

統計学 01

早稲田大学政治経済学部

第7回

西郷 浩

本日の目標

- 母集団と標本
 - 全数調査 vs 標本調査
 - 無作為標本抽出
- 母数 parameter と統計量 statistics
- 統計量の標本分布

母集団と標本(1)

■ 母集団:

□ 調査対象全体:

- 例:有権者全体

■ 標本:

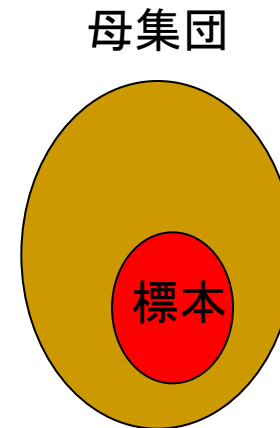
□ 母集団の一部

- 例:有権者のうちの3000人

■ 注:

□ 有限母集団:有権者全体

□ 無限母集団:サイコロをふる試行の全体



母集団と標本(2)

■ 全数調査と標本調査

- 全数調査：母集団の全体を調査する。
 - 例：国勢調査、経済センサス
- 標本調査：母集団の一部（標本）を調査する。
 - 例：家計調査、労働力調査、商業動態調査
- 余談：
 - わが国の調査体系
 - 世帯統計：国勢調査で確定した調査区・住戸をつかって標本調査をおこなうのが標準的。
 - 事業所・企業統計：経済センサスで確定した調査区・事業所・企業をつかって標本調査をおこなうのが標準的。

母集団と標本(3)

■ 全数調査の長短

□ 長所:

- 標本誤差(全体の一部しか調べないために生じる誤差)なし。
- 標本調査のための名簿(フレーム)整備

□ 短所:

- 費用・時間がかかる。
- 詳細な調査は不可能。
- 非標本誤差(標本誤差以外の誤差)が大きくなる。
- そもそも全数調査が実際上不可能なものがある。
 - 例: 缶詰の抜き取り検査

母集団と標本(4)

■ 標本調査の長短

□ 長所

- 費用・時間が節約できる。
- 詳細な調査が可能である。
- 非標本誤差が(相対的に)小さい。

□ 短所

- 標本誤差が必ず発生する。
 - ただし、その理論的な扱いは、非標本誤差よりもずっと易しい。
- 良質な母集団名簿(フレーム)が必要である。

母集団と標本(5)

