pd-3. 相関，相関係数，t 検定

（Python によるデータサイエンス演習）

URL: https://www.kkaneko.jp/ai/pd/index.html

金子邦彦

統計的分析の基礎：相関、標本、t検定

相関と相関係数

2変数間の関連性を−1から1の範囲で数値化し，正の相関・負の相関・相関なしを判断

母集団と標本

調査対象全体（母集団）から一部を選ぶサンプリングと，標本サイズの重要性

t検定とp値

2つの標本の平均値の差が偶然によるものか，統計的に有意かを判断する手法

意義

データ間の関係性分析や統計的仮説検定の実践力

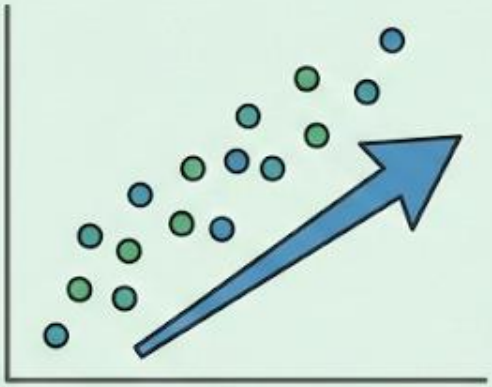
図解は Nano Banana Pro を用いて作成

相関（そうかん）とは

変数X → 関連性があるかを示す → 変数Y

Xが変化→Yも変化する傾向

正の相関（せいのそうかん）



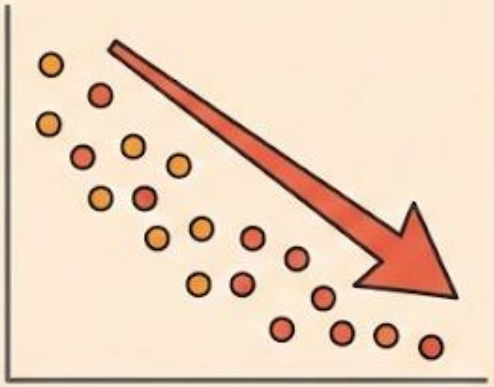
Xが増える→Yが増える傾向

勉強時間が増える

得点が上がる

例：勉強時間と得点

負の相関（ふのそうかん）



