統計学ノートNo. 6
標本分布

担当教員：蛭川雅之
研究室：紫英館3階345研究室
メール： hirukawa@econ.ryukoku.ac.jp
オフィスアワー：毎週金曜日3講時

9/20/2021

1　母集団と標本

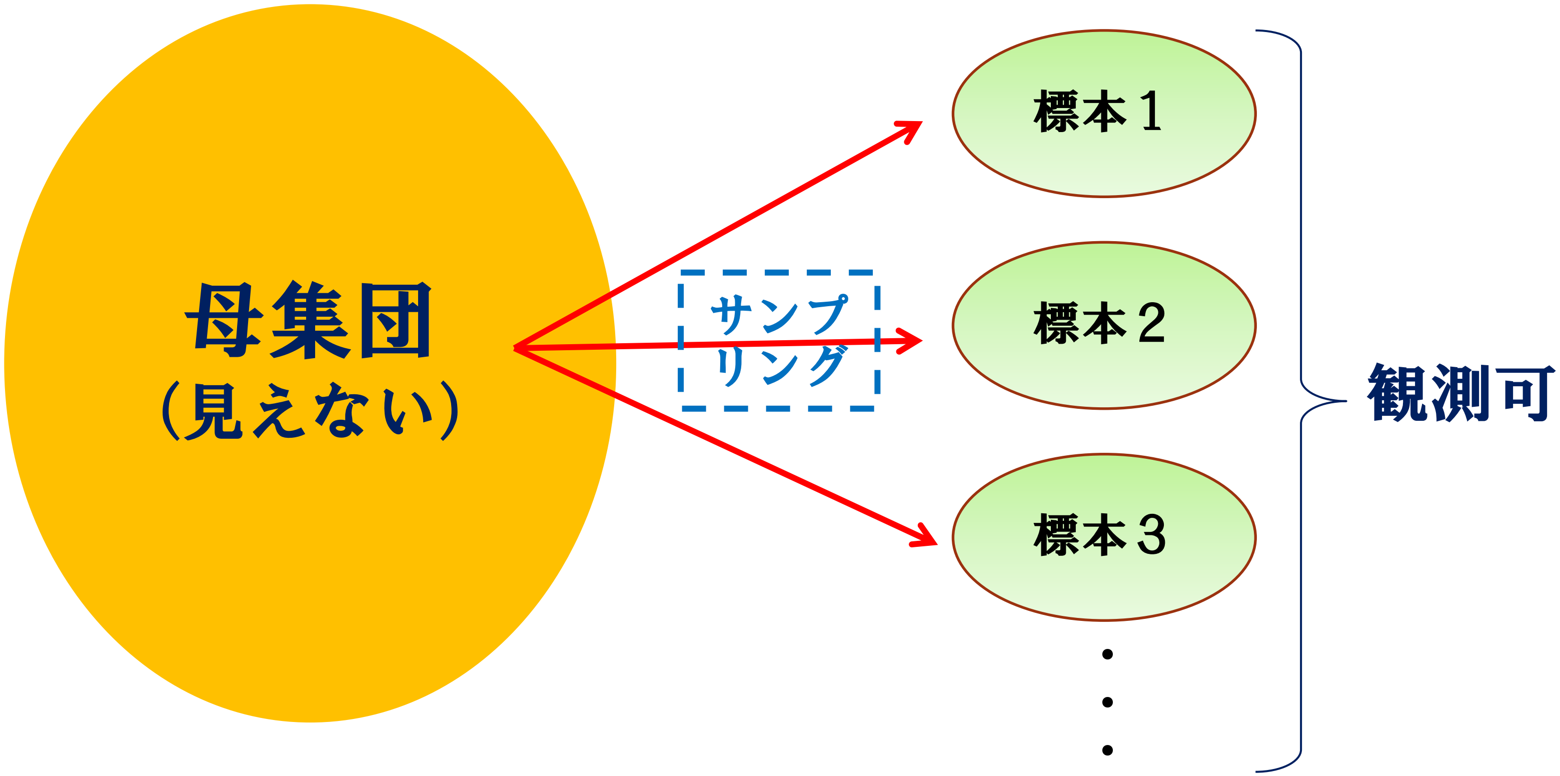
記述統計から推測統計へ

1. **記述統計:**
   - 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。
2. **推測統計:**
   - 全数調査（センサス）はコストがかかる。
   - 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。