■ 確率標本抽出

□ 標本抽出 sampling: 母集団から標本を選ぶこと

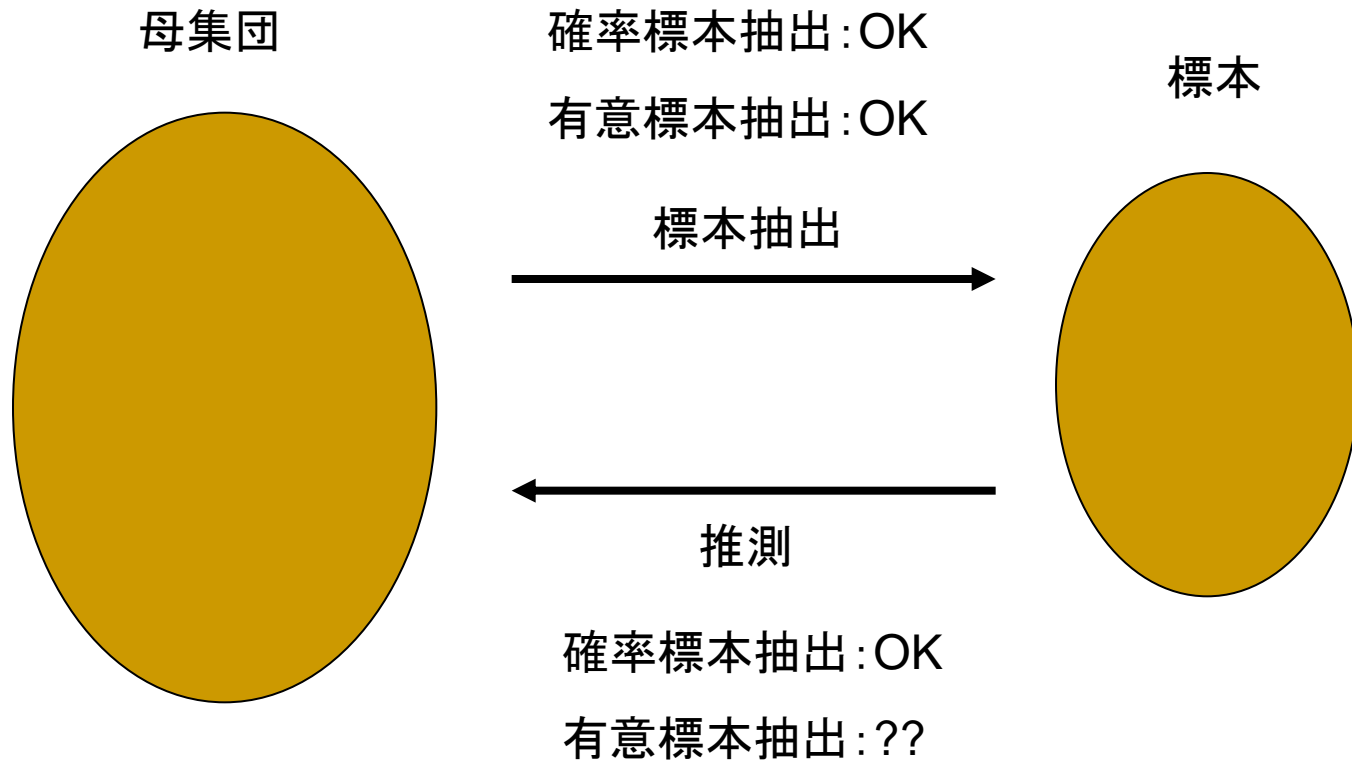
■ 確率標本抽出 probability sampling

- ある特定の標本が抽出される確率があらかじめコントロールされているような抽出方法
- 代表例: 無作為標本抽出(後述)

■ 有意標本抽出 purposive sampling

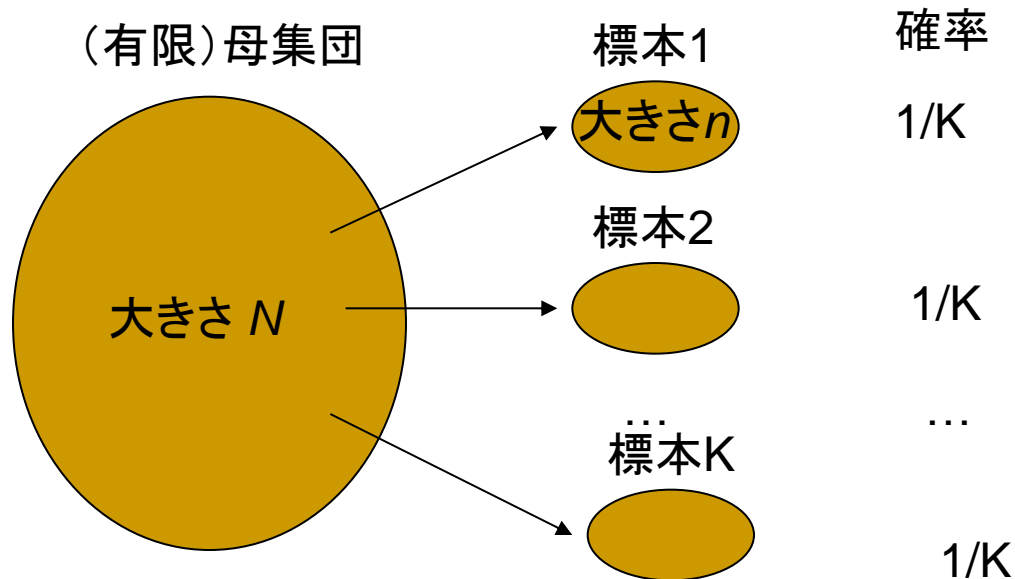
- ある特定の標本が抽出される確率があらかじめコントロールされていないような抽出方法
- 代表例: 街頭インタビュー調査
- この場合の「有意」は「意図がある; 無作為でない」という意味。

母集団と標本(6)



母集団と標本(7)

- 無作為標本抽出(確率標本抽出の一種)
 - どの標本も出現する確率が等しい標本抽出



- 無限母集団のとき:「独立な試行の結果の集まり」と考える。

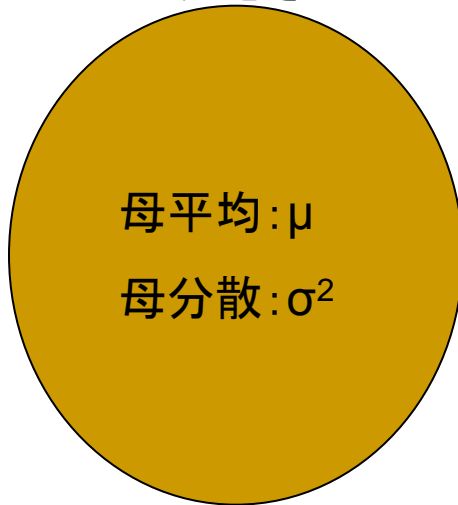
母数と統計量(1)

- 母数 parameter: 母集団の分布の特徴をあらわす値
 - 母集団情報が既知であるときにのみ計算可能
 - 母平均: 母集団における平均(ないし期待値)
 - 例: 2012年4月1日現在早大在籍者を母集団としたときの在籍者全員の平均身長
 - 母分散: 母集団における分散
- 統計量 statistics:
 - 標本情報(既知)から計算できる値
 - 標本平均
 - 例: 標本抽出された200人の早大在籍者の平均身長
 - 標本分散
 - 例: 標本抽出された200人の早大在籍者の身長分散

母数と統計量(2)