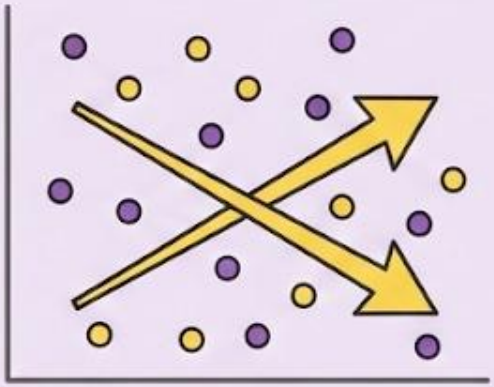
Xが増える→Yが減る傾向

ガソリン代が上がる

車の利用が減る

例：ガソリン代と車の利用

相関なし（そうかんなし）



XとYに関係がない

足のサイズ

勉強時間

例：足のサイズと勉強時間

図解は Nano Banana Pro を用いて作成

作成ツール：Nano Banana Pro

相関係数とは

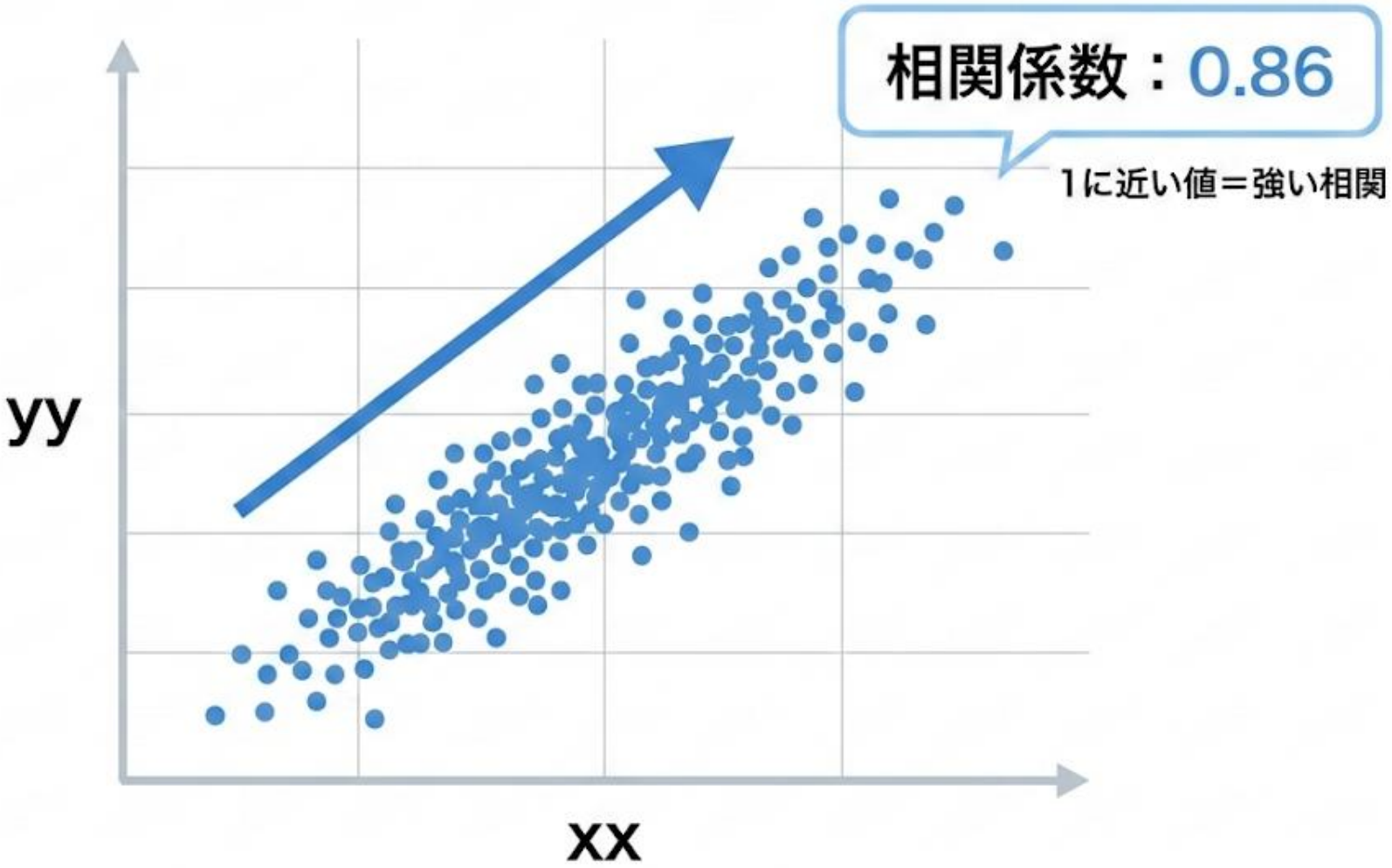
相関係数は、相関を数値化したものである。範囲は-1から1までである。

- 1に近い値：相関あり（正の相関）
- 0に近い値：相関なし
- -1に近い値：相関あり（負の相関）

相関係数を算出することで、変数間の関係を定量的に分析できる。

正の相関

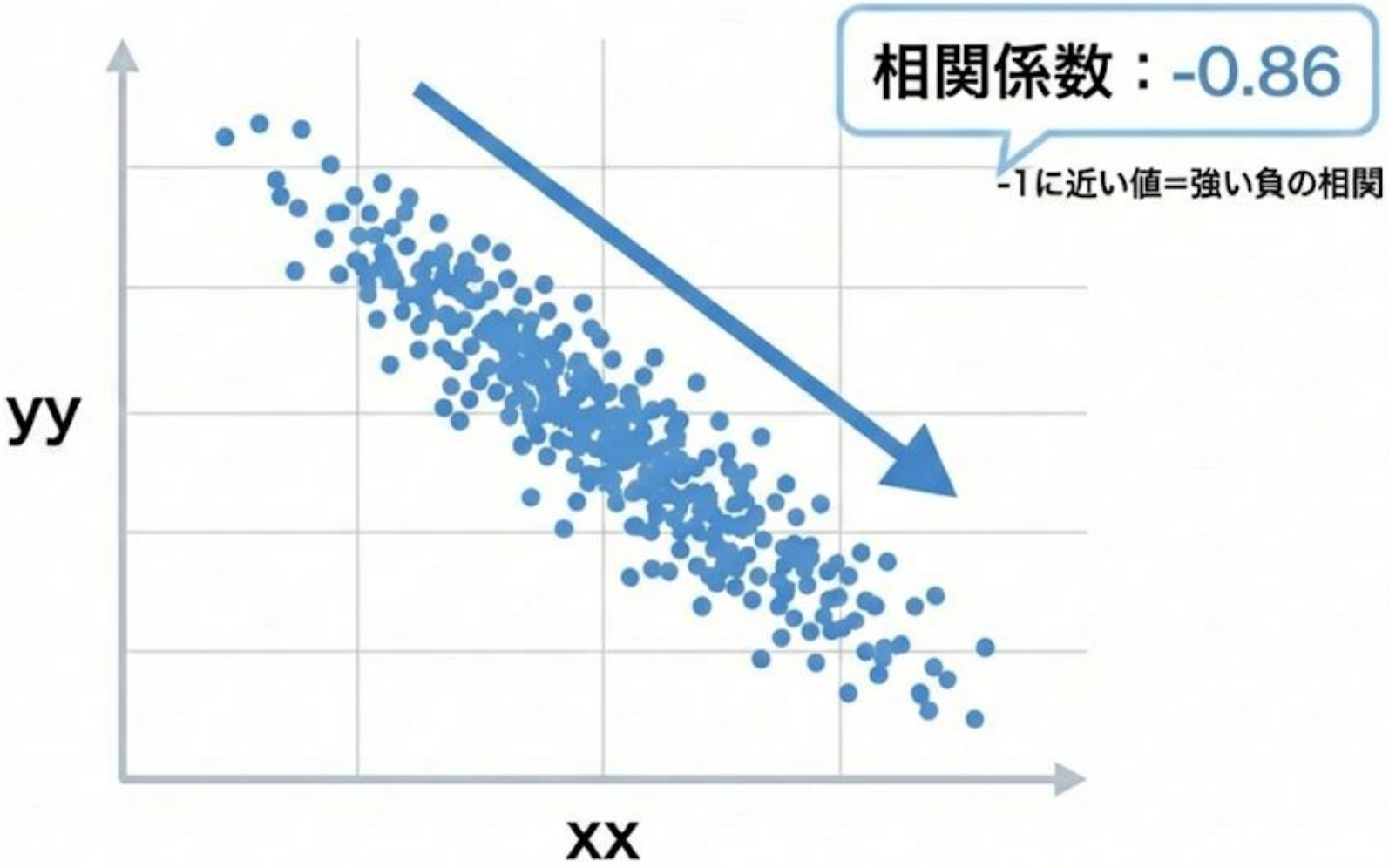
片方の値が増えると、もう片方も増える傾向



図解は Nano Banana Pro を用いて作成

負の相関

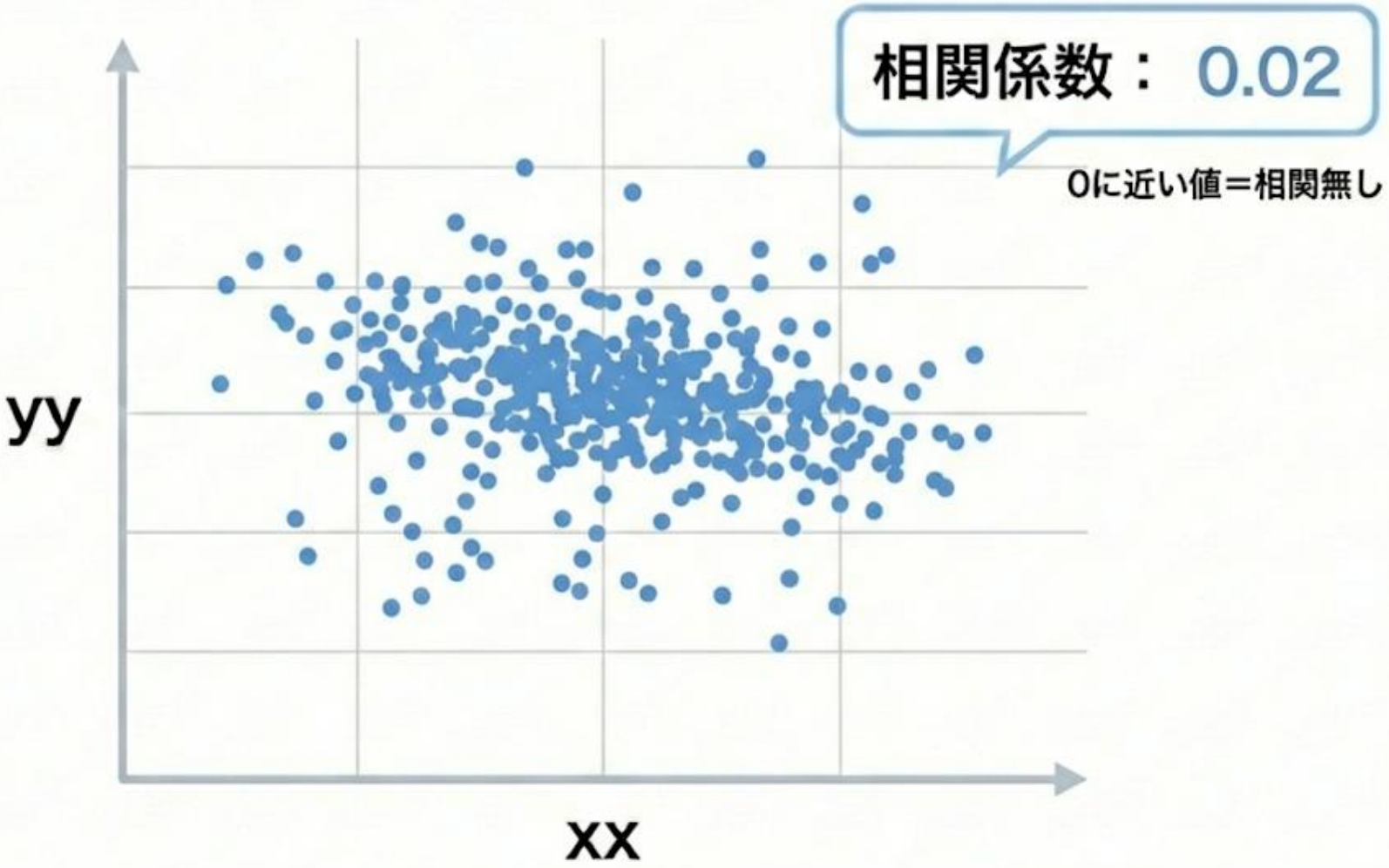
片方の値が増えると、もう片方も減える傾向



