情報科学
【AI・データサイエンス】

第7回
相関と検定

相関
統計的検定

相関

データの広がりと分散
相関

「相関」を一言で！

2つの量の関係性を説明する方法

- 2つの量の関係性
  - 科学的な分析では因果関係を知りたいことが多い
  - 因果とまでいかなくても，関係性の有無が分かると嬉しい

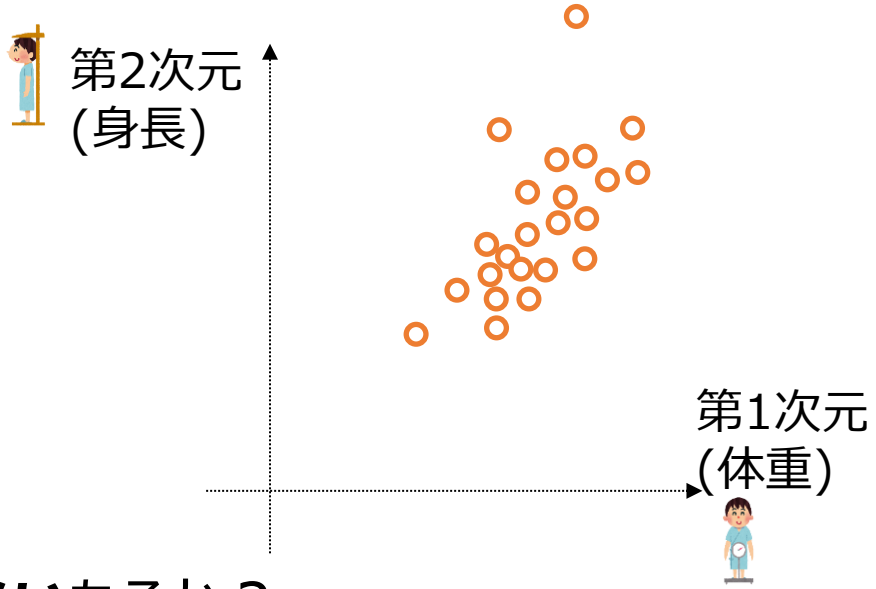
- 例
  - 気温とアイスの売り上げ高の関係性
  - 大気中の二酸化炭素濃度と海水面の高さの関係性
  - 読書量と試験の点数の関係性
  - 土壌の微生物数と作物の成長度合いの関係性
  - 年収額と幸福度の関係性

データの広がりと分散

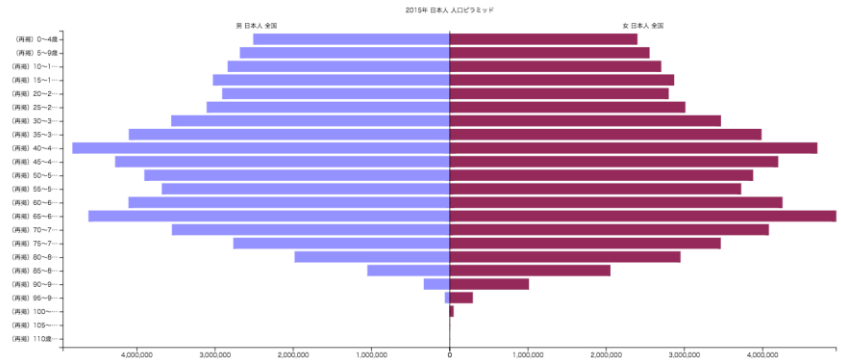
データの分布

データと分布

- データ(デジタル大辞泉より)
  1. 物事の推論の基礎となる事実
     また，参考となる資料・情報
     「―を集める」「確実な―」
  2. コンピューターで，プログラムを
     使った処理の対象となる記号化・
     数字化された資料

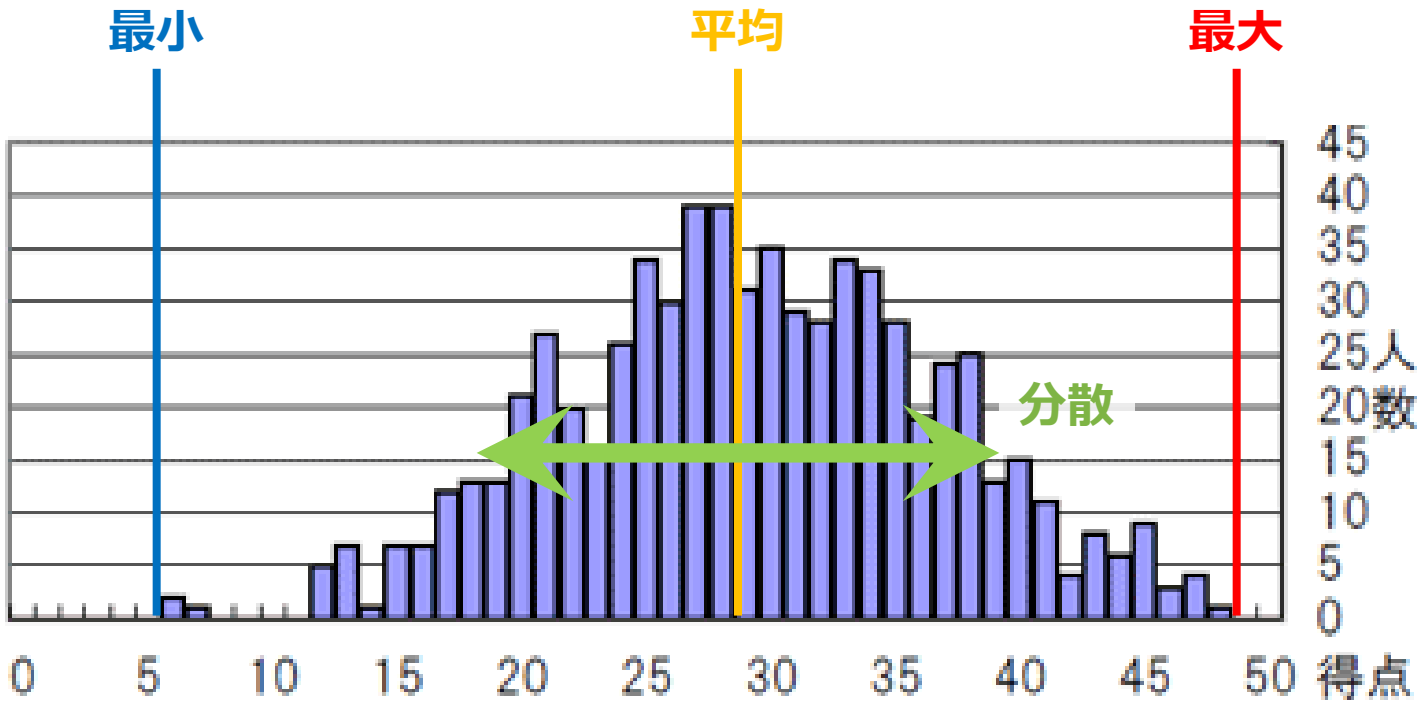


- 分布 ＝ **どんな**データが，**どのくらい**あるか？
  - 身体測定結果の分布
  - 年齢別人口分布
  - 成績の分布
  - 顧客の分布
  - 生産物の分布



