

第2章 統計的な推測

第2節 統計的な推測 (推測統計の基礎)

第1節の統計

第1節では、確率分布から平均などの統計量を正確に算出していた。

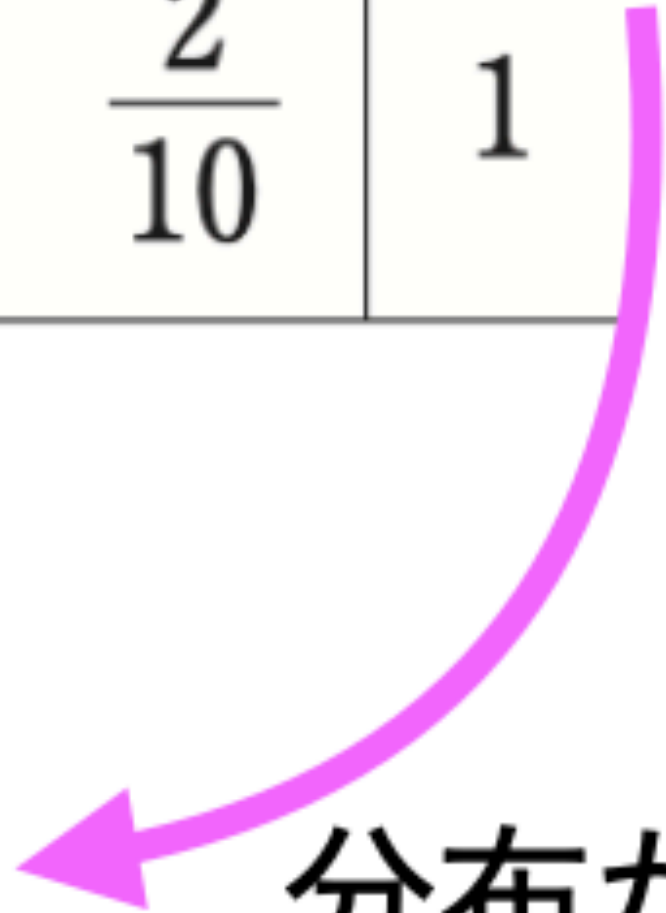
X	0	1	2	計
P	$\frac{5}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	1

