統計学 2

11. 標本平均と母平均 (1)

矢内　勇生

2019年5月23日

高知工科大学　経済・マネジメント学群

今日の目標

たいすう
- 大数の法則を理解する
- 標本分布と標準誤差を理解する

大数の法則
(Law of Large Numbers: LLN)

- ある試行（例：コイン投げ）においてある事象が起こる（例：表が出る）確率がπ（例：0.5）だとする（さらに、各試行は他の試行に影響を及ぼさない[各試行は独立である]とする）

- **大数の法則**：試行回数を増やすにつれて、ある事象が起こる割合（比率）はπに近づく

★ 厳密には、「大数の弱法則（Weak Law of Large Numbers）」と「大数の強法則（Strong Law of Large Numbers）」を区別する必要がある

LLNの例1：コイン投げ（1）

- 正しいコインを投げたとき、表が出る確率 $\pi = 0.5$
  - 実際にコインをn 回投げたとき、表が出る割合 p は0.5になる？

LLNの例1：コイン投げ（2）

表が出る割合 p は0.5になるとは限らない！

- 実際にコインを投げてみると
  - 1回目に表が出た → p = 1/1 = 1
  - 2回目も表が出た → p = 2/2 = 1
  - 3回目も表が出た → p = 3/3 = 1
  - 4回目は裏が出た → p = 3/4 = 0.75

LLNの例1：コイン投げ（3）

- コインを投げる回数が少ないとき、表が出る割合p は0.5から離れた値になりやすい

▸ LLN：投げる回数 n を大きくすれば、p が0.5に近づく！（シミュレーションで示す）

★ コイン投げの回数 = 標本サイズ

標本分布（sampling distribution）

- 母集団からサイズ n の標本を単純無作為抽出するとき、選ばれる個体の組み合わせは何通りもある

- 抽出可能な組み合わせをすべて考え、それぞれの標本で統計量（例：女性の割合）を求めると、その統計量は分布する

- こうして求められる分布（統計量の標本ごとの分布）を**標本分布**と呼ぶ

標本分布の例（1）

個人	A	B	C	D	E
身長(cm)	158	159	160	161	162

5人の母集団からサイズ2の標本を選ぶ：
全部で${}_5C_2$=10 通りの選び方

標本	{A, B}	{A, C}	{A, D}	{A, E}	{B, C}	{B, D}	{B, E}	{C, D}	{C, E}	{D, E}
標本平均	158.5	159	159.5	160	159.5	160	160.5	160.5	161	161.5