年齢別人口分布

分布から分かること



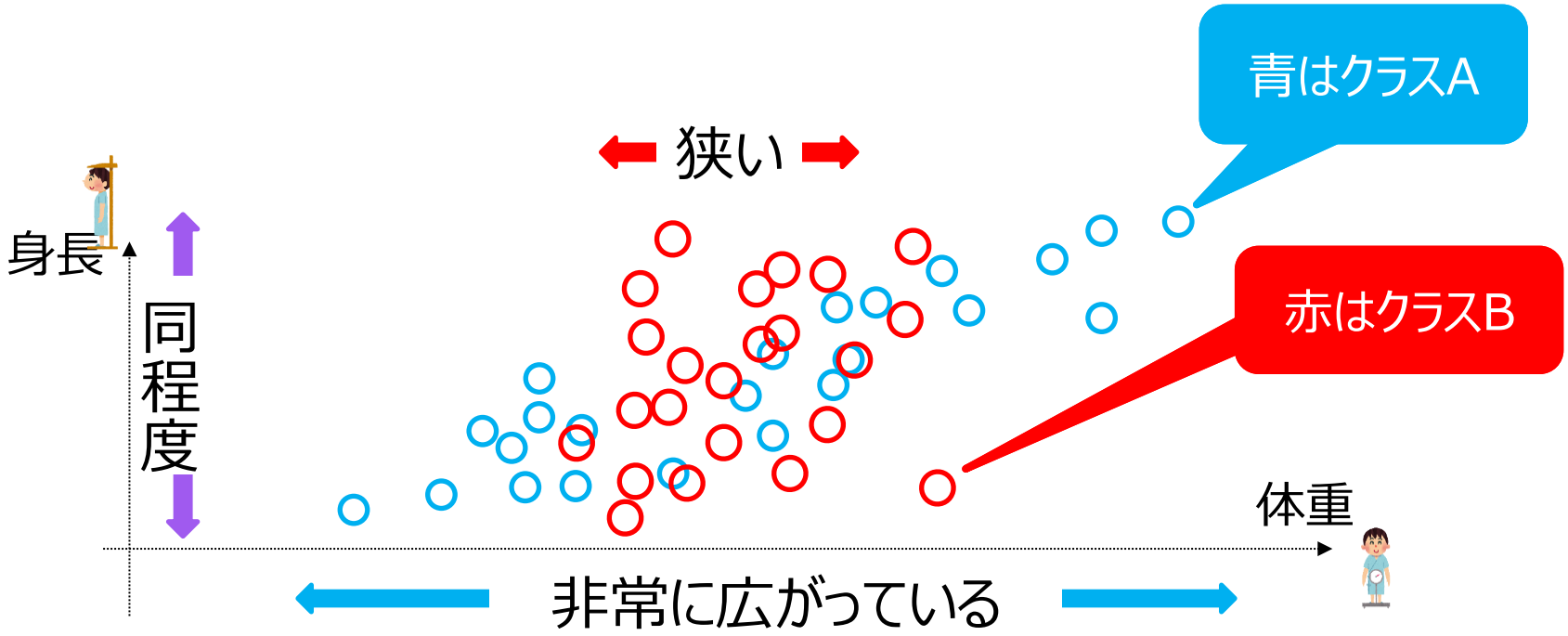
どんな事が見えてくるか？

- 最小，最大
- 平均，分散

分布から分かること ＝ 集団の性質

- どのような個体から構成されてるか
- 例）3年1組の特徴，3年3組の特徴
- 例）日本人口の特徴，中国人口の特徴

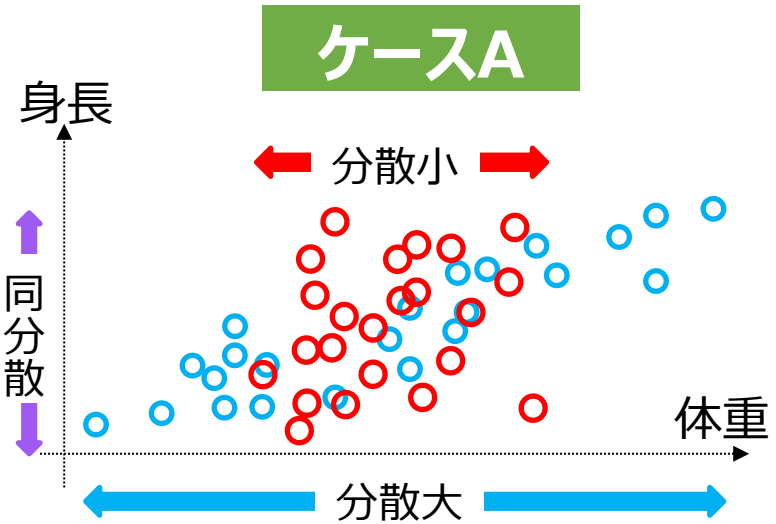
分散＝データの広がり具合

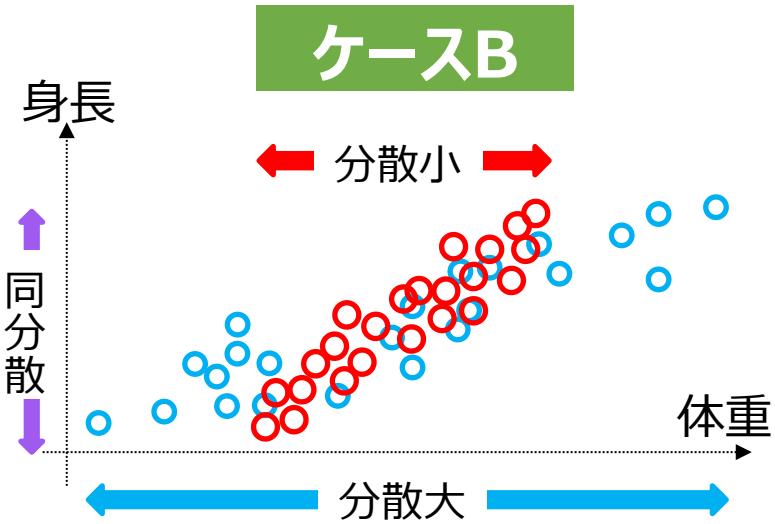


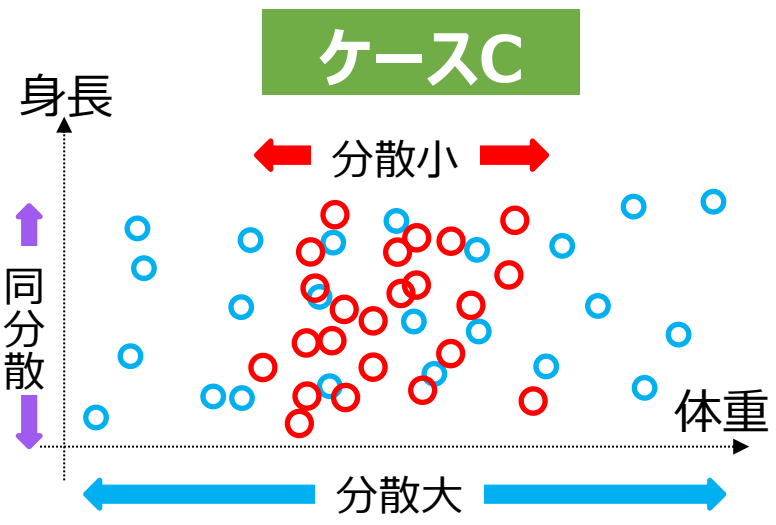
- 分散＝広がり具合を数値で表現(＝定量化)したもの
  - 体重の分散は「**青** ＞ **赤**」
  - 身長の分散は「**青** ≒ **赤**」