平均： m

分散： σ^2

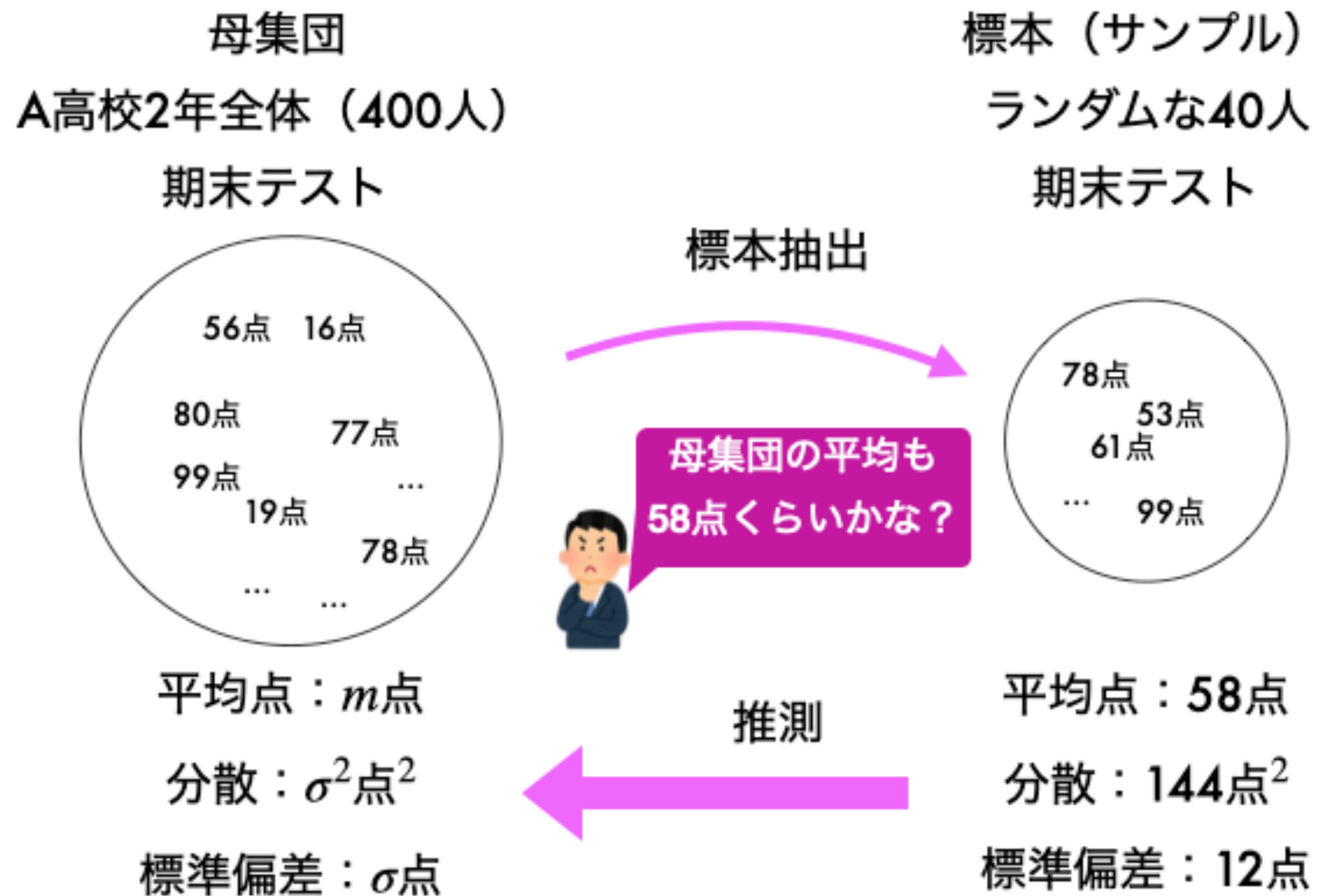
標準偏差： σ

分布から
算出



第2節のモチベーション（推測統計）

第2節では、母集団から抽出されたデータの集合の統計量から母集団の特徴を推測することをモチベーションとする。



全数調査と標本調査

- **全数調査**

- 調べたい対象全体のデータを集める調査方法
- 国勢調査や、ある高校内の授業アンケートなど
- 最も理想的な調査だが、多くのコストがかかり、現実的な手法でない場合も多い



全国民へ。
職業は？
住所は？

国勢調査

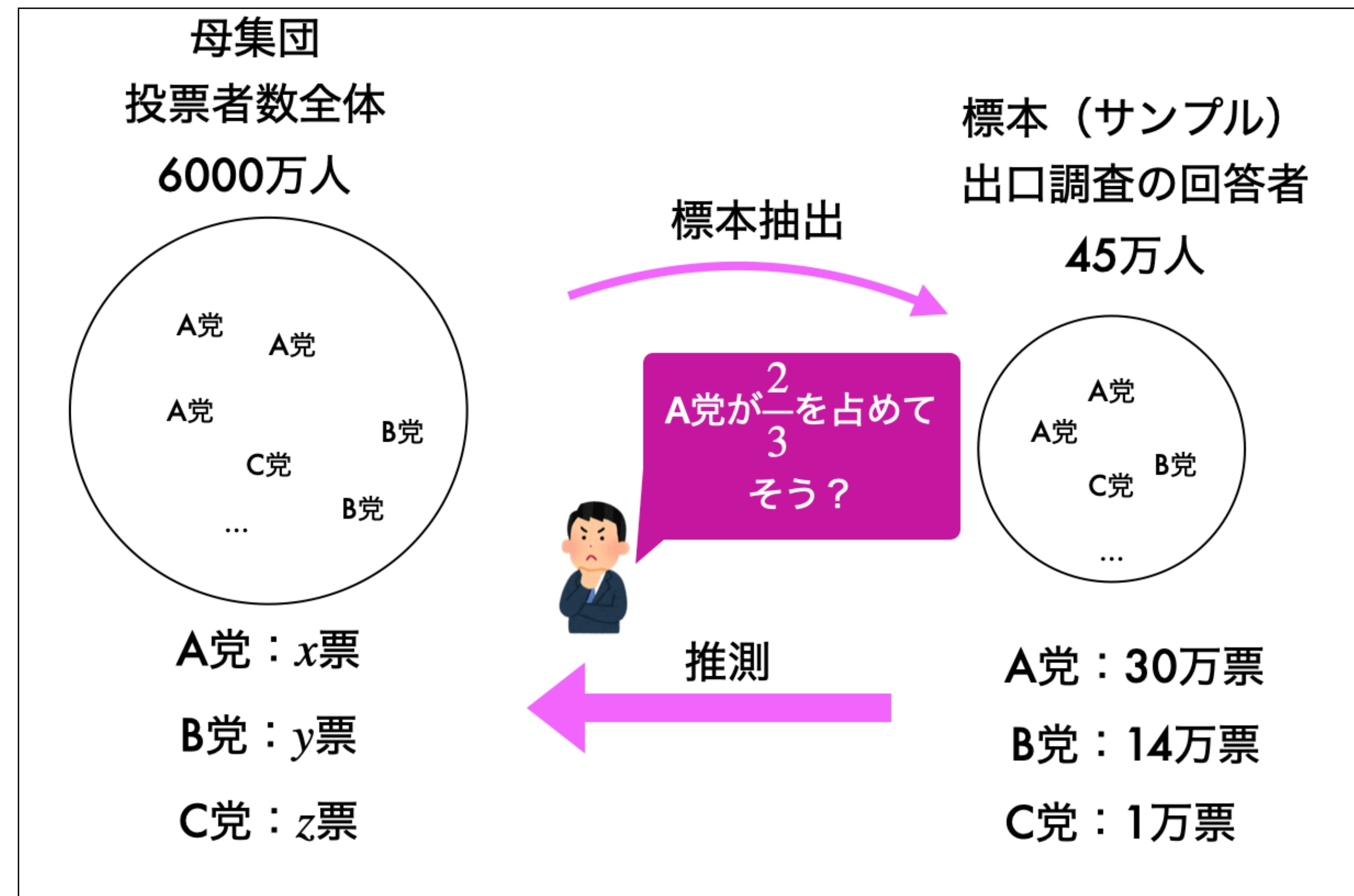
全数調査と標本調査

• 標本調査

- 調べたい対象から一部抽出された標本（サンプル）を使う調査方法
- 例えば、選挙の出口調査など
- 現実的な調査方法ではあるが、特定の属性を持つデータを多く含むことなどがないようにする
- 例えば、40代ばかりに調査をする、関東のみ調査する、などは避けたい



どこの党に入れましたか？



標本（サンプル）

- 母集団

- 調べたい対象全体の集合
- 母集団の要素の総数のことを母集団サイズと呼び、 N で表す

- 標本（サンプル）

- 母集団から抽出された要素の集合
- ランダム抽出であることが望ましい
- 例えば、A高校2年期末テストの結果を母集団とすると、2年A組のテスト結果は標本になる。ただし、この場合はランダム抽出ではない。

- 標本サイズ、サンプルサイズ

- 標本の大きさ、要素数のこと
- n で表す

*サンプルサイズと似た言葉にサンプル数がありますが、別の用語です。

サンプル数は集合の個数を指し、サンプルサイズは集合の要素数を指します。巷では非常に誤用が多いです...