図解は Nano Banana Pro を用いて作成

相関無し

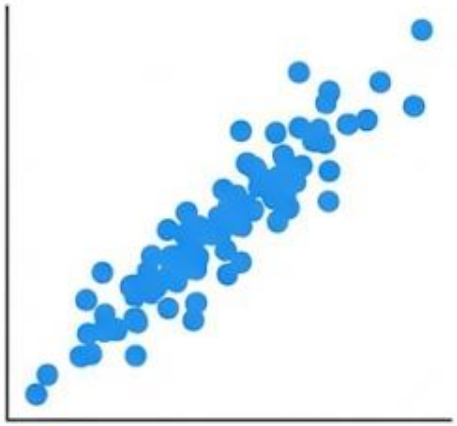
片方の値が変わっても、もう片方の値は変化しない傾向



図解は Nano Banana Pro を用いて作成

相関係数のまとめ

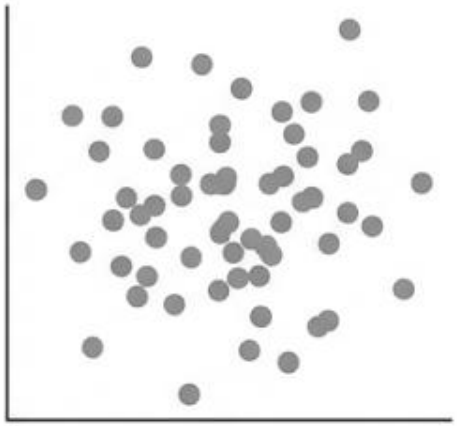
正の相関



1に近い値

0.86

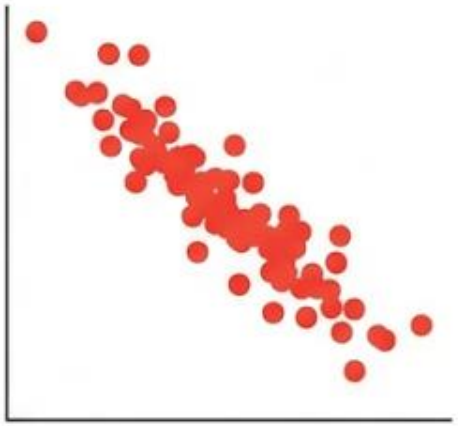
相関なし



0に近い値

0.12

負の相関



-1に近い値

-0.86

図解は Nano Banana Pro を用いて作成

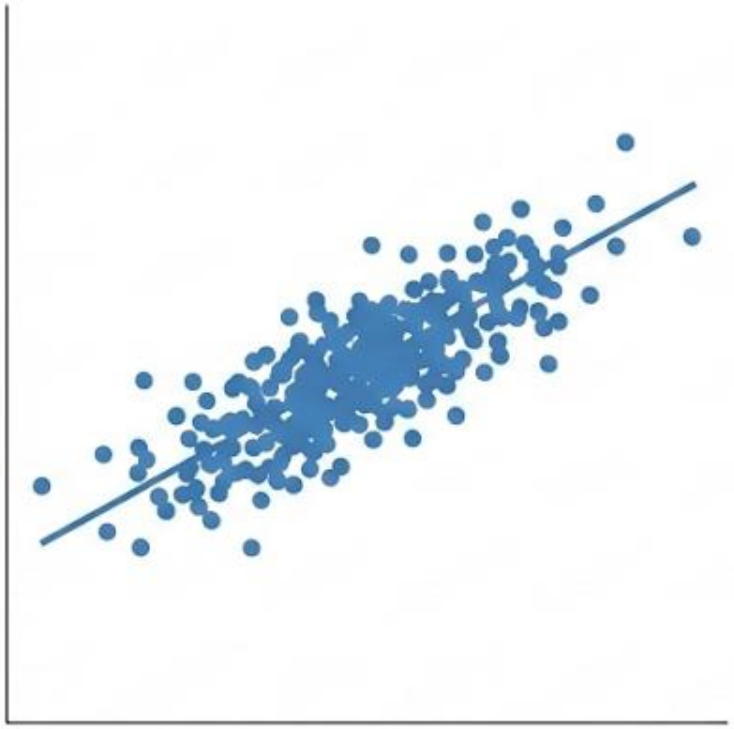
相関係数の活用例

相関係数は、2つの量の間の関係性を分析する際に活用される。

- 広告を増やすと、売上高が増えそうか
- 相関が高い複数の金融商品を扱うと、リスクが高いか
- 遺伝子と疾患に関係がありそうか

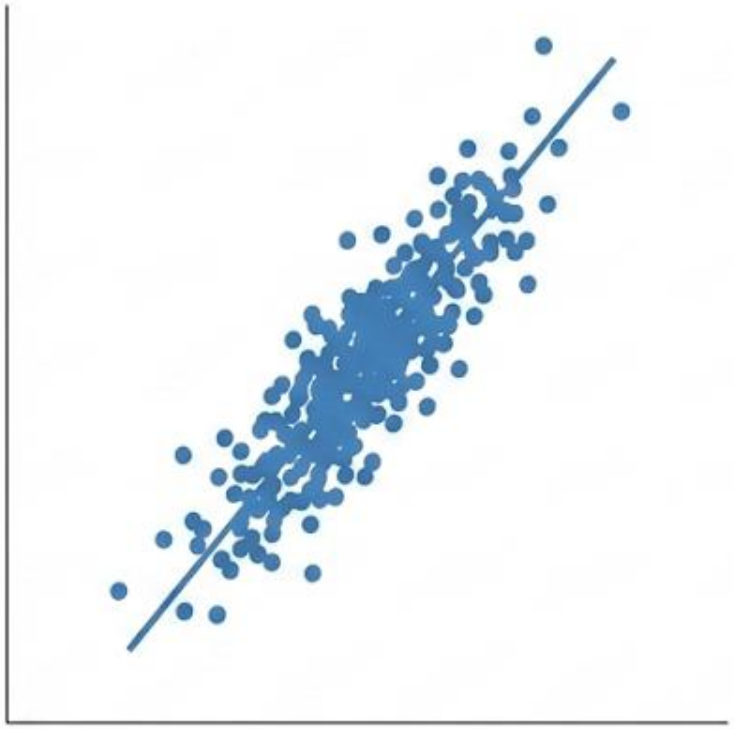