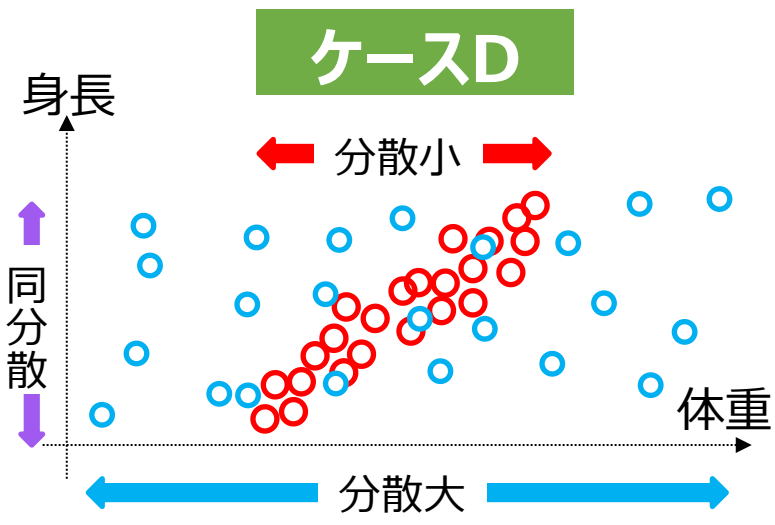
⇨ 定量化することで集団の比較が可能に

分散だけで十分か？


ケースA
身長
分散小
同分散
体重
分散大


ケースB
身長
分散小
同分散
体重
分散大

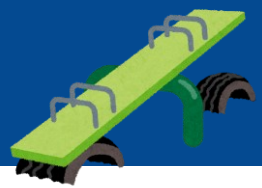

ケースC
身長
分散小
同分散
体重
分散大


ケースD
身長
分散小
同分散
体重
分散大

相関

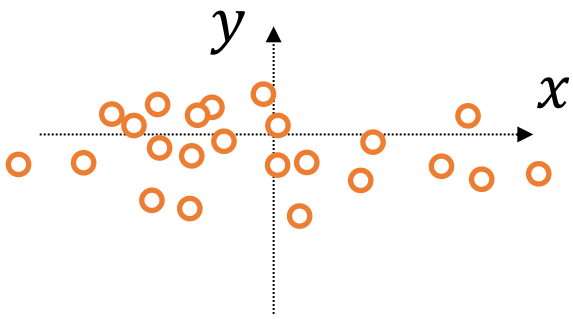
Xが大きくなると，Yも大きくなる

Xが大きくなると，Yは小さくなる

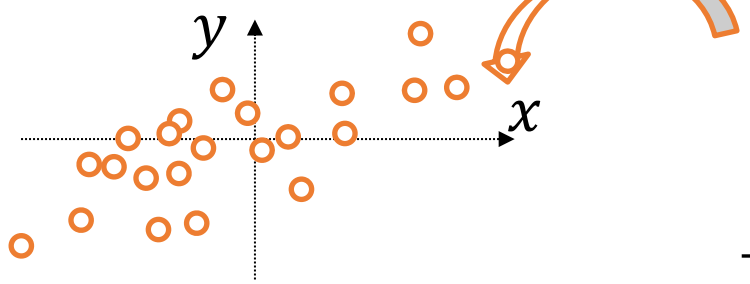


相関：二つの量の間の関係

- Case 1：無相関
  - x →大，y →特段の傾向無し
  - 要するに，xとyは無関係
  - 身長と数学の点数

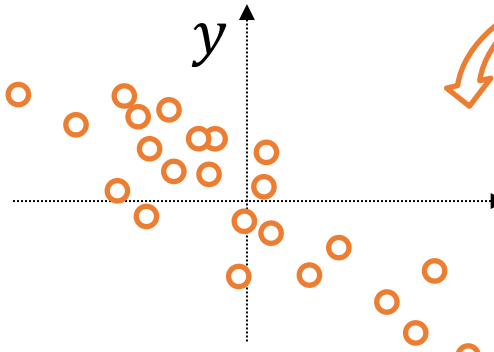


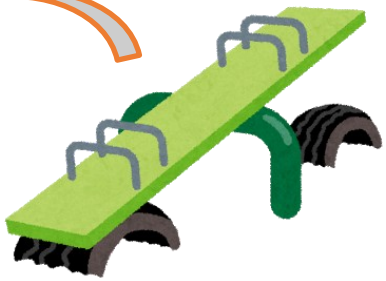
- Case 2：正の相関
  - x →大，y →大
  - 身長と体重



二人三脚的

- Case 3：負の相関
  - x →大，y →小
  - 身長とバレーボール攻撃失敗率

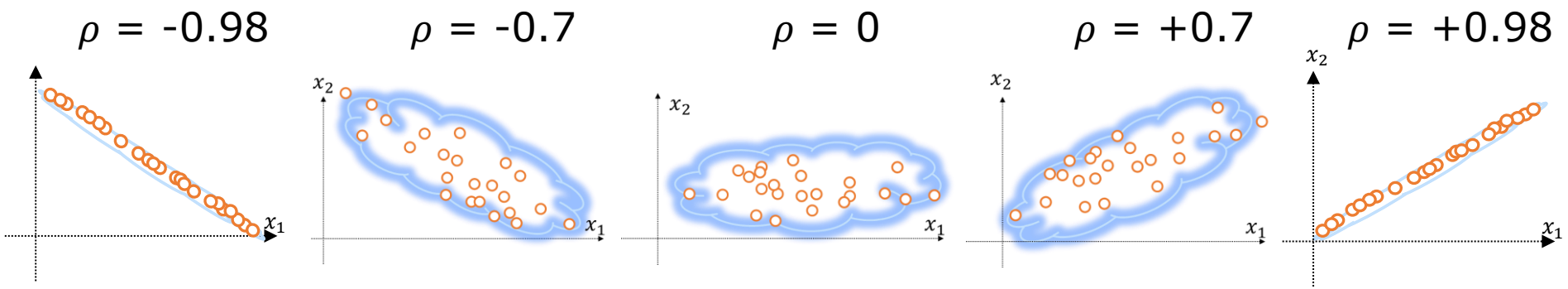




シーソー的

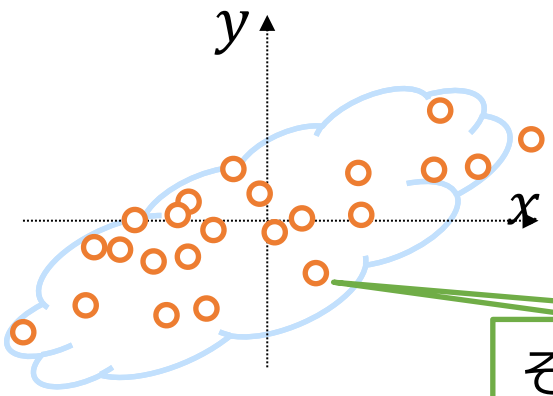