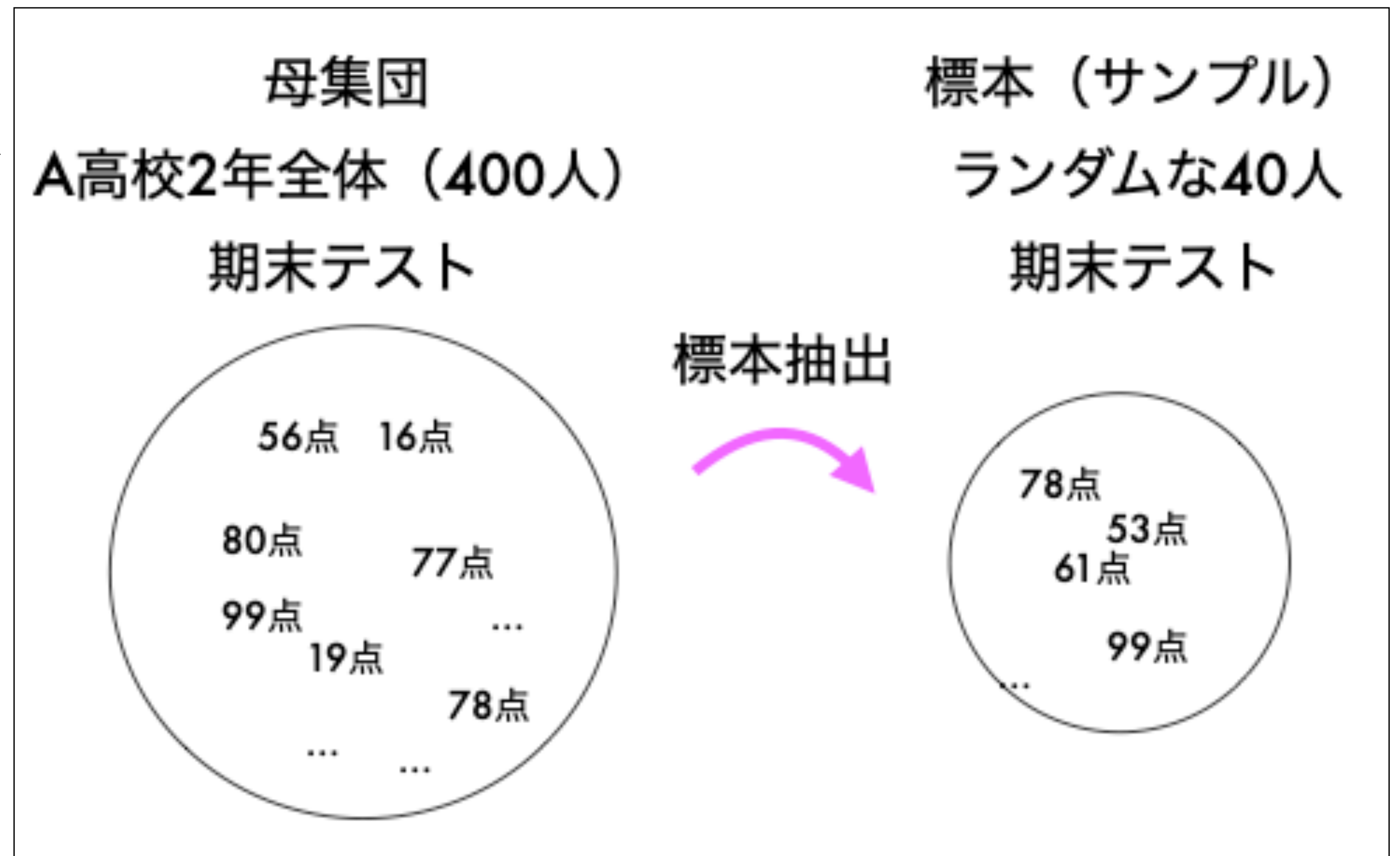
標本抽出

- 標本（サンプル）抽出

- 母集団から標本（サンプル）を抜き出すこと

- 無作為抽出

- 母集団の各要素を等しい確率で抽出すること
- 標本抽出のあるべき姿だが、現実的にはなかなか難しい



母数（パラメタ）

- 母数（パラメタ）

- 母集団の特徴を表す値
- 例えば
 - ◆ 母集団における平均（母平均）
 - ◆ 母集団における分散（母分散）
 - ◆ 母集団における標準偏差（母標準偏差）など
- 一方で、標本（サンプル）の平均は標本平均であり、これと区別する
- $B(n, p)$ の試行回数 n や成功確率 p も母数であり、 $N(m, \sigma^2)$ の m, σ^2 も母数