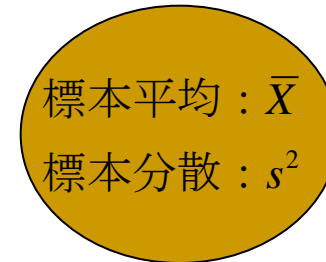
(有限)母集団

大きさ N



標本

大きさ n



$$\mu = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$

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

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

母数と統計量(3)

- 母数≠統計量(標本誤差の存在)

$$\mu \neq \bar{X}$$

$$\sigma^2 \neq s^2$$

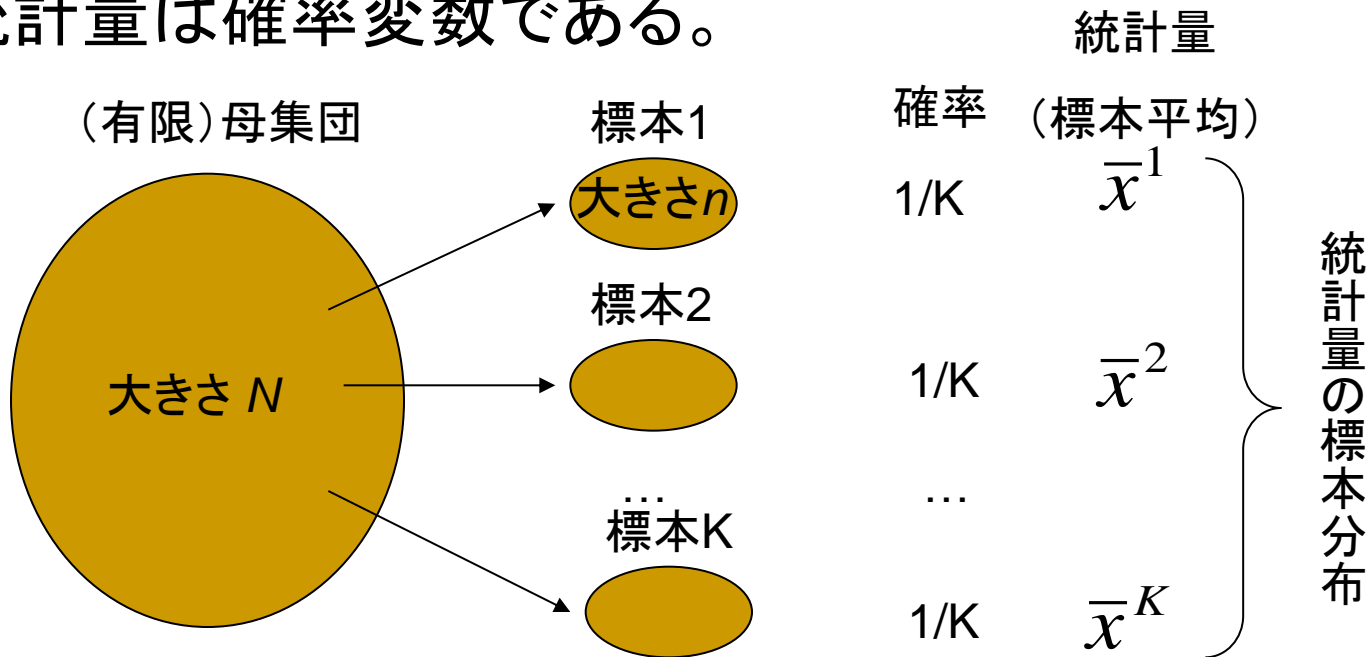
- この乖離をコントロールする方法はあるか？

- Yes !

- 「標本抽出が、あらかじめコントロールされた確率にもとづいておこなわれる」という事実を生かす。
- つまり、「統計量は、標本抽出にともなって確率的に変化する確率変数である」という点に着目する。

統計量の標本分布(1)

- 無作為標本抽出と統計量
 - 統計量は確率変数である。



統計量の標本分布(2)

- 統計量の標本分布 sampling distribution
 - 標本抽出にともなう、統計量の確率分布
 - (標本抽出を何度も何度も繰り返したら観察できる)統計量の確率的な挙動。
 - 1回の標本抽出の結果(統計量のひとつの実現値)のみを見ていては、「統計量と母数とは等しくならない」ということしかわからない。
 - しかし、統計量の標本分布(=確率変数としての統計量の確率的挙動)と母数とであれば関連付けられる。

統計量の標本分布(3)