相関係数の性質：「傾き」ではなく「強弱」

緩やかな傾き



相関係数 ≈ 1（強い正の相関）

急な傾き

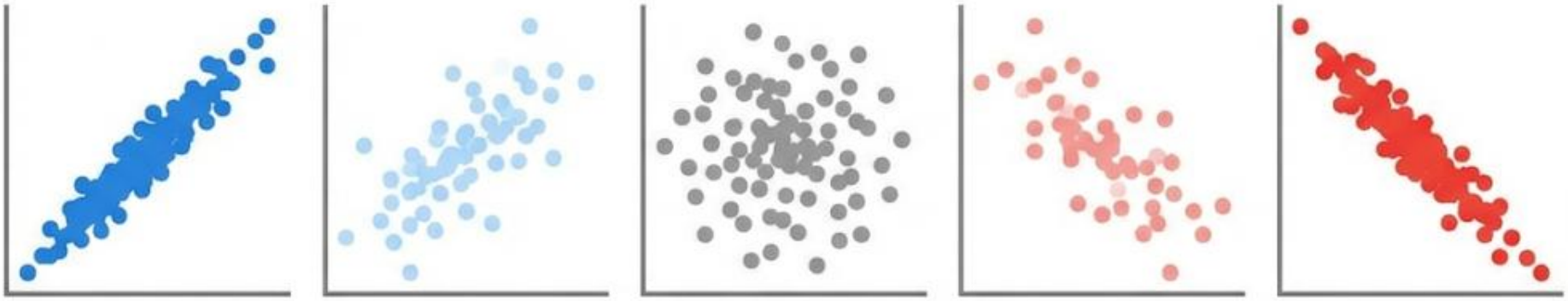


相関係数 ≈ 1（強い正の相関）

傾きが異なっても、点の密集度（相関の強弱）が同じなら、相関係数は同じ値になる。

図解は Nano Banana Pro を用いて作成

相関係数の例



図解は Nano Banana Pro を用いて作成

ここまでのまとめ（相関）

- 相関がある場合、一方が変化すると、もう一方も変化する傾向にある
- 1に近い値：**相関あり（正の相関）**
- 0に近い値：**相関なし**
- -1に近い値：**相関あり（負の相関）**

Pythonでの相関係数の算出

1. ライブラリの準備 (NumPy)

```
import numpy as np
```

2. データの用意 (身長と体重)

```
h = np.array([160, 170,
              160, 180, 170])
w = np.array([60, 70,
              50, 70, 60])
```