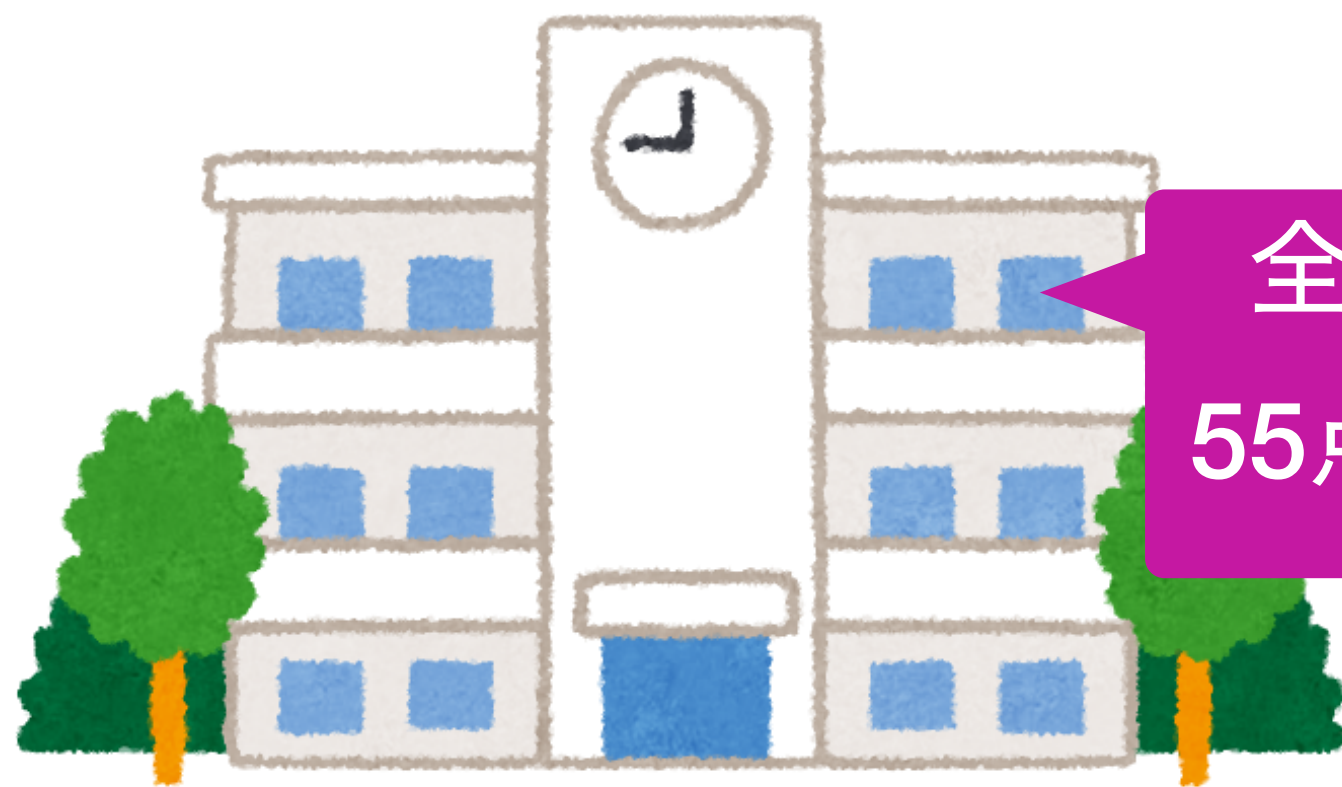
共通テスト当日に全国平均点（母平均）を知りたいところですが、
現実的にかなり難しいですね。
一般に、母数を知るのは難しく、母数を推定するのが推測統計です。

母集団の要素数のことを母数と呼ぶ誤用が巷では目立ちます。

SNSなどでこの誤用をすると統計警察がサイレン鳴らして飛んできます。気をつけましょう。

母平均と標本平均

ある高校の全校生徒300人のテストの点数を母集団とする。300人のテスト平均点は母平均になるが、ランダムな5人の標本の平均は標本平均となる。
標本平均はどうしてもデータに偏りが出る可能性がある。



全校生徒の平均点は
55点です！（母平均）

僕ら5人の平均点は
80点です！（標本平均）



2年生
80点



1年生
50点



3年生
90点



3年生
90点



3年生
90点

母数を確率変数で表す

母数（母集団の特徴を表す値）である母平均を m 、母分散を σ^2 、母標準偏差を σ とする。

母集団から1つ無作為抽出した時の値を確率変数 X とすると、この X の期待値、分散、標準偏差は以下のように各母数と一致する。

$$E[X] = m \quad V[X] = \sigma^2 \quad \sigma[X] = \sigma$$



母数を今までの確率変数で表現できる、
ということをいっています！
(当たり前だろ、と思うかも知んですが...)

例題

0と書かれたカードが5枚、1と書かれたカードが3枚、2と書かれたカードが2枚ある。この10枚のカードを母集団とみて、1枚カードを引き、出た数字を確率変数 X とおく。

母集団分布は大きさ1の無作為標本の値 X の確率分布と一致するが、

(1) 母平均 m は？ (2) 母標準偏差 σ は？

X	0	1	2	計
P	$\frac{5}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	1

解答

$$m = E[X] = 0 \cdot \frac{5}{10} + 1 \cdot \frac{3}{10} + 2 \cdot \frac{2}{10} = \frac{7}{10}$$

$$\sigma = \sigma[X] = \sqrt{\left(0^2 \cdot \frac{5}{10} + 1^2 \cdot \frac{3}{10} + 2^2 \cdot \frac{2}{10}\right) - \left(\frac{7}{10}\right)^2} = \frac{\sqrt{61}}{10}$$