■ 「確率変数としての標本平均」と母平均との関係

- 無作為標本抽出のもとで、

$$E(\bar{X}) = \mu$$

$$V(\bar{X}) = \sigma^2/n$$

- 母集団の分布型が正規分布であれば、

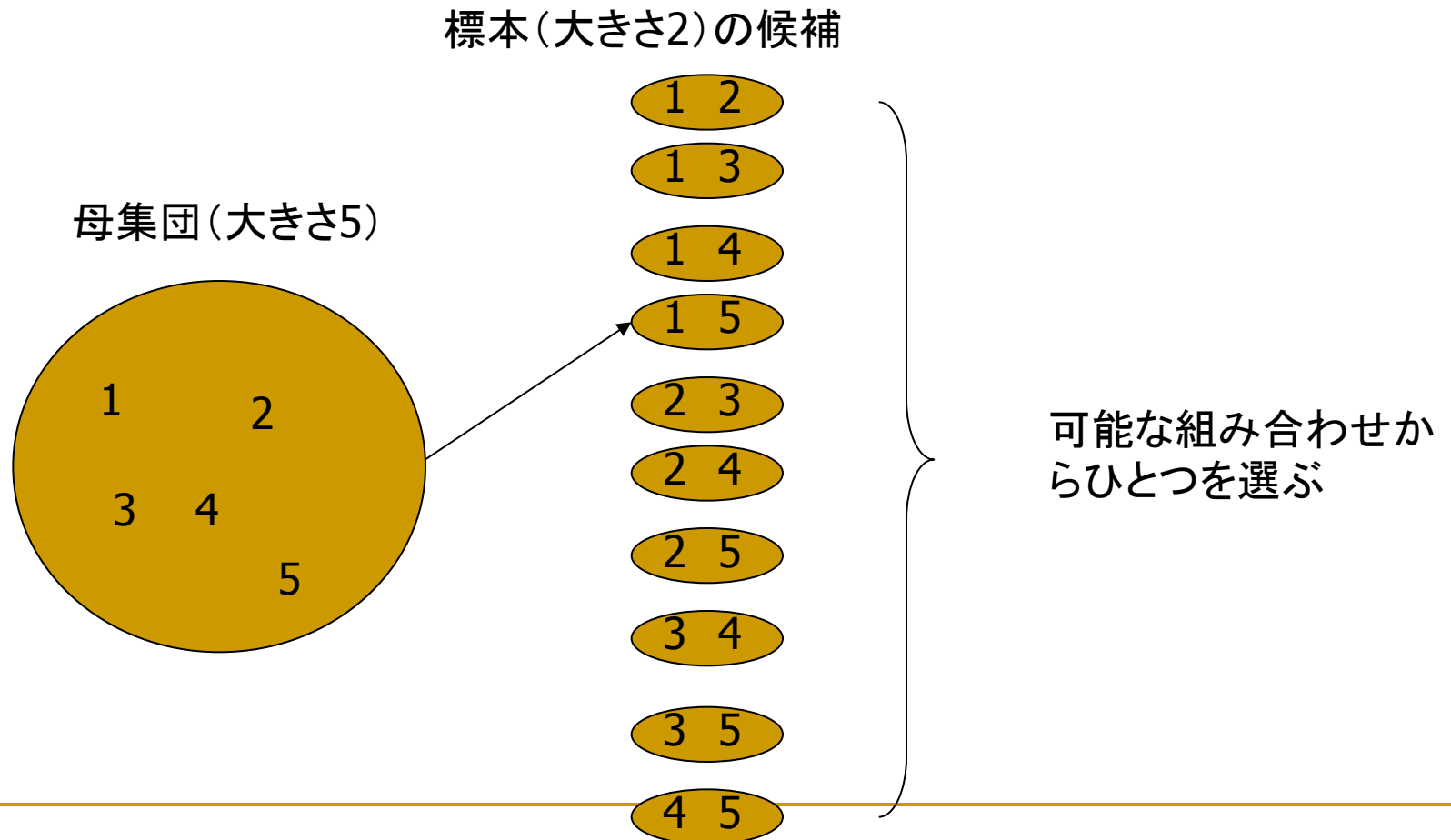
- $\bar{X} \sim N(\mu, \sigma^2/n)$

- さらに、標本の大きさ n が十分に大きければ、母集団の分布の型にかかわらず(中心極限定理)