相関を数値で表現する(定量化する)

- **相関係数** ρ ＝ 相関の度合いを-1～+1の範囲の実数で表したもの
  - ρが負: 負の相関
  - ρが0: 無相関
  - ρが正: 正の相関



相関係数ρの定義

- xとyの相関を考える
- 簡単のためにxもyも平均ゼロとする ＝ 原点周りに散らばっている



それぞれx座標，y座標がある

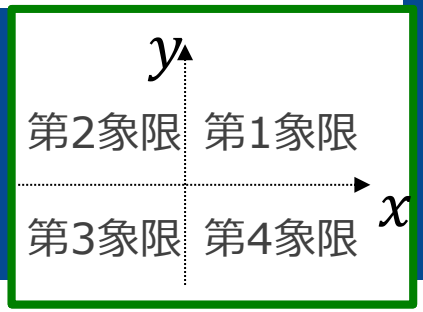
- この時，相関係数を(なるべく簡単な言葉で)定義すると…

$$\rho = \frac{(xy)\text{の平均値}}{\sqrt{x\text{の分散} \cdot y\text{の分散}}}$$

分子で符号が決まる

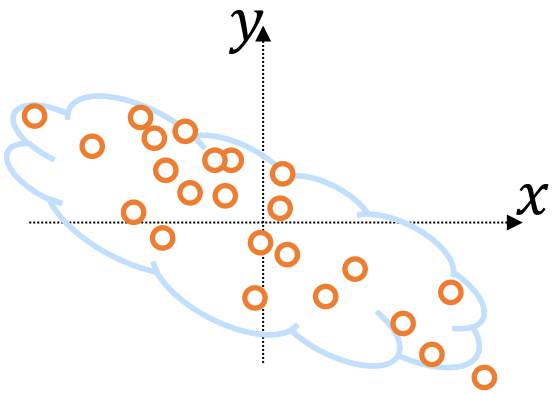
分母は正規化の役目
(大きさを調整するだけ)

相関係数ρの分子 ＝xyの平均値



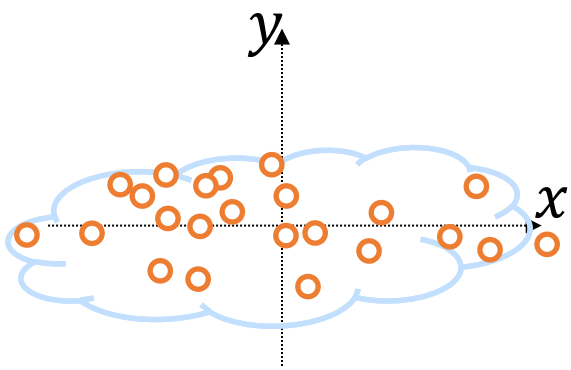
- この分子の意味を3つの場合で考えてみる

負の相関



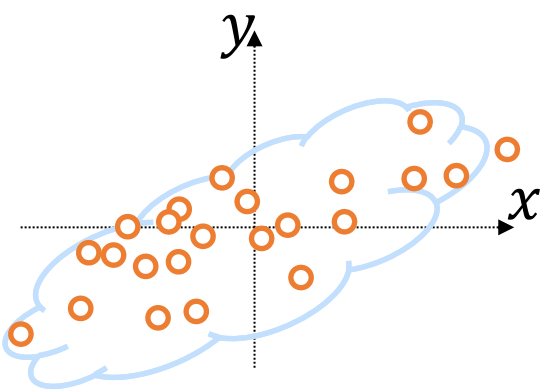
x と y の符号は
反対になり易い場合.
つまり xy の符号は
負になる場合が多数.
平均すると負になるはず.

無相関



相関が無い場合なので,
x によらず, y が
正の場合と負の場合は
同程度存在するはず.
つまり xy も正と負が
同程度存在し,
平均して 0 に近くなるはず.