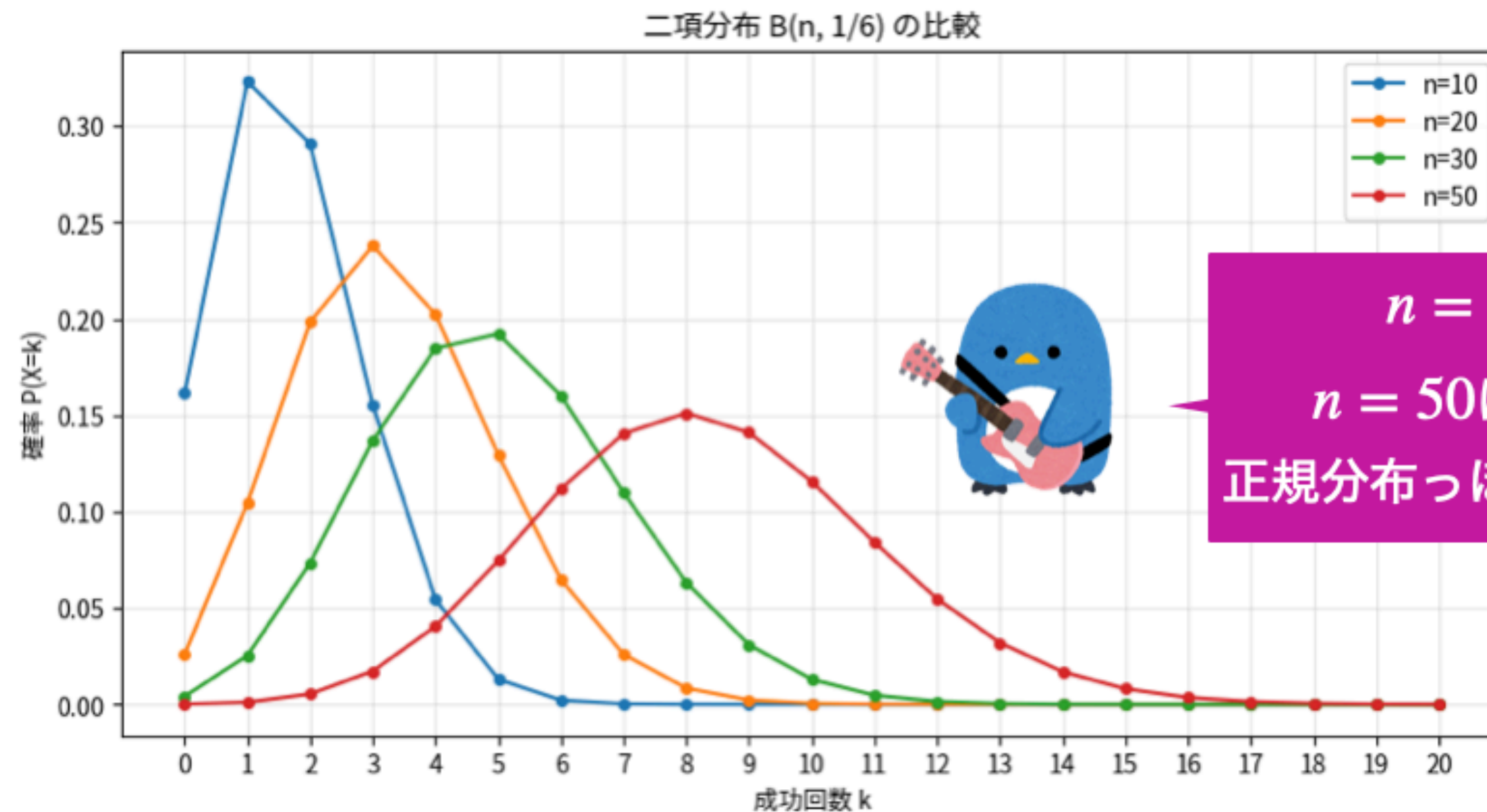
母集団分布の連続型確率分布による近似

母集団サイズは非常に大きいことが多く、離散データであっても連続型確率変数の分布で近似されることが多い。

例えば、二項分布の正規分布による近似など

二項分布の正規分布による近似

二項分布 $B(n, p)$ に従う確率変数 X は、 n が大きい時、近似的に正規分布 $N(np, npq)$ に従う。ただし、 $q = 1 - p$ である。