- $\bar{X} \xrightarrow{D} N(\mu, \sigma^2/n)$

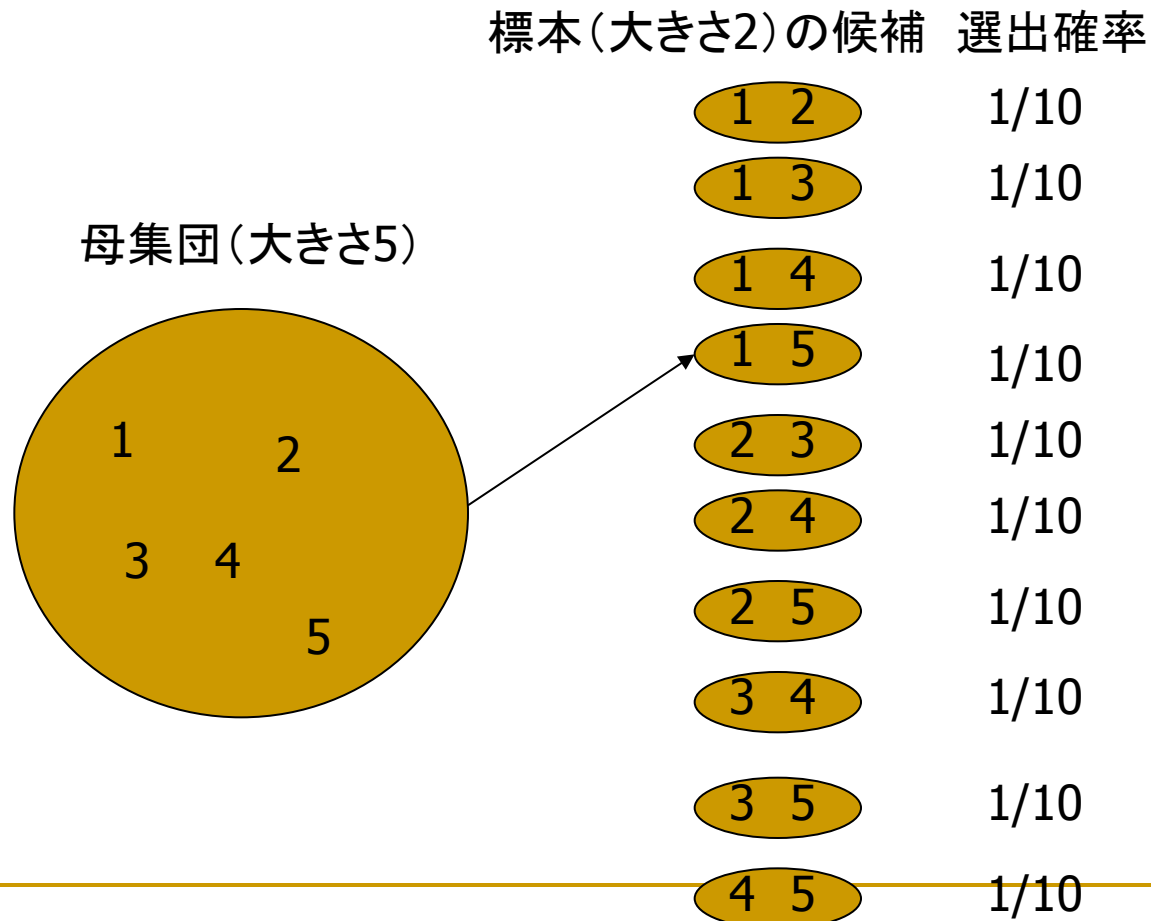
統計量の標本分布(4)

図1: 母集団と標本との関係(数値例)



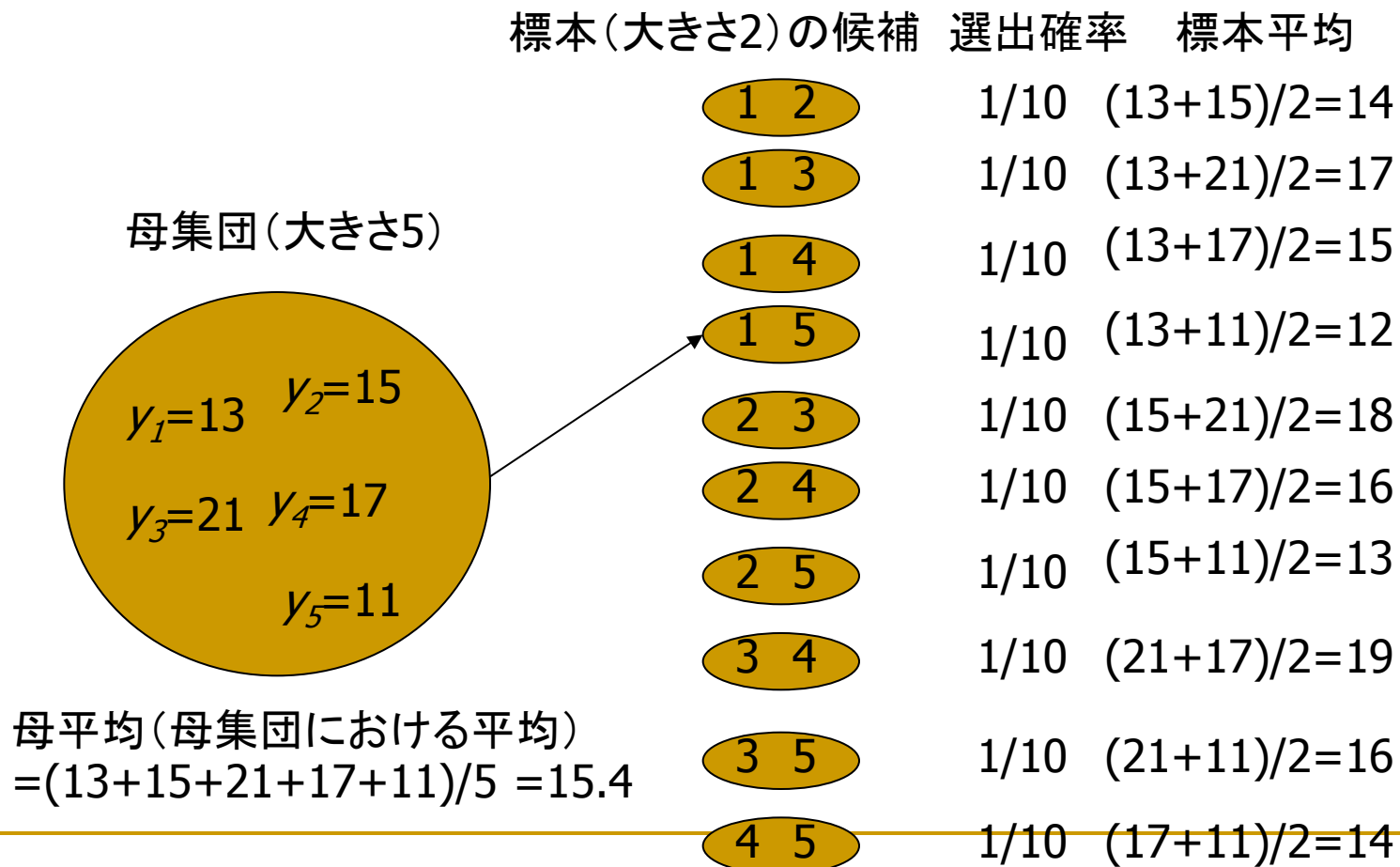
統計量の標本分布(5)