標本分布の例（2-1）

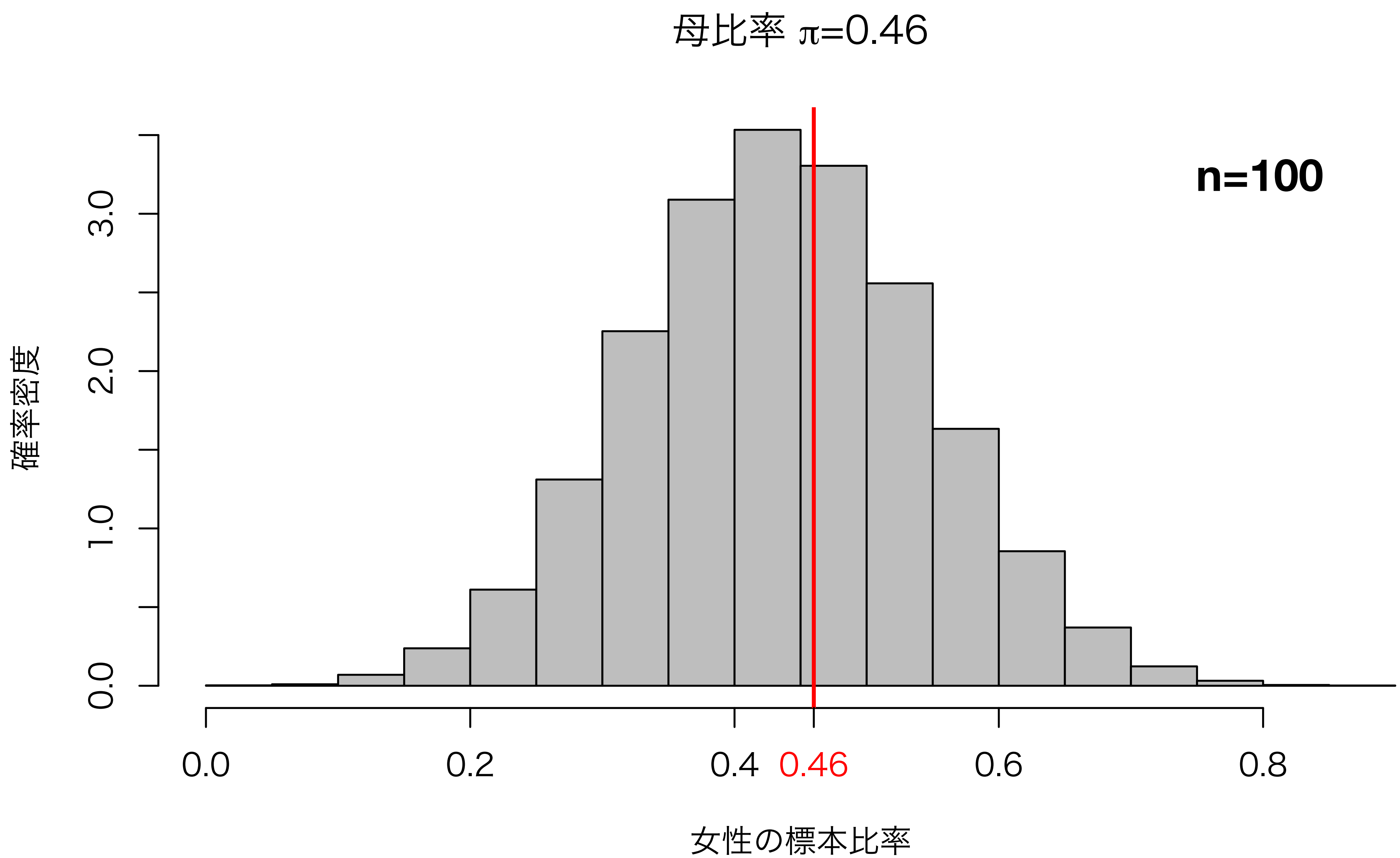
- 男5400人、女4600人の計1万人から100人選ぶ

- 1万人から100人を選ぶ方法は全部で約6.5×10^{241}通り
  → 全部の組み合わせ：標本分布

➡ 数が多いので全部の組み合わせを考えるのは困難

➡ コンピュータ・シミュレーションでn=100のサンプルを100万個抽出し、擬似的な標本分布を得る

標本分布の例（2-2）



標本分布と標本サイズ

- 標本サイズnを変えれば、得られる標本分布も変わる
    - 10人の母集団から2人を選ぶ：45通り
    - 10人の母集団から3人を選ぶ：120通り
    - 10人の母集団から9人を選ぶ：10通り

➡ 可能な組み合わせが異なれば、統計量の分布も変化する

標準誤差 (standard error: SE)

- 標準誤差 = 標本分布に現れるばらつき

  = **統計量の標準偏差**

- 母集団が十分大きい（目安：母集団が標本サイズnの100倍以上）とき、標準誤差 SE は

$$\mathrm{SE} = \frac{\text{母標準偏差}}{\sqrt{\text{標本サイズ}}} = \frac{\sigma}{\sqrt{n}}$$

または

$$\mathrm{SE} = \frac{\text{母標準偏差の推定値}}{\sqrt{\text{標本サイズ}}} = \frac{s}{\sqrt{n}}$$