身長(h)	体重(w)
160	60
170	70
160	50
180	70
170	60

3. 相関係数の計算 (corrcoef関数)

```
print(np.corrcoef(h, w))
```

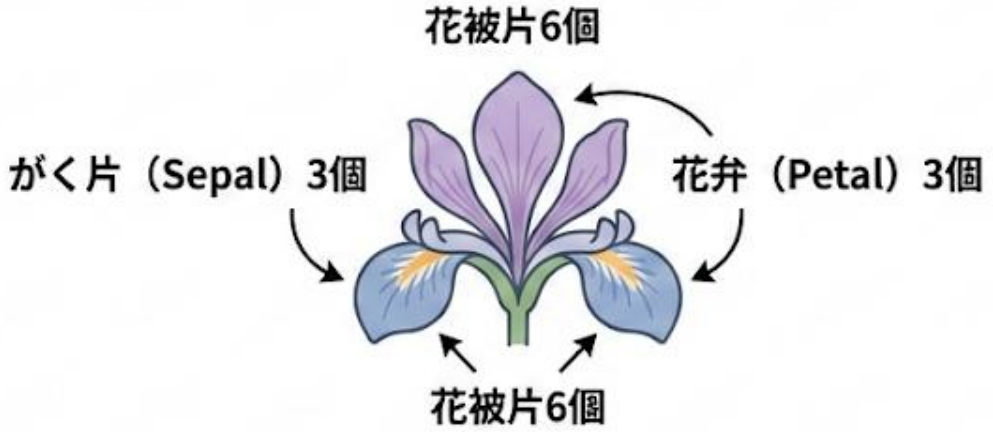
右上と左下が相関係数（同じ値）

図解は Nano Banana Pro を用いて作成

Irisデータセット概要

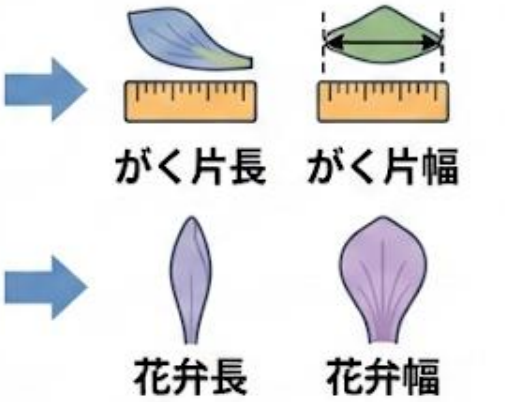
アヤメ属（Iris）

多年草



Irisデータセット

対象3種



がく片、花弁の幅・長さ計測

データ数：150
（50 × 3種）

1936年

作成者：Ronald Fisher

図解は Nano Banana Pro を用いて作成

Irisデータセット：特徴量と種類（ラベル）

`x`（特徴量データ：2次元配列）

がく片の長さ (cm)	がく片の幅 (cm)	花弁の長さ (cm)	花弁の幅 (cm)
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
6.4	3.2	4.5	1.5
5.1	3.0	1.6	0.3
4.7	3.0	4.6	1.5
⋮	⋮	⋮	⋮
6.4	3.2	4.5	1.5

各行＝1つのアヤメの測定値（4つの値）

`y`（ターゲットデータ：1次元配列）

0
0
1
2
0
1
2
⋮
...

0 = Setosa（種類A）

1 = Versicolor（種類B）

2 = Virginica（種類C）

各値＝xに対応する正解ラベル（アヤメの種類）

データセットの構成：特徴量（x）＋正解ラベル（y）

機械学習で分類を学習するためのペアデータ

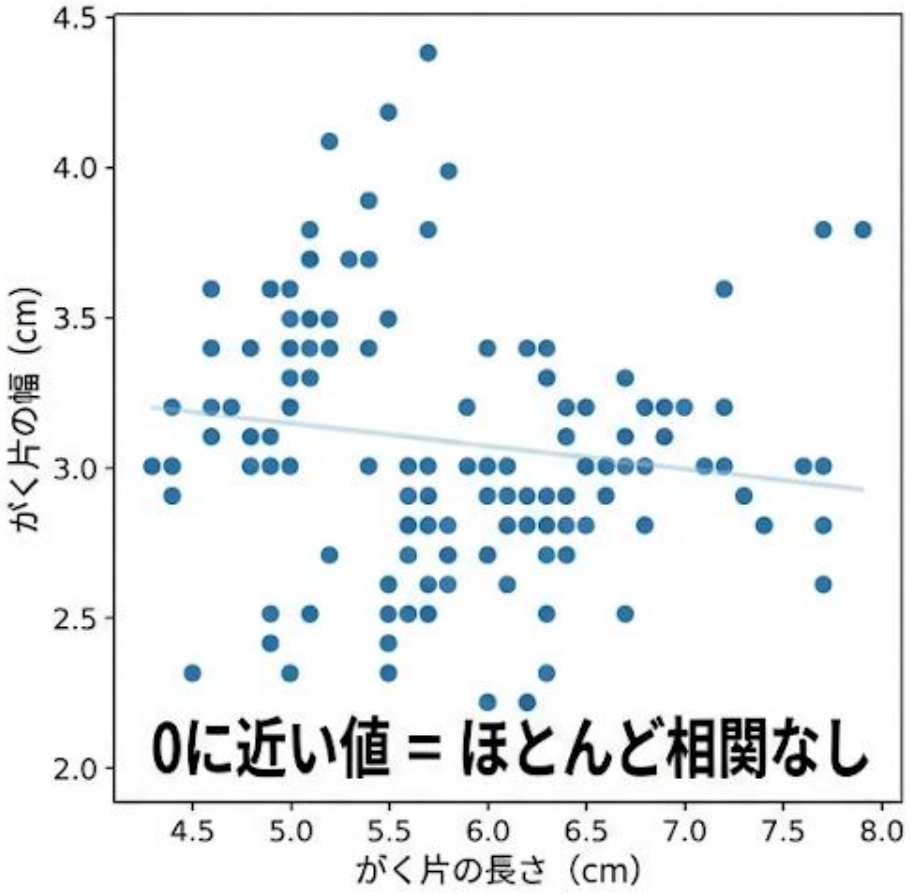
図解は Nano Banana Pro を用いて作成

Irisデータセット：がく片の長さと幅の相関関係

がく片の長さ
(Sepal Length, 列0)

がく片の幅
(Sepal Width, 列1)

アヤメの測定データから、2つの特徴量の関連性を調べる。



Pythonコード

```
from sklearn.datasets import load_iris
                              ← データの準備

iris = load_iris()
x = iris.data                 ← 特徴量の抽出

import numpy as np            ← 相関係数の計算
np.corrcoef(x[:, 0], x[:, 1])
```