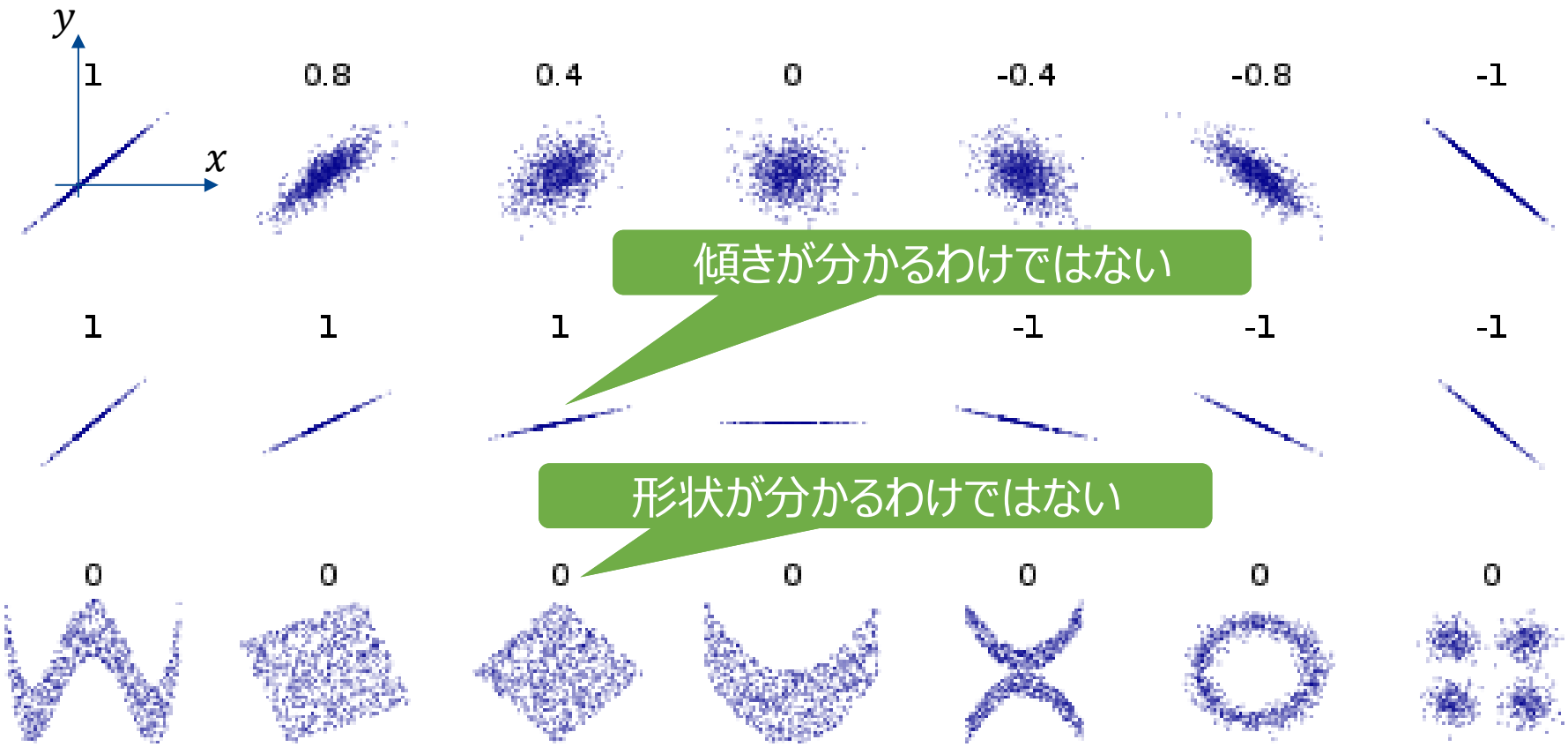
正の相関



x と y の符号は
同じになり易い場合.
つまり xy の符号は
正になる場合が多数.
平均すると正になるはず.