図2: 無作為抽出(数値例)



統計量の標本分布(6)

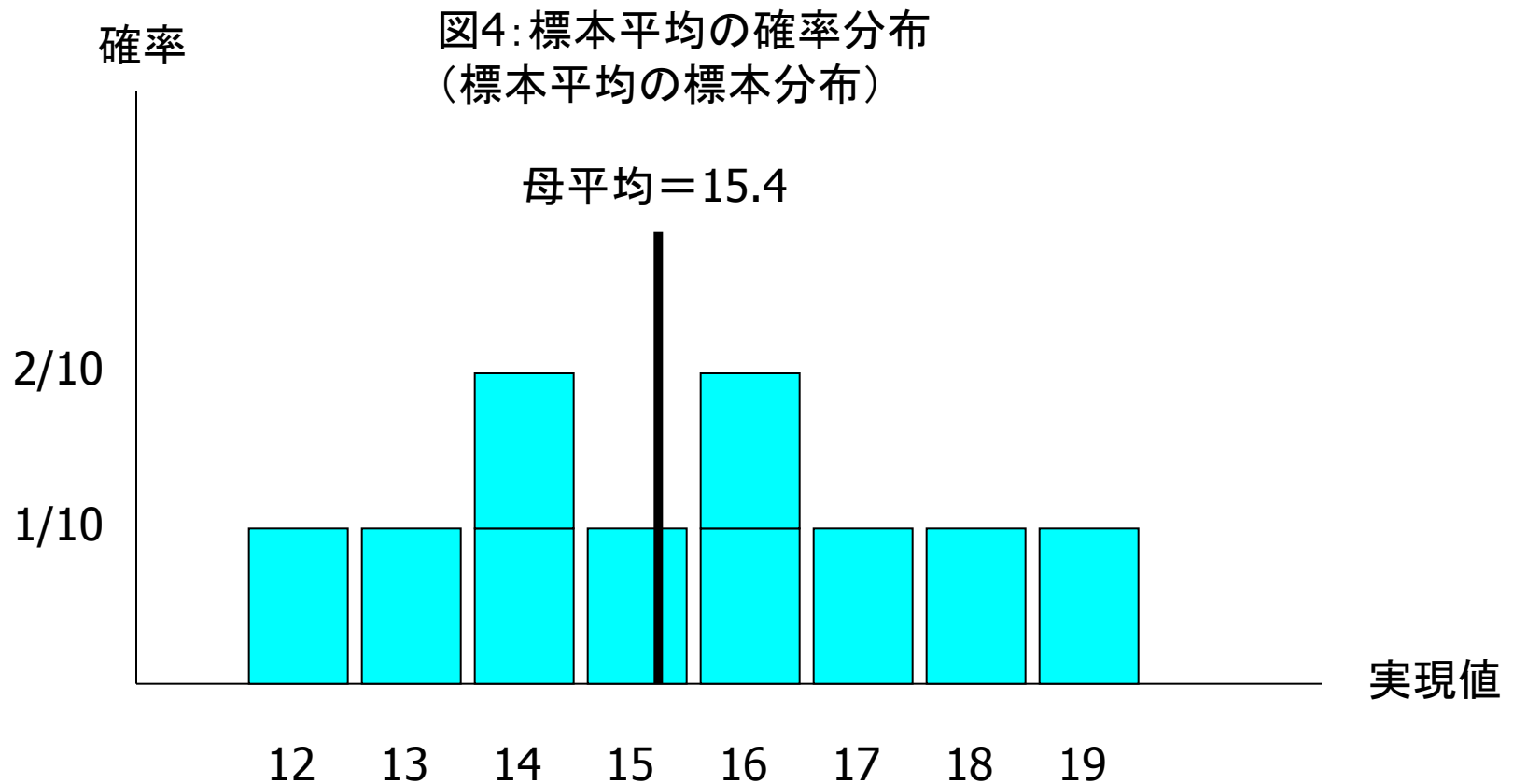
図3: 標本平均の値(数値例)



統計量の標本分布(7)