結果：相関係数は約 -0.1

結論

がく片の長さと幅の間には、明確な直線的な関連性（相関）は見られない。データ点は散らばっている。

図解は Nano Banana Pro を用いて作成

母集団（ぼしゅうだん）とは？

調査・研究の対象となる**全体の集団**

把握と理解が不可欠

例1：

人類全体

例2：

20歳以上の人類全体

図解は Nano Banana Pro を用いて作成

サンプリングと標本

例：ランダムに1000名を選ぶ

例：全対象者を調べることが困難な場合

母集団から一部を選ぶこと

適切なサンプリングにより、標本から母集団の性質を推測する。

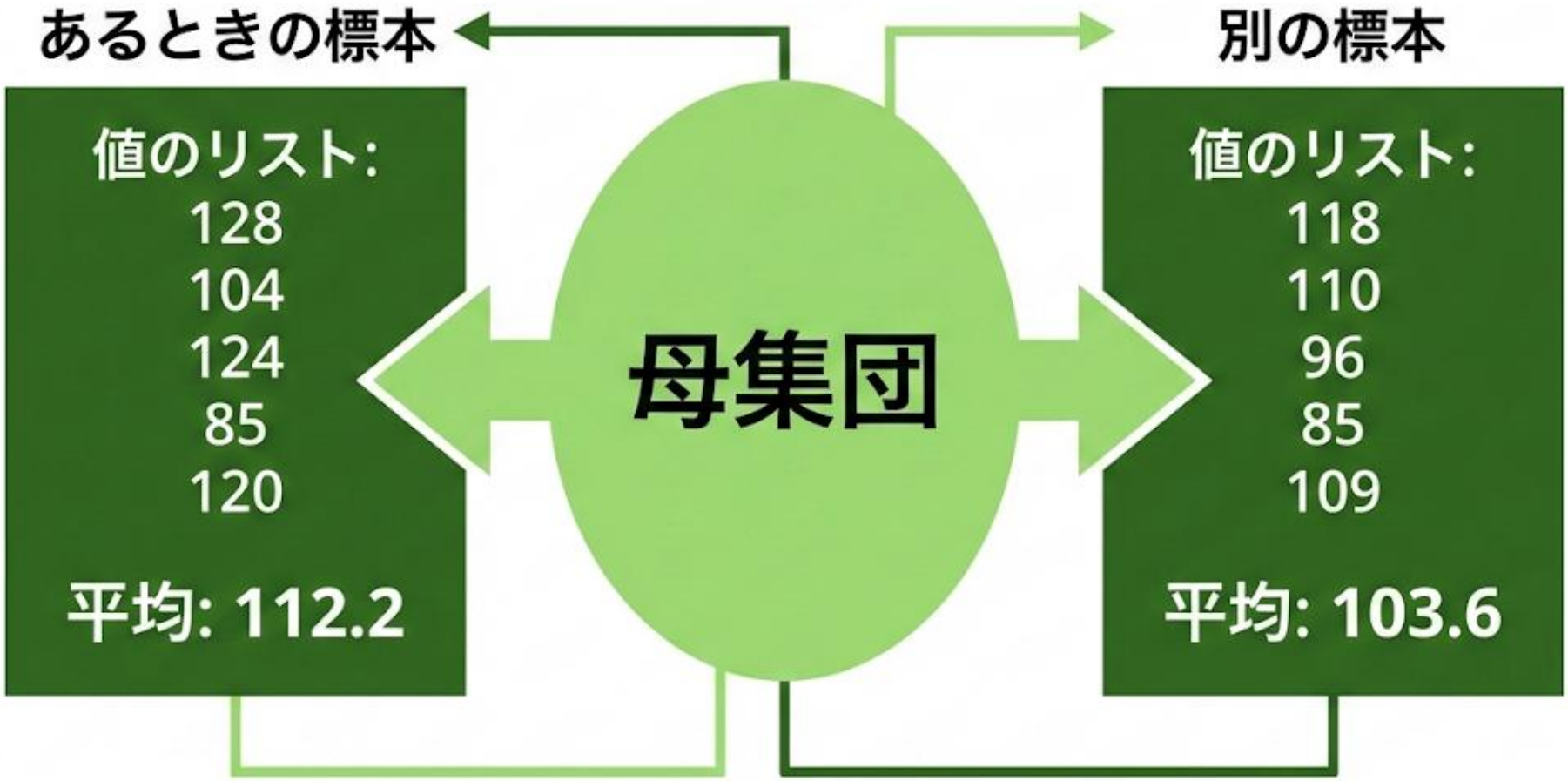
図解は Nano Banana Pro を用いて作成

十分な数の標本が必要

- 標本の大きさが小さいと、結果の信頼性が下がる
- 十分な数の標本を得ることが重要
- 標本の大きさの決定は簡単に決めることができない
- 母集団の特徴、調査や研究の目的によって、適切な標本サイズが異なる

標本による値の違い

選ばれた標本によって、値や平均が異なります。



図解は Nano Banana Pro を用いて作成

母集団と標本のまとめ

- 母集団：調査や研究の対象となる全体の集団
- サンプリング：母集団全体を調べることが困難な場合、母集団から一部を選ぶ。母集団の特徴や性質を推測することが可能となる。
- 標本：母集団からサンプリングで選ばれた母集団の一部。標本から得られたデータを分析し、母集団全体の性質や傾向を推測可能。

注意点：十分な標本サイズの確保が必要。ランダムに選択するなどの考慮が重要。

t検定とは

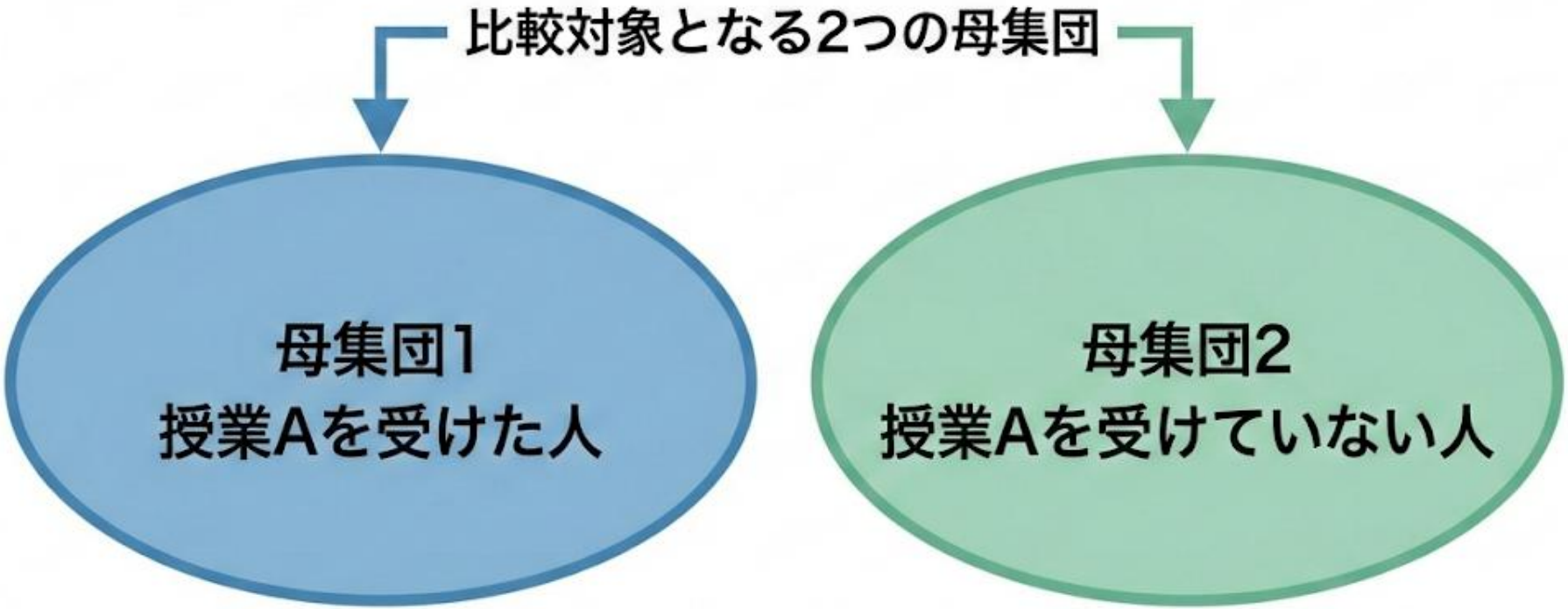
t検定は、2つの標本の平均値が統計的に有意に異なるかどうかを判断するための統計手法である。