標準誤差の特徴

- 標準誤差の大きさは、標本サイズn の平方根に反比例する

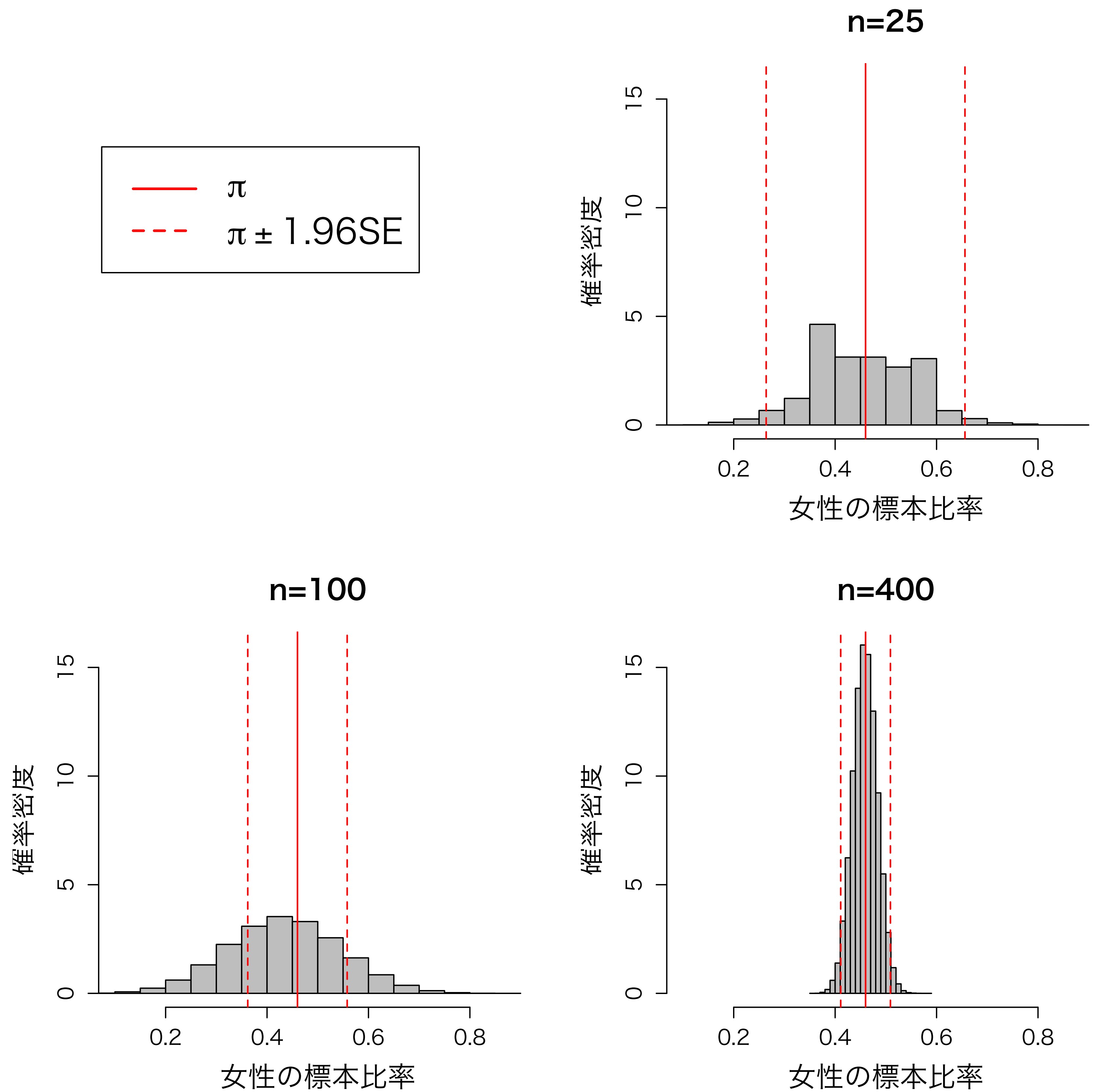
- nを4倍にすれば、SE は半分になる！

➡ nが大きいほど、統計的推測が正確になる

標準誤差の例（1）

- 女性比率が0.46 の100万人の母集団から n人の標本を抽出し、女性比率を求めるときの標準誤差
  - n=25のとき： $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{25}} \approx \frac{0.5}{5} = 0.1$
  - n=100のとき： $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{100}} \approx \frac{0.5}{10} = 0.05$
  - n=400のとき： $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.46(1-0.46)}}{\sqrt{400}} \approx \frac{0.5}{20} = 0.025$