- 観察できること
 - どの標本平均も母平均と異なる。
 - 標本誤差がある。
 - ひとつひとつの標本平均と母平均とをばらばらに比べているかぎり、「標本誤差がある」ということ以外はなにもいえない。
- 見方を変えて...
 - 標本平均は確率的に変化するもの(確率変数)

統計量の標本分布(8)



統計量の標本分布(9)

- 標本平均(確率変数)と母平均との関係
 - 標本平均の出方をあらわした確率分布の中心の位置に母平均が位置している。
 - 標本平均の期待値
$$\begin{aligned} &= (1/10) \times 12 + (1/10) \times 13 + (2/10) \times 14 \\ &\quad + (1/10) \times 15 + (2/10) \times 16 + (1/10) \times 17 \\ &\quad + (1/10) \times 18 + (1/10) \times 19 \\ &= 15.4(\text{母平均}) \end{aligned}$$
 - 標本平均の出方全体(確率分布)と母平均には対応関係が見出せる。

統計量の標本分布(10)

■ 実験

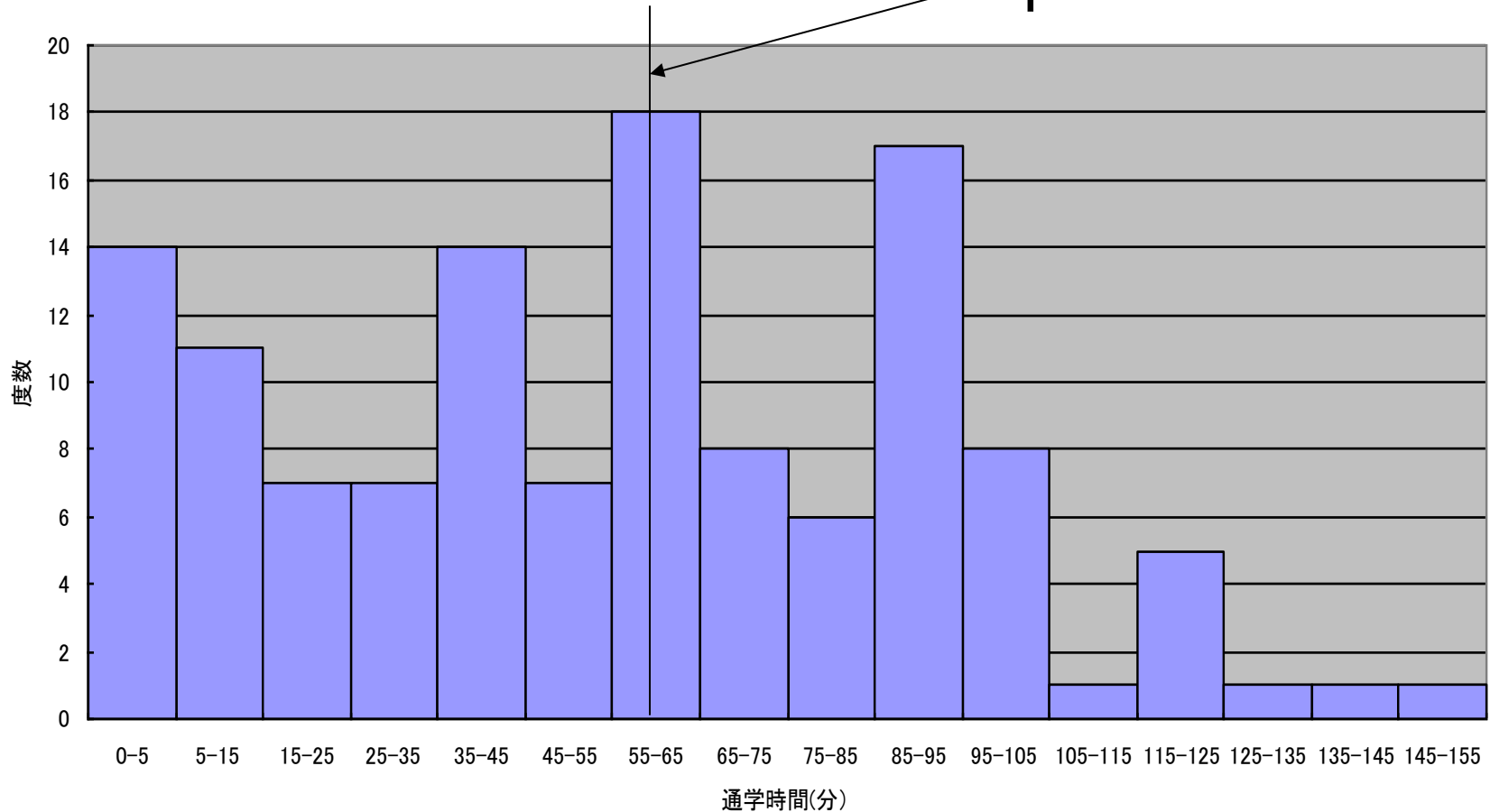
- 母集団：
2002年X月X日の「統計理論」出席者
- x : 通学時間(分)
- 実験の内容：
 - 「 n 人の出席者を無作為抽出して、標本平均を求める」を300回繰り返す。
 - 300回求めた標本平均の度数分布(経験的な標本分布)を描く。