注意点

- 標本が正規分布に従っていること
- 外れ値が存在する場合は、取り除いたり、適切に修正すること
- 十分な標本サイズを確保すること（小さな標本サイズでは結果の信頼性が下がる可能性がある）

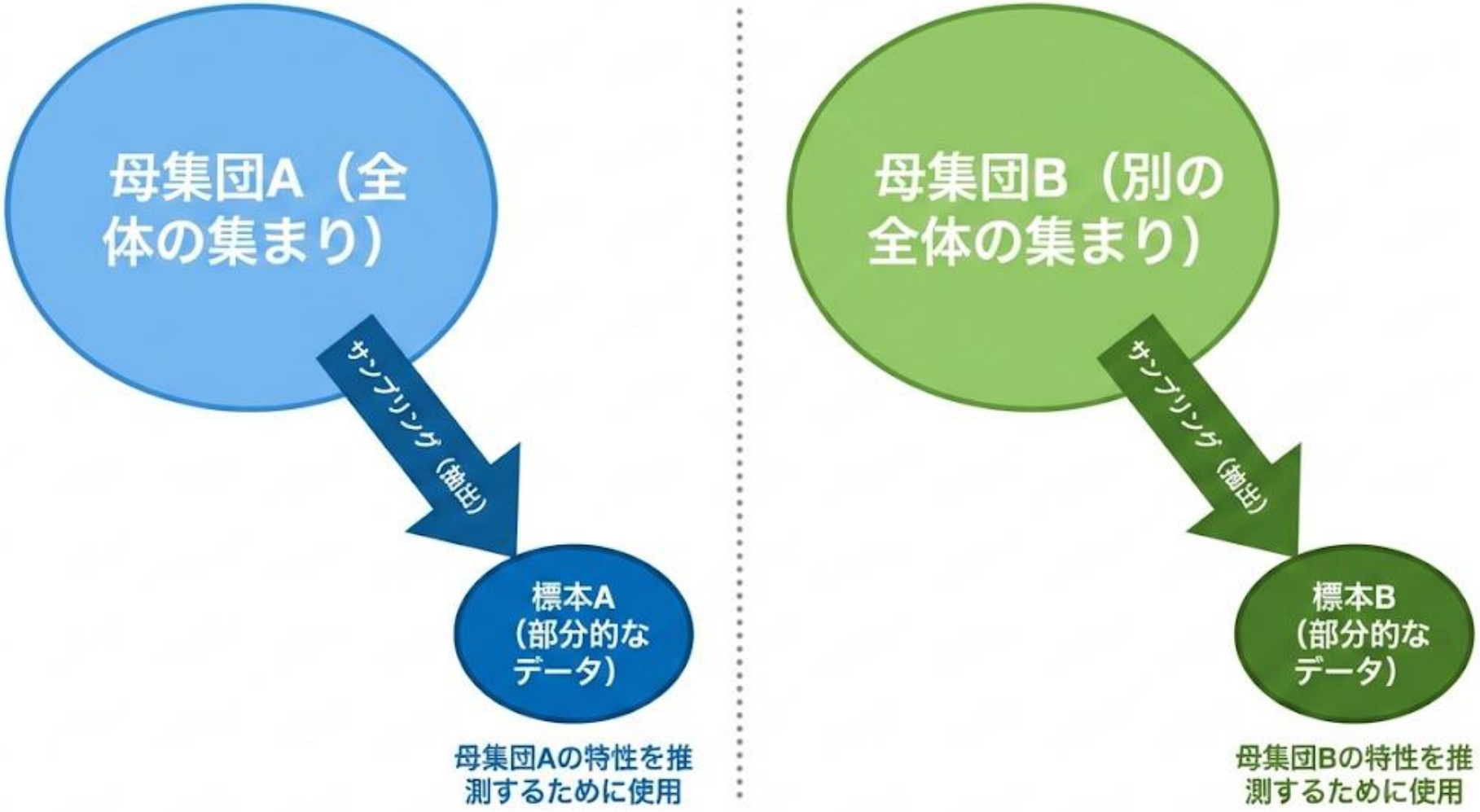
母集団が複数あるという考え方

例：授業Aの受講有無による比較



図解は Nano Banana Pro を用いて作成

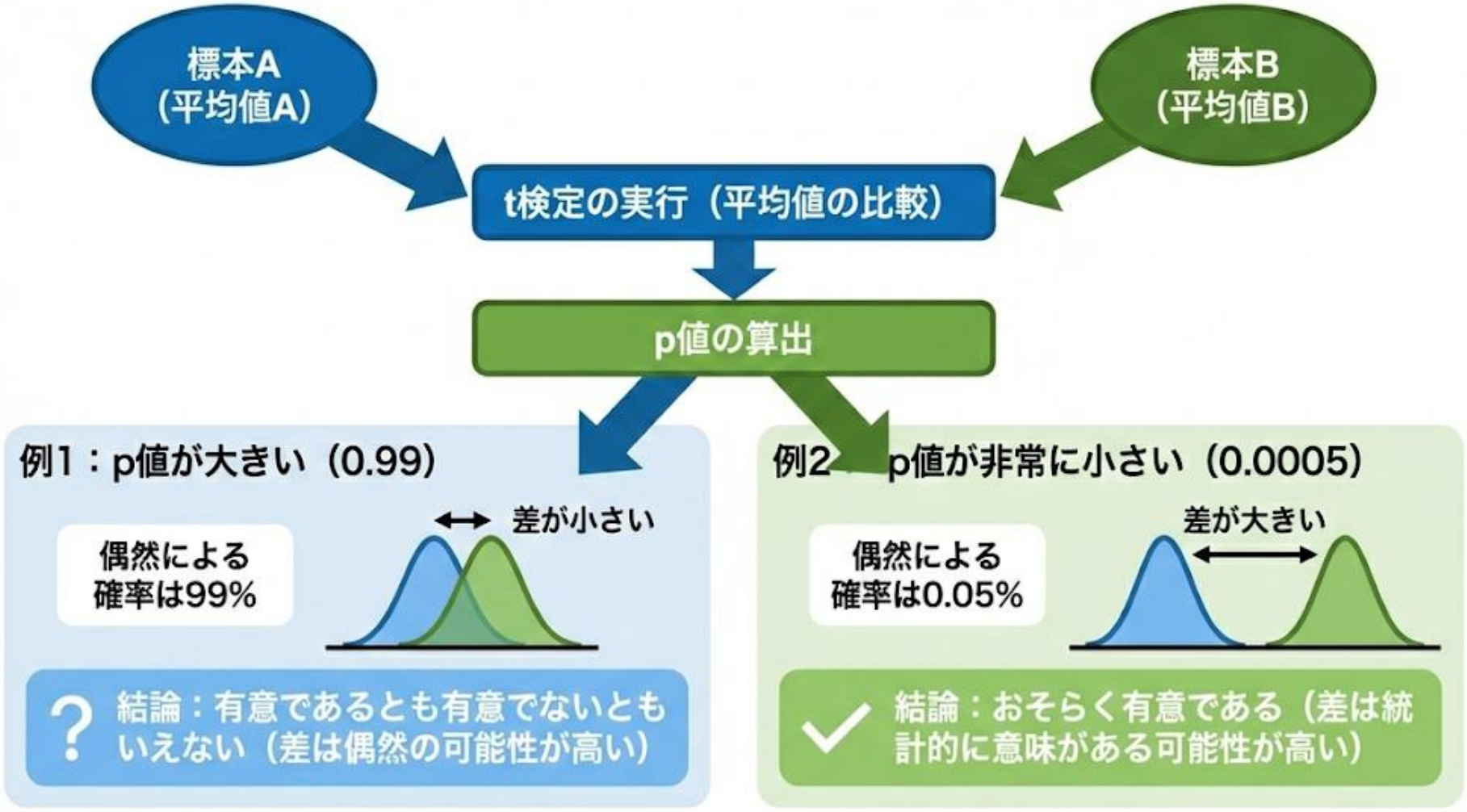
2つの母集団と2つの標本



図解は Nano Banana Pro を用いて作成

統計的推測：t検定とp値（平均値の比較）

2つの標本の平均値が統計的に有意に異なるかを判断するプロセス



p値は、観察された差が『偶然』によって生じる確率を示し、この確率が低いほど『有意な差』であると判断されます。

図解は Nano Banana Pro を用いて作成

p値と有意性：t検定による解釈の論理構成

p値は、2つの標本の差が「偶然」によって生じる確率を示す

p値の解釈（偶然の確率）

p値が小さい場合（例：< 0.05）

✓ 論理：とても偶然とは思えず、「有意である」と考える

t検定の例1（有意な差）

標本1：128, 104, 124, 85, 120

標本2：180, 191, 189, 131, 130, 150

→ t検定のp値 = 0.006908

✓ 結論：2つの標本の差が偶然による（有意でない）確率が低い。

p値が大きい場合（例：> 0.05）

？ 論理：偶然であるとも、偶然でないとも言えないと考える

t検定の例2（有意とは言えない差）

標本1：128, 104, 124, 85, 120

標本2：100, 106, 89, 89, 105

→ t検定のp値 = 0.1541

？ 結論：有意であるとも有意でないとも言えない。

このように、p値の大小によって、観測された差が統計的に意味のある「有意な差」であるかどうかの判断が分かれます。

図解は Nano Banana Pro を用いて作成