相関係数ρと分布の形

- 相関係数ρが分かれば，分布の形を少し想像することができる

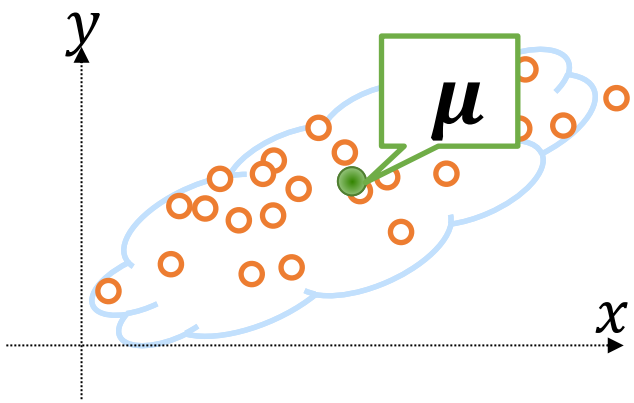


Wikipedia “相関係数”より

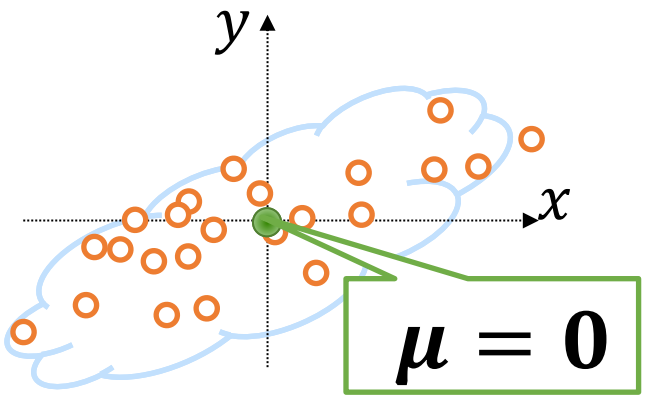
補足: データが原点周りに無い場合

- データを並行移動してから計算すればよい

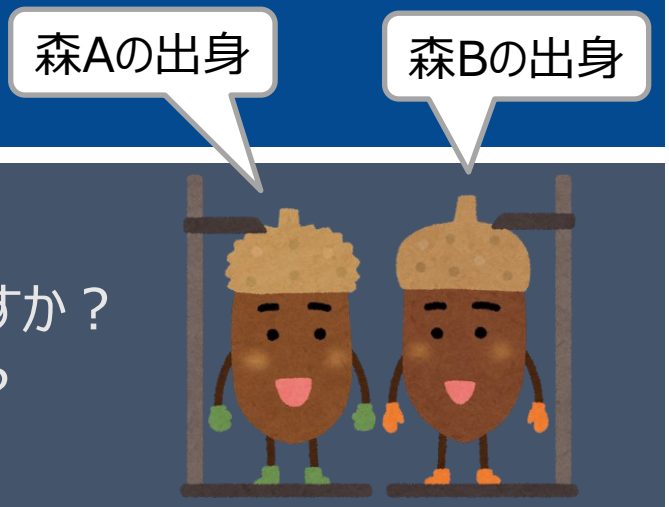
並行移動しても
データの広がり(相関)
には影響がない



x座標，y座標から
それぞれの平均を引き算



統計的検定



その差，本当に意味ありますか？
たまたまじゃないですか？

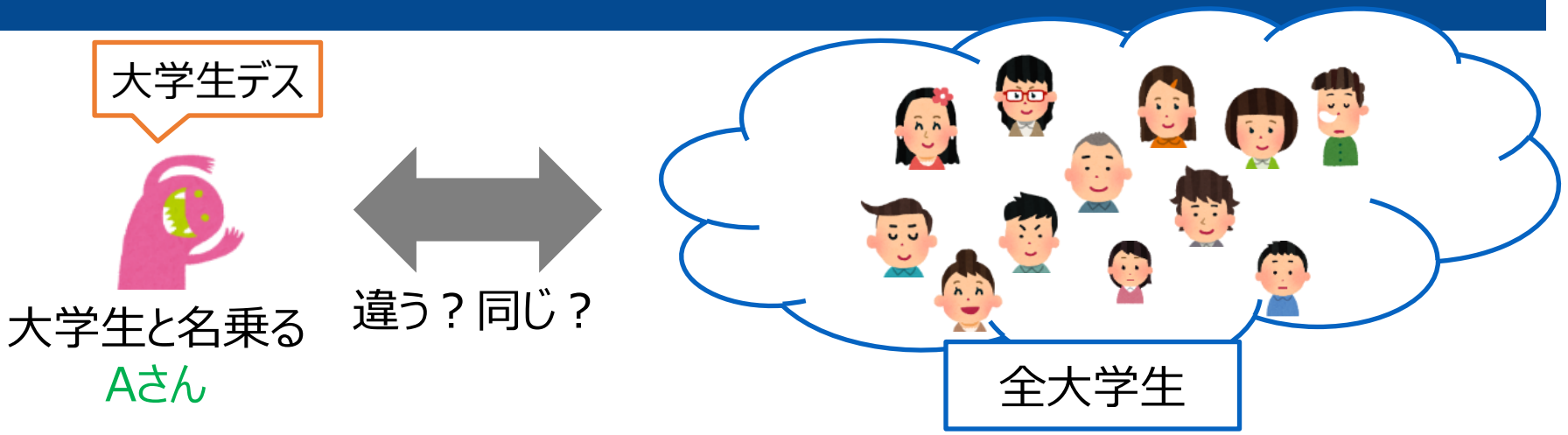
「統計的検定」を一言で！

統計的に差を評価する枠組み

- 統計的な評価
  - 観察して得られるデータは数量的に不十分なことが多い
  - データが限られた状況でも「差がある」ことを示したい

- 例
  - 東京と福岡で平均所得に差があるか？
  - あるトレーニングを行った前後で能力に差が出たか？
  - あるダイエット食品に効果があるか？
  - ダイエット食品と食べた群と食べなかった群に差があるか？
  - ある遺伝子がある病気の原因になるか？

あなたは本当に大学生？

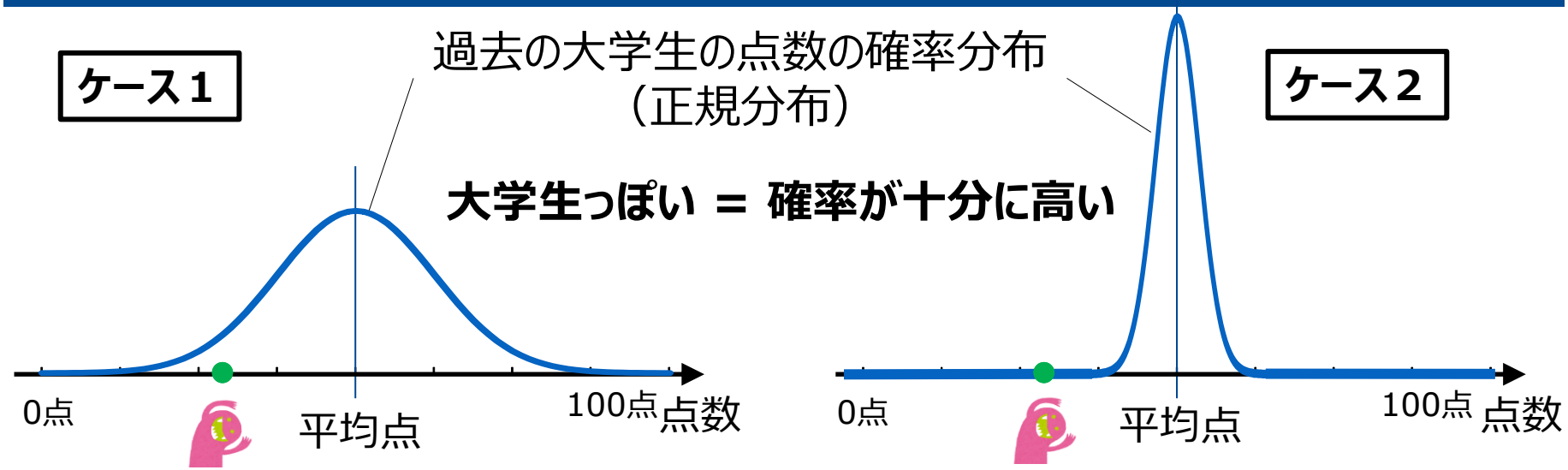


- Aさんの主張は本当だろうか？どの程度信頼できるだろうか？
- テストの点数で判断してみよう
  - 試しにAさんにあるテストを受けてもらう
  - 最近の大学1年生の平均点と比べて，大学生っぽい点数なら信じる

- では「大学生っぽい」とは?