標準誤差の例（2）



π
π ± 1.96SE
n=25
確率密度
0
5
10
15
0.2
0.4
0.6
0.8
女性の標本比率
n=100
確率密度
0
5
10
15
0.2
0.4
0.6
0.8
女性の標本比率
n=400
確率密度
0
5
10
15
0.2
0.4
0.6
0.8
女性の標本比率

母集団の大きさと標本サイズ（1）

- 都道府県ごとに増税に賛成か反対か調べたい
  - 東京都の人口：約1300万人
  - 岩手県の人口：約130万人

- 東京の標本サイズは岩手の標本サイズの10倍にするべき？

母集団の大きさと標本サイズ（2）

- 仮に、東京でも岩手でもちょうど半分の人が増税に賛成（反対）だとする
- 標本サイズを100にすると
  - 東京の標準誤差 = 0.5/10 = 0.05
  - 岩手の標準誤差 = 0.5/10 = 0.05
- 母集団の人口が10倍でも、同じ標本サイズで同じ精度の調査ができる

➡（母集団が十分大きいとき）標本誤差は母集団の大きさに依存しない！

母集団の大きさと標本サイズ（3）

- Good news：東京の調査を岩手と同じ精度で行うためには、岩手と同じサイズの標本を抽出すればよい
- Bad news：岩手の調査を東京の調査と同じ精度にするためには、東京と同じサイズの標本を抽出しなければならない

標本平均と母平均

- 標本平均：標本から求められる平均値（統計量）
- 母平均：母集団の平均値（母数）
- 例：日本の成人男性（母集団）の身長を知るために、単純無作為抽出で1000人抽出した
    - 標本平均：1000人の身長の平均値

↓ 推測する！

- 母平均：日本の成人男性身長の平均値（未知、興味の対象）

標本平均と母平均の差

- 標本平均の誤差 = 標本平均 – 母平均
- 「標本平均 = 母平均」になるとは限らない

➡ 標本平均には誤差がある

➡ 標本平均は分布する！

標準誤差 ≤ 母標準偏差

- 標準誤差 ≤ 母標準偏差

➡標本平均のばらつき ≤ 母集団のばらつき

- なぜ？
  - シミュレーションで示す

標本分布のシミュレーション

★母集団の身長が平均170、 標準偏差5.5の正規分布に従うことを知っているとする

★このとき、標本平均の分布をシミュレーションで手に入れる（標本の数 [標本サイズではない！] =10万）

- 目的：標本平均のばらつきが母集団のばらつき以下になる理由を理解する

標本サイズ（*n*）が1のとき

- 標本平均は標本として抽出された値そのもの
- 標準誤差　$\mathrm{SE} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{\sigma}{\sqrt{1}} = \sigma$

➡ 「標本平均のばらつき=母集団のばらつき」になるはず

標本サイズ（*n*）が2のとき

- 標本平均は標本として抽出された2つの値の平均値
- 標準誤差　$\mathrm{SE} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{\sigma}{\sqrt{2}} \approx 0.7\sigma$

➡「標本平均のばらつき=母集団のばらつきの約0.7倍」になるはず

標本サイズ（*n*）が10のとき

- 標本平均は標本として抽出された10個の値の平均値
- 標準誤差　$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{10}} \approx 0.3\sigma$

➡「標本平均のばらつき=母集団のばらつきの約0.3倍」になるはず

標準誤差の特徴

- 標準誤差の大きさは、標本サイズ*n* の平方根に反比例する
  - *n*を4倍にすれば、SE は半分になる！

➡ *n* が大きいほど、統計的推測が正確になる（大数の法則による）

今日のまとめ

- 大数の法則：標本サイズが大きくなるほど、誤差が小さくなる
- 標本には誤差がつきもの
    - 標本分布と標準誤差
    - 「平均すれば」うまく推定できる
    - しかし、1つひとつの標本平均は信用できない
    - では、どうする？（次回のテーマ）