母集団と標本

- **観測するデータの背後には膨大なデータがある。**
  - **膨大なデータセットの一部を観測しているにすぎない。**
- **いくつかのデータを観測することで、背後にある全データに関して何か推測することはできないか？**
  - **「部分から全体へ」：推測統計の考え方**

母集団と標本（つづき）



例：全投票者の投票結果　　　出口調査の結果

統計的推論のイメージ

- **部分から全体への推論（例）：**
  - **味噌汁やスープの味をスプーン１杯分の味で判断する。**
  - **よく混ざっている限り、スプーン１杯分の味を確認すれば全体像はわかる。**
- **統計的推論も、部分が母集団全体の様子を反映していることを前提とする。**
  - **多少のズレは覚悟する。**

母集団に関する疑問１

- **母集団のデータ数は有限か無限か？**
  - **有限母集団：有権者数、袋の中の豆…**
    - **利用にあたり、有限母集団修正と呼ばれる補正が必要。**
  - **無限母集団：蝶の体長、機械から生産される製品全体、企業の日々の売上、胃がん患者数…**

⇒　**無限母集団を取り扱う。**

母集団に関する疑問２

- **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。**

⇒　**無作為抽出を取り扱う。**

２　標本平均の考え方

1個のデータから母平均を推測する

- **母集団の平均（母平均）をμ、分散（母分散）をσ^2と表記する。**

- **母分散σ^2を知っているという前提で、1個のデータXを使って母平均μを推測したい。**
  - **Xが正規母集団からの標本である場合、μは**

$$X - 2\sigma \leq \mu \leq X + 2\sigma$$

の範囲に存在する可能性が極めて高い。

なぜ標本平均をとるのか？

- **母分散σ^2を知っているという前提で母平均μを推測する問題を再度考える。**
  - **今回は同じ母集団からの標本が（1個でなく）n個あるとする。**
- **標本n個の平均（標本平均）をとると、母平均μに近い値が出るのではないか？**
  - **直感的には、平均をとることで偶然によるデータの散らばりを消す効果が期待できる。**

サイコロ投げ：1個 vs. 2個

- 偏りのない正六面体のサイコロ1個を無限回投げて出る目を記録し、これを母集団とする。
  - 1~6それぞれの目の出る確率は$\frac{1}{6}$であるから、母平均μは

$$\mu = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3$$

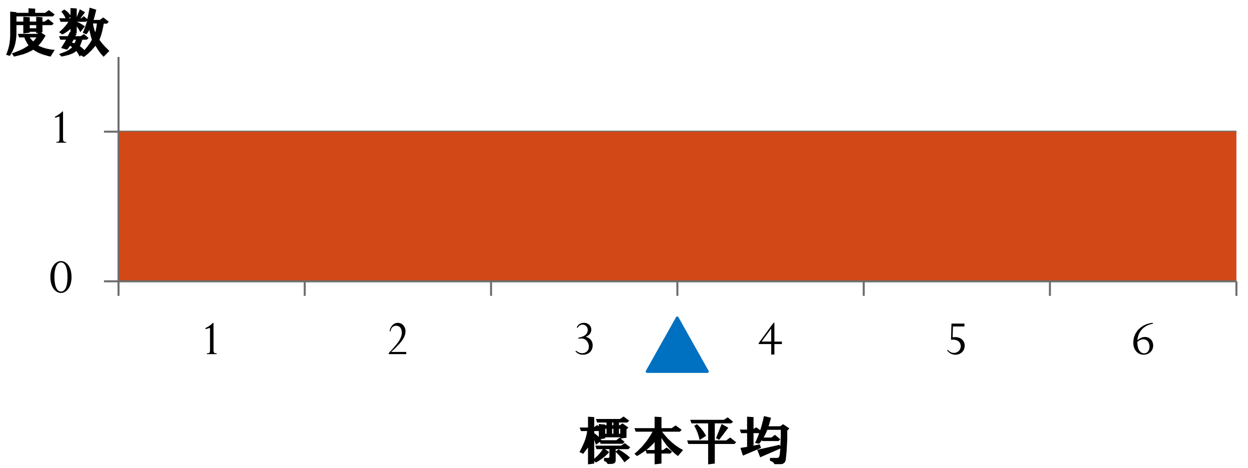
$$+ \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = \frac{7}{2} = 3.5$$