$n = 30$ や
 $n = 50$ は確かに
正規分布っぽいですね！

中心極限定理という、
統計学で一番重要な定理とも言える定理で
この命題は証明が可能です、高校範囲を
超えるので、ここではfactとします。

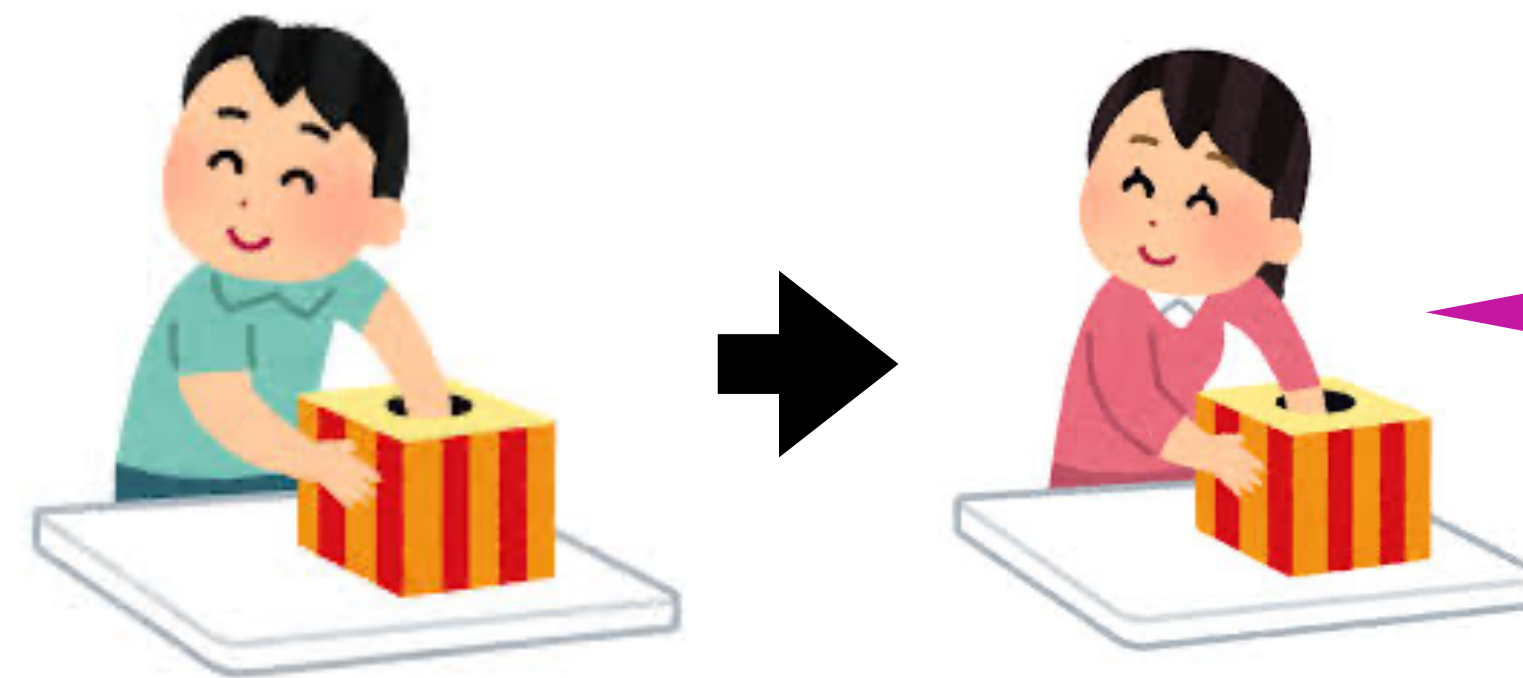
復元抽出と非復元抽出

- 復元抽出

- 母集団から標本（サンプル）を抽出するとき、毎回元に戻しながら次のものを1個ずつ取り出すこと
- 反復試行での抽出は復元抽出
- 独立な試行になる

- 非復元抽出

- 母集団から標本（サンプル）を抽出するとき、毎回元に戻さずに次のものを取り出すこと
- 従属な試行になる



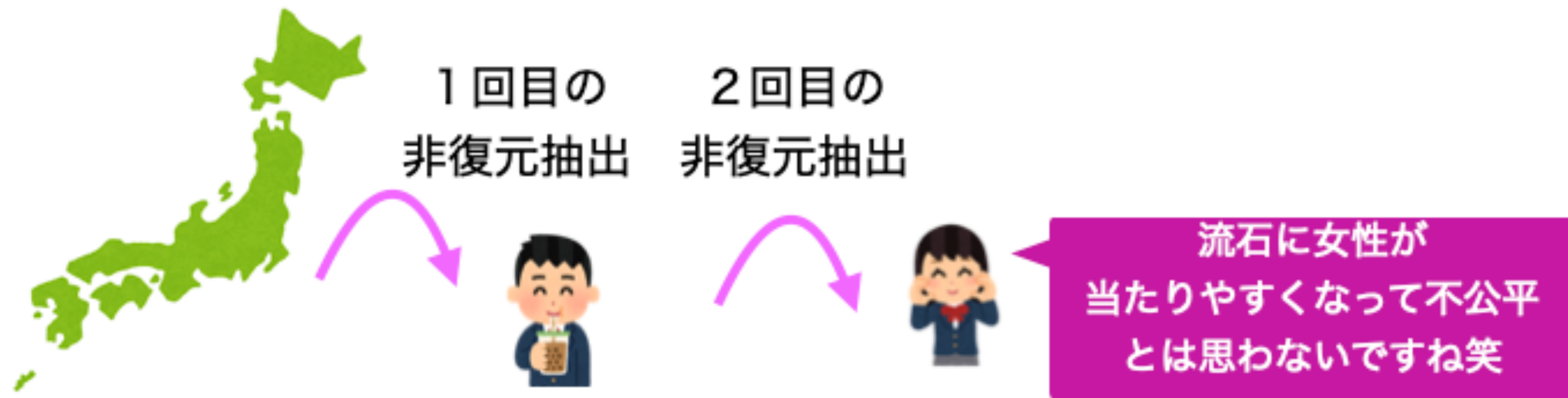
非復元抽出

前のやつが当たり引いたせいで
当たりの確率が下がったぞ😡

母集団サイズが大きいときの非復元抽出

母集団サイズがサンプルサイズに比べて十分に大きい場合、非復元抽出と復元抽出の差異は小さくなる。

例えば、日本全国から元に戻さず1人ずつ抽出し、男性であるかを調べたい。
最初が男性であろうと女性であろうと、1億3000万人の日本人口レベルだと2回目の非復元抽出に影響をほぼ与えない。