➡ 統計的検定の出番

どんな点なら大学生っぽい？



- ケース1
  - 大学生として考えると，平均よりは低い
  - しかし，確率的には十分起こりそう

➡ 否定は難しい

- ケース2
  - ケース1と比べ確率は格段に低い
  - こんな点数をとる大学生が存在するだろうか？

➡ 否定しても問題なさそう

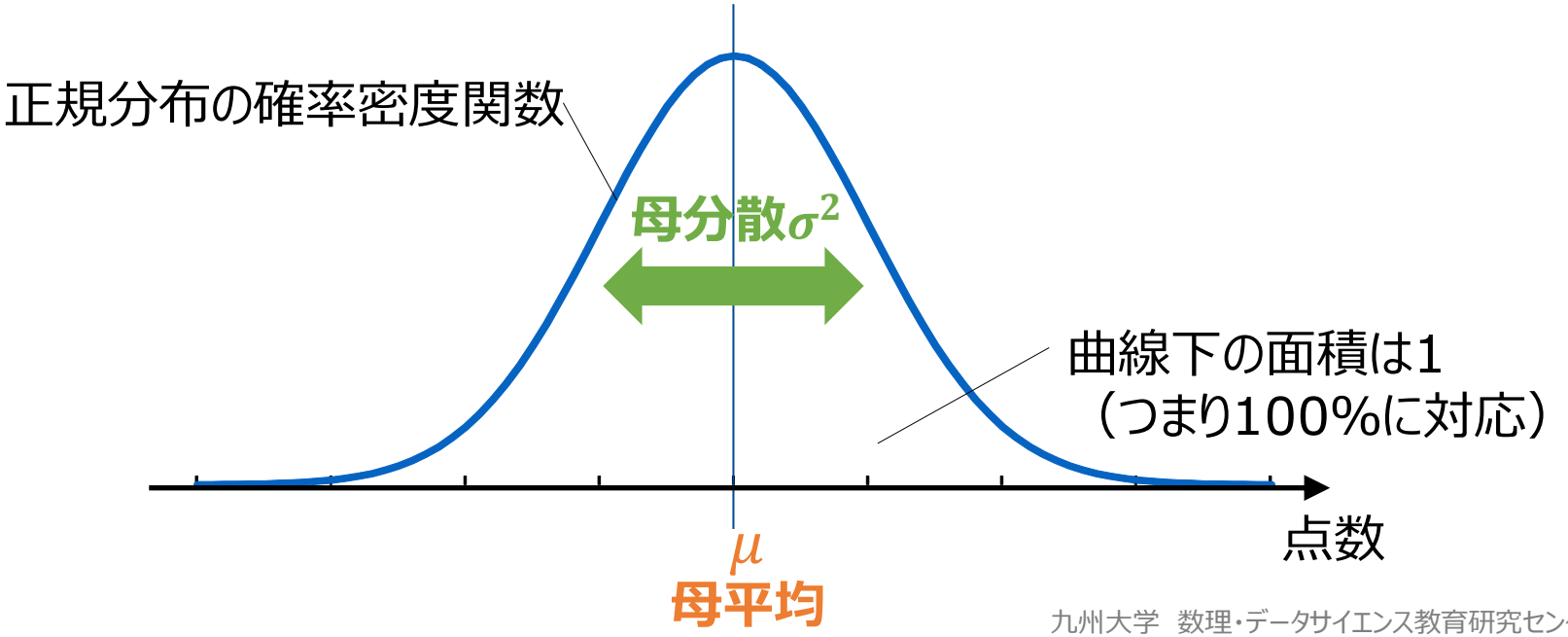
統計的検定の基本アイディア

帰無仮説の反対は**対立仮説**と呼ぶ

1. 「差がない」という仮説 (**帰無仮説**) を考える
   - Aさんは大学生である (Aさんと大学生に差がない)
2. 仮説を信じてみる
   - Aさんが大学生だと仮定する
3. 信じた上でそれが起きる確率を過去のデータから計算する
4. 基準とする確率 (**有意水準**) と比較する
   - 基準より低い
     - → 基準から考えるとあり得ないことが起きている
     - → そんなはずはない，仮定した仮説がそもそも間違ってたのでは?
     - → 帰無仮説を棄却する (大学生とは言えない)
   - 基準より高い
     - → 基準から考えると起きてもおかしくないことが起きている
     - → どこにも問題はなく，帰無仮説を棄却できない
     - → 大学生でないとは言い切れない (大学生とも言い切れない)

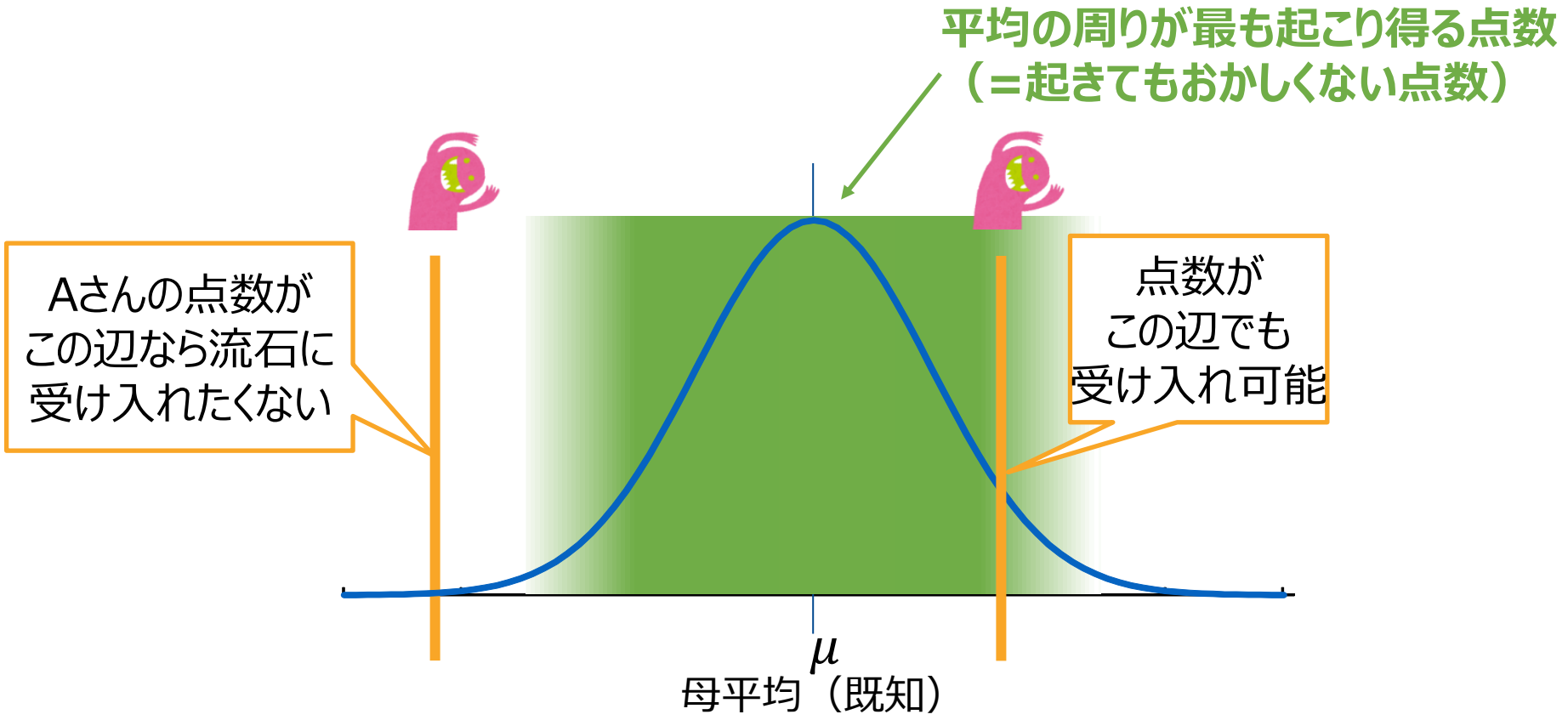
確率分布の仮定

- 母集団（過去のテストを受けた大学生達）の分布は（**母平均μ**，**母分散σ^2**）はわかっているとする
  - あまり現実にはないが，今回はシンプルな例ということで
- 成績の分布は正規分布と仮定する
  - データの種類に合わせて，適切な分布を使う