である。

サイコロ投げ：1個 vs. 2個（つづき）

- **まず、サイコロ1個を投げて、出る目（＝標本平均）を記録する。**

D1	1	2	3	4	5	6
標本平均	1	2	3	4	5	6

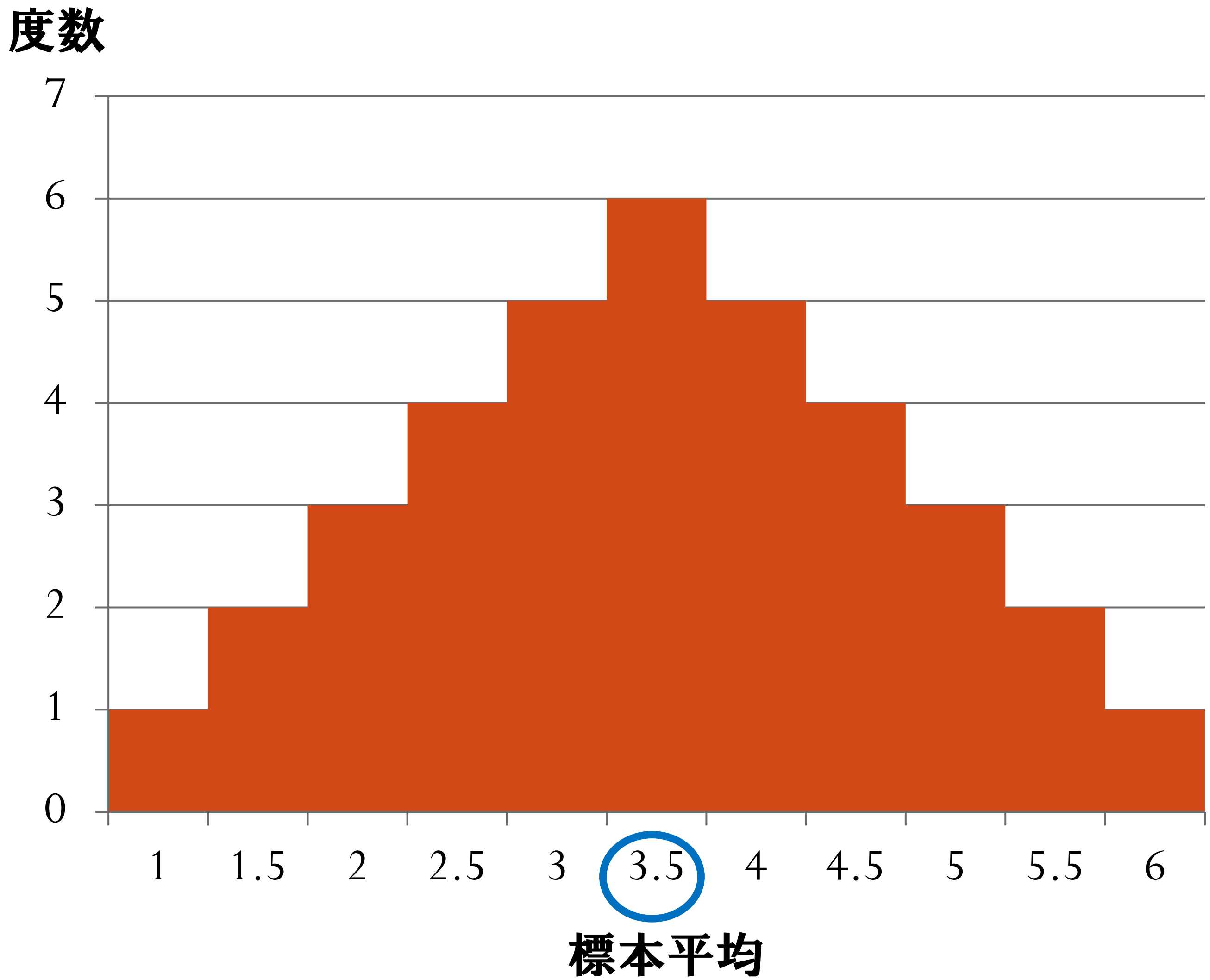


サイコロ投げ：1個 vs. 2個（つづき）

- 今度は、サイコロを2個同時に投げて、出た目の平均（＝標本平均）をとる。

D1＼D2	1	2	3	4	5	6
1	1	1.5	2	2.5	3	3.5
2	1.5	2	2.5	3	3.5	4
3	2	2.5	3	3.5	4	4.5
4	2.5	3	3.5	4	4.5	5
5	3	3.5	4	4.5	5	5.5
6	3.5	4	4.5	5	5.5	6