標本（サンプル）の統計量

標本（サンプル）サイズ n の標本について、各標本の値を $X_1, X_2, \dots, X_n$ とすると、標本平均 $\bar{X}$ と標本分散 S^2 、この標本標準偏差 S は

$$\bar{X} := \frac{1}{n}(X_1 + X_2 + \dots + X_n) \quad S^2 := \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2 \quad S = \sqrt{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2}$$

であり、 $\bar{X}, S^2, S$ も確率変数である。

標本標準偏差が別のものを指す場合もあります。詳しく言うと不偏分散 s^2 と呼ばれる分散の平方根 s です。



数1のデータの分析で扱った内容ですね

標本における統計量

母平均 m 、母分散 σ^2 、母標準偏差 σ の母集団から大きさ n の無作為標本を抽出するとき、標本平均 $\bar{X}$ の期待値、分散、標準偏差は

標本平均が
期待されている値

$$E[\bar{X}] = m \quad V[\bar{X}] = \frac{\sigma^2}{n} \quad \sigma[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

証明

データ1つ1つの値を $X_1, X_2, \dots, X_n$ としているので、

$E[X_1] = E[X_2] = \dots = E[X_n] = m$ より、

$$E[\bar{X}] = E \left[\frac{X_1 + X_2 + \dots + X_n}{n} \right] = \frac{1}{n} (E[X_1] + E[X_2] + \dots + E[X_n]) = \frac{nm}{n} = m \quad (\text{続く})$$

標本における統計量

母平均 m 、母分散 σ^2 、母標準偏差 σ の母集団から大きさ n の無作為標本を抽出するとき、標本平均 $\bar{X}$ の期待値、分散、標準偏差は

$$E[\bar{X}] = m \quad V[\bar{X}] = \frac{\sigma^2}{n} \quad \sigma[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

標本平均の
データの散らばり

証明

$V[X_1] = V[X_2] = \dots = V[X_n] = \sigma^2$ と、 $X_1, X_2, \dots, X_n$ が互いに独立であることから

$$V[\bar{X}] = V\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{1}{n^2}(V[X_1] + V[X_2] + \dots + V[X_n]) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\text{よって、} \sigma[\bar{X}] = \sqrt{V[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad \blacksquare$$