「よくあること」の範囲

- 平均からどのくらい離れているかを考える

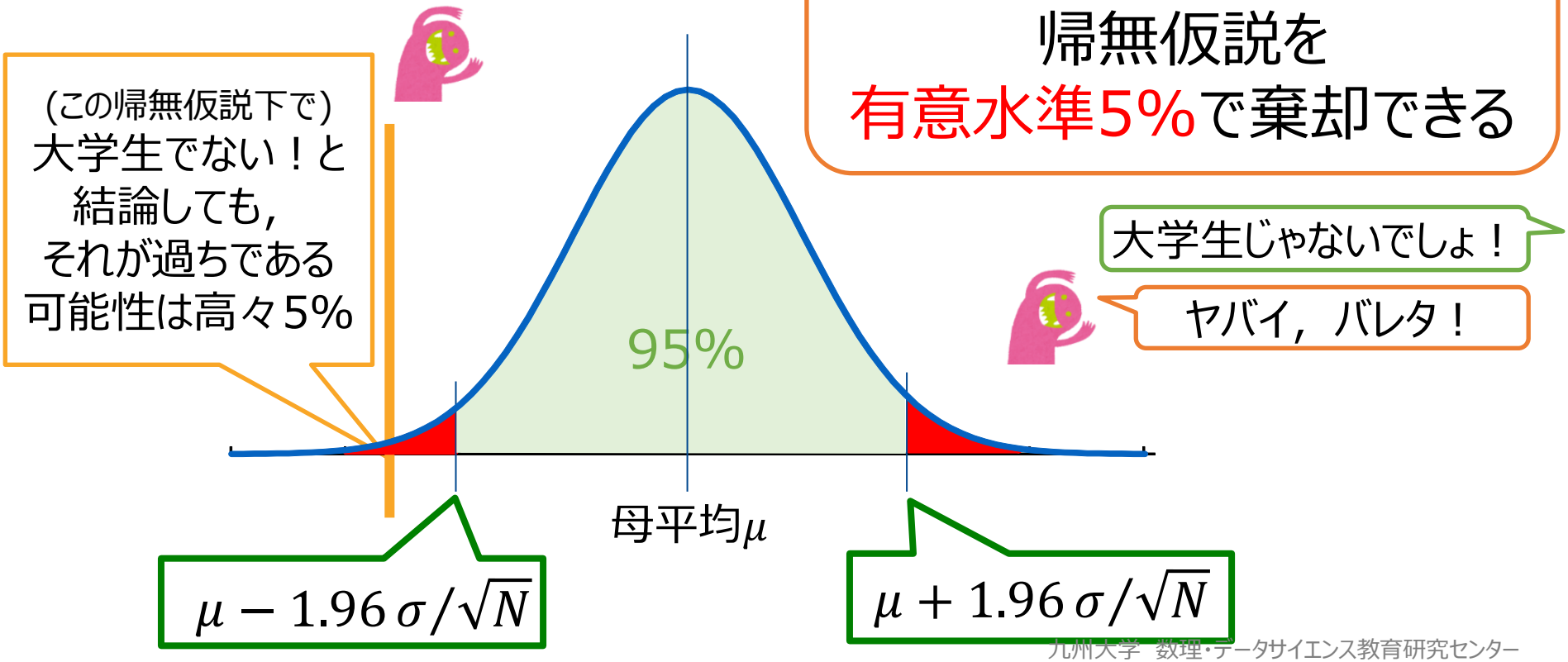


有意水準5%による判定

- 端っこだったら，仮説を棄却！
  - 間違って棄却してしまう可能性も少々あるが…

「大学生である」という
帰無仮説を
有意水準5%で棄却できる

(この帰無仮説下で)
大学生でない！と
結論しても，
それが過ちである
可能性は高々5%



有意水準1%による判定

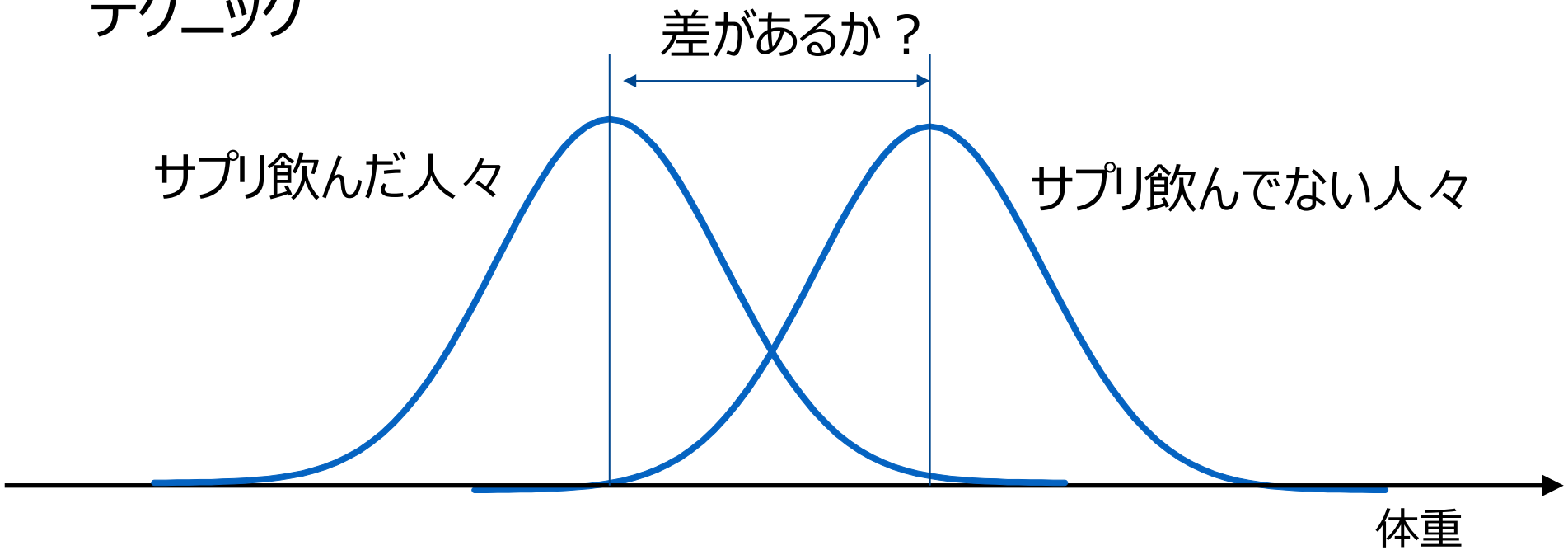
- もっと端っこなら，もっと自信をもって棄却できる

「大学生である」という
帰無仮説を
有意水準1%で棄却できる

大学生でない！と
結論しても，
それが過ちである
可能性は高々1%

この考えをさらに広げると，2つの集団に差があるかどうかの検定もできる

- 効果の有無等を検証する際に，極めてよく用いられるテクニック



まとめ

- データのもつ様々な性質
  - 分布
  - 平均，分散(広がり)
  - 相関

- 統計的検定
  - 確率に基づいて差の有無を論じる手法
  - 帰無仮説，対立仮説
  - データが得られる確率を評価