1, 0.05)$ に従う：すなわち，

$$P(X_i = 1) = 0.05, \qquad P(X_i = 0) = 0.95$$

となる．さらに，

$$X_1 + X_2 + \cdots + X_{10,000}$$

は二項分布 $Bi(10,000, 500)$ となる．しかし，この確率を具体的に計算するのは困難である（なぜだろうか？）．そこで，正規分布による近似をする．まず，

$$\begin{aligned} \mu &= \mathbb{E}[X_1] = 0.05, \\ \sigma &= \sqrt{\mathbb{VAR}[X_1]} = \sqrt{0.05 \times 0.95} = \sqrt{0.0475} = 0.22\,(0.2179449) \end{aligned}$$

となることに注意する．中心極限定理を使えば，

$$\frac{\bar{X}_{10,000} - 0.05}{0.22/\sqrt{10,000}}$$

は標準正規分布で近似できるので，ほとんど確実に

$$\frac{\bar{X}_{10,000} - 0.05}{0.22/\sqrt{10,000}} \leq 3$$

となる．これをとけば，

$$\bar{X}_{10,000} - 0.05 \leq 0.0022 \times 3 \Longleftrightarrow \sum_{i=1}^{10,000} X_i - 500 \leq 220 \times 3$$

$$\Longleftrightarrow \sum_{i=1}^{10,000} X_i \leq 500 + 3 \times 220 = 1160$$

の不良品を覚悟すれば十分である．