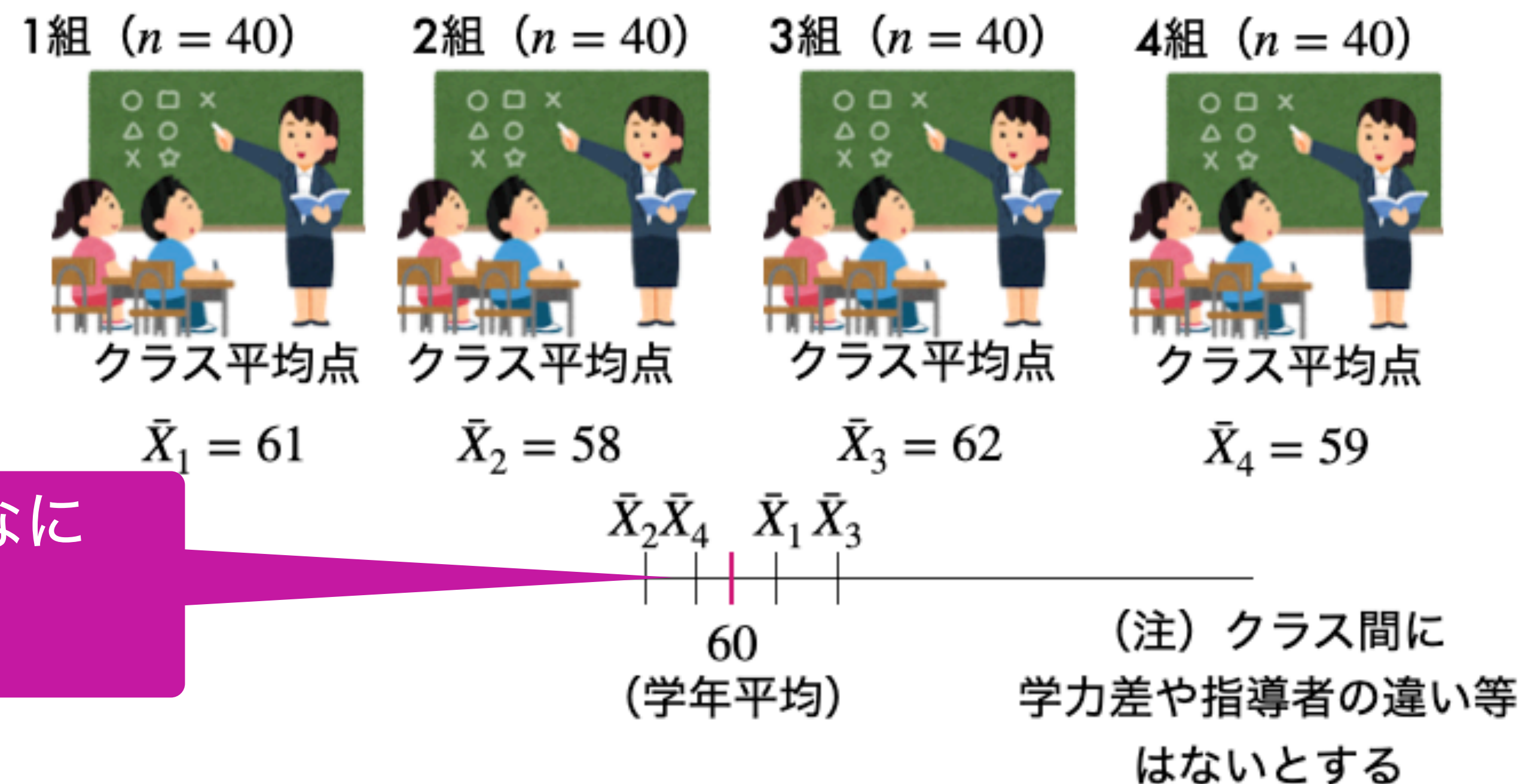
標本平均の分散はなぜnで割るのか

$$V[\bar{X}] = \frac{\sigma^2}{n} \text{は標本平均の散らばりである。}$$

標本サイズ n が大きい場合は、標本平均の値は母平均に近づきそう

→ データがあまり散らばらない

例えば、学年平均60点のテストで
40人のクラス平均点が80点になるような
ことはほぼない。



40人の平均を取ればそんなに
散らばらない

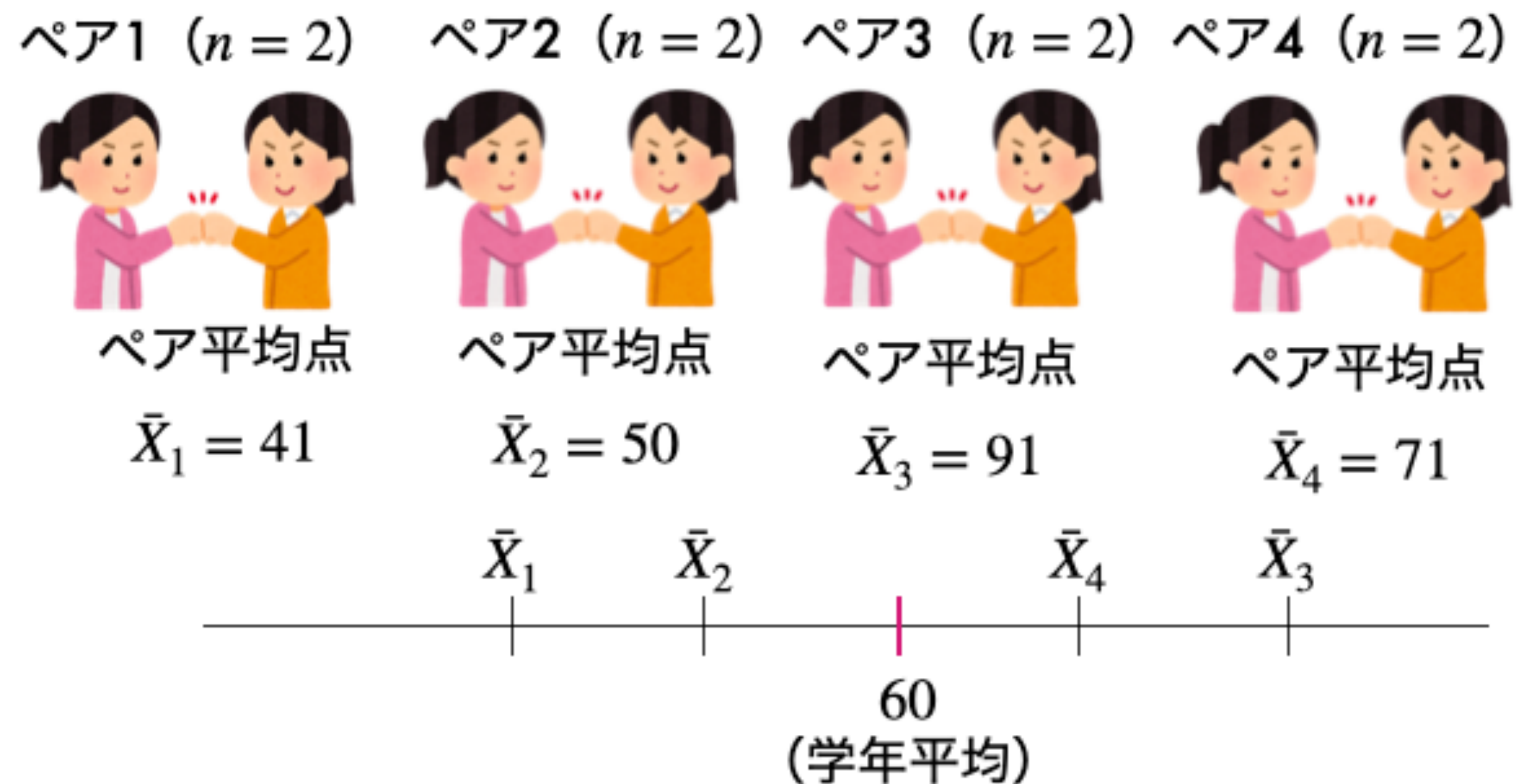
標本平均の分散はなぜnで割るのか

$$V[\bar{X}] = \frac{\sigma^2}{n} \text{は標本平均の散らばりである。}$$

標本サイズ n が小さい場合は、標本平均はやや散らばる

例えば学年平均60点のテストで
二人ペアの平均点は、
ペアによっては91点になったり
41点になることもしばしばある。

ただ、それでも母分散 σ^2 よりは
散らばりは小さい



例題

母平均60、母分散25、母標準偏差5の母集団から、大きさ25の標本を無作為抽出するとき、その標本平均 $\bar{X}$ の期待値と分散、標準偏差は？

解答

$$E[\bar{X}] = 60$$

$$V[\bar{X}] = \frac{\sigma^2}{n} = \frac{25}{25}$$

$$\sigma[\bar{X}] = \frac{\sigma}{\sqrt{n}} = \frac{5}{5} = 1$$