サイコロ投げ：1個 vs. 2個（つづき）



大数の法則

- 同一の母集団からn個のデータ

$$X_1, X_2, \ldots, X_n$$

を取り出し、その標本平均

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

をとる。nが大きければ大きいほど、標本平均は母平均μに近い値をとる可能性が高くなる。

3　正規分布からの標本平均

サイコロ投げのヒストグラムからわかること

- 2個のデータを観測する場合の標本平均の分布は、1個のデータを観測する場合の標本平均の分布に比べて、より母平均$\mu = 3.5$の周辺に集中している。
- 2個のデータを観測する場合の標本平均の分布は母集団の分布と明らかに異なる。
  - 標本分布の形状がnと共に変化するため、母平均を「間違うリスク」を一定水準に保つのは容易でない。

正規分布の再生性

- **独立な正規確率変数**

$$X_1 \sim N(\mu_1, \sigma_1^2), \ldots, X_n \sim N(\mu_n, \sigma_n^2)$$

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。**

正規母集団からの標本平均の分布

- **正規分布の再生性から、平均μと分散σ^2の正規母集団$N(\mu,\sigma^2)$から取り出したn個のデータの標本平均$\bar{X}$の分布は再び正規分布となることがわかる。**
- **具体的には、標本平均$\bar{X}$の分布は平均μと分散$\frac{\sigma^2}{n}$の正規分布**

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

となる。

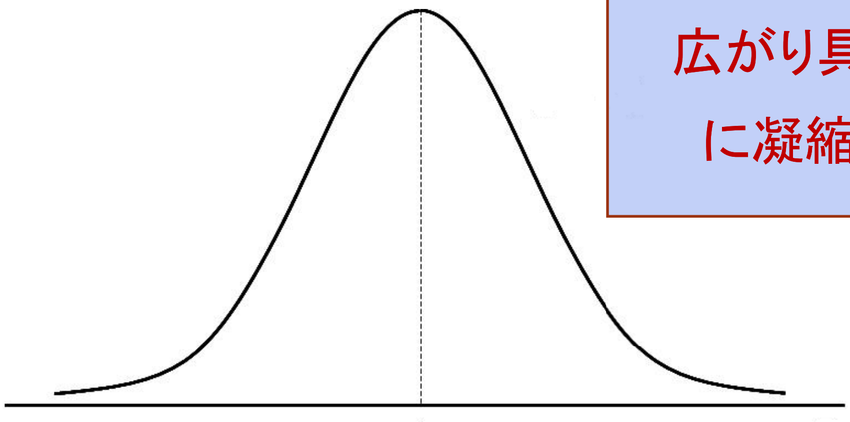
正規母集団からの標本平均の分布（つづき）

母集団の分布

$N(\mu, \sigma^2)$

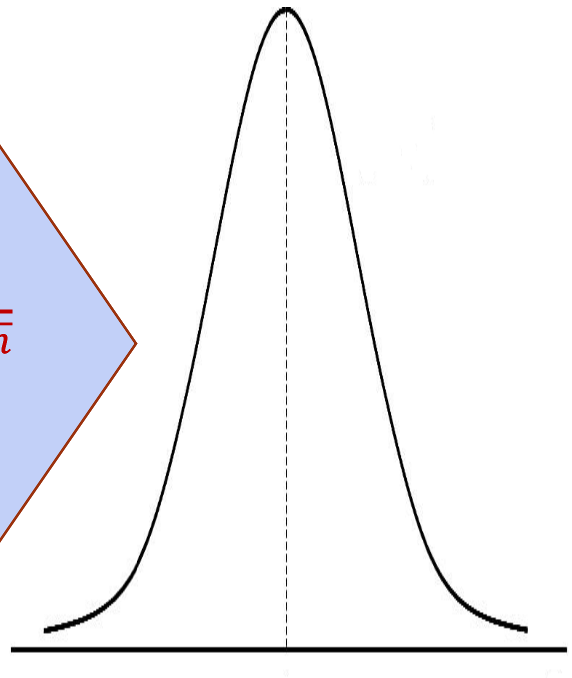
標本平均の分布

$N\left(\mu, \dfrac{\sigma^2}{n}\right)$



μ

広がり具合が$\frac{1}{\sqrt{n}}$に凝縮される



μ