統計量の標本分布(11)

図5: 母集団における通学時間の分布

図1: 母集団における通学時間の分布

$\mu=60$

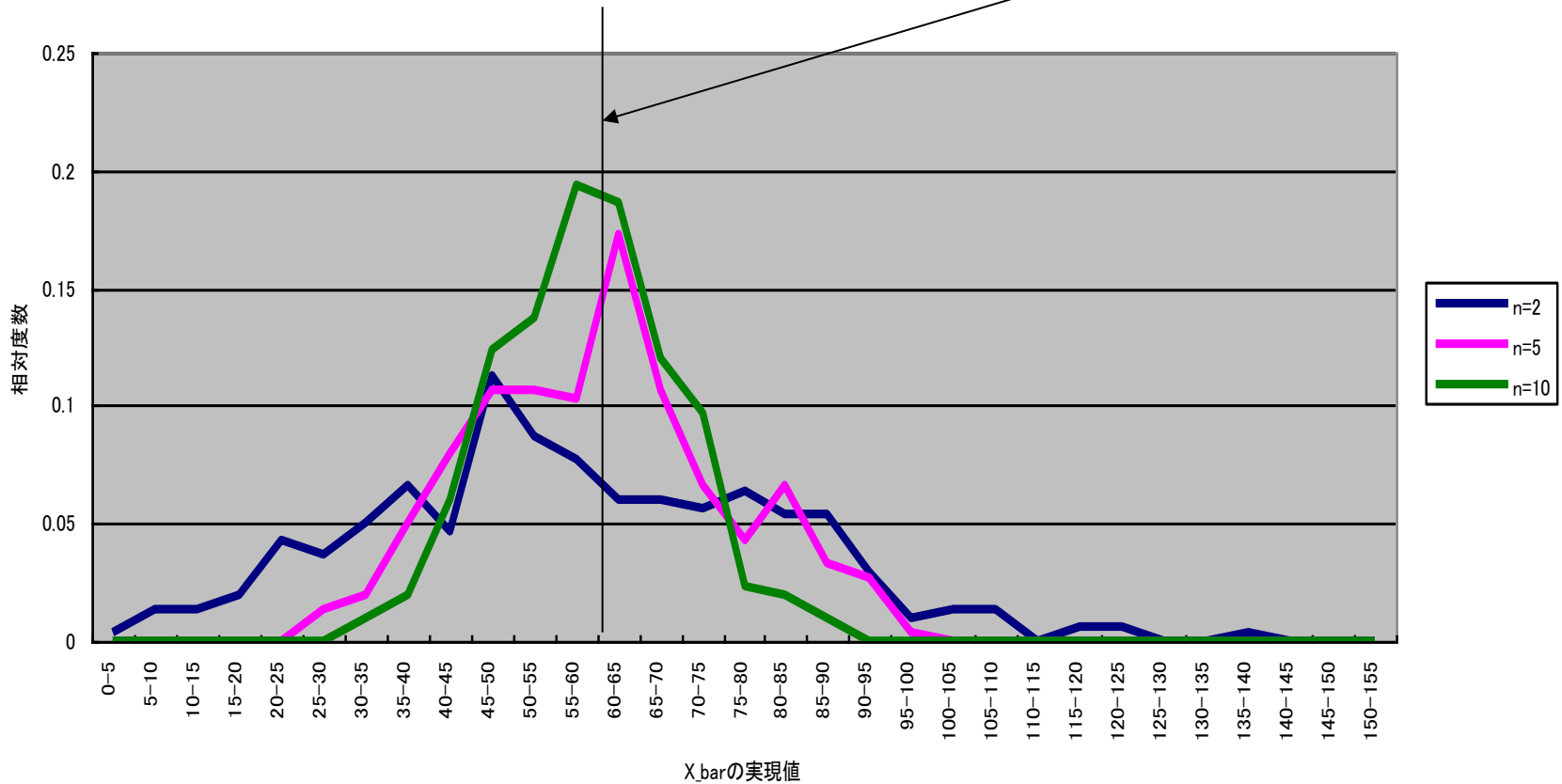


統計量の標本分布(12)

図6: 標本平均の出方(標本平均の標本分布)

図2: 標本平均 \bar{X} の標本分布(シミュレーションによる近似)

$\mu=60$



統計量の標本分布(13)

- 同様に、標本分散 s^2 の標本分布も考えられる。
 - ただし、標本平均の標本分布に比べると、利用頻度は格段に少ない。
 - とはいえ、次の事実は重要：
 - $E(s^2) = \sigma^2$

統計量の標本分布:まとめ

- 統計量の標本分布とは?
 - 標本抽出にともなって生じる統計量の確率分布
 - 標本抽出を何度も繰り返したら観察できるであろう、統計量(たとえば標本平均)の確率分布
- 統計量の標本分布がなぜ必要か?
 - 統計量のひとつの実現値だけを見ている、統計量と母数との関係は(両者が異なるということ以外)わからない。
 - 統計量の確率変数としての挙動全体(確率分布)を利用すると、母数との関係がわかる。