t検定のまとめ

t検定

- 2つの標本の平均値の統計的な有意性を判断する統計手法
- 標本が正規分布に従い、外れ値を適切に扱い、十分な標本サイズを確保することが重要

t検定のp値

- 2つの標本の差が偶然である確率
- p値が低いとき、差が統計的に有意である可能性が高まる

p値の解釈

- p値が小さいとき：「差は統計的に有意であり、偶然とは考えにくい」
- p値が大きいとき：「差は統計的に有意であるとは言いきれない。偶然であるとも、偶然でないとも言えない」

Pythonによるt検定

① ライブラリのインポート

```
from scipy import stats
```

② データの準備

```
sample1 = [86, 93, 91, 89, 92]
sample2 = [77, 81, 84, 85, 80]
```

③ t検定の実施（ttest_ind）

```
t, p = stats.ttest_ind(sample1, sample2)
```

④ p値の表示

```
print(p)
```

```
from scipy import stats
sample1 = [86, 93, 91, 89, 92]
sample2 = [77, 81, 84, 85, 80]
t, p = stats.ttest_ind(sample1, sample2)
print(p)
```

```
0.001670538621824343
```

図解は Nano Banana Pro を用いて作成

演習

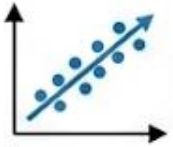
1. サイズが5以上の数値データを、2個準備しなさい

2. 1のデータについてt検定を行い、そのp値を求めなさい

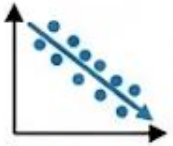
統計学の基礎：相関、母集団、t検定

1. 相関と相関係数

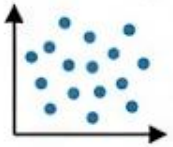
2変数の関連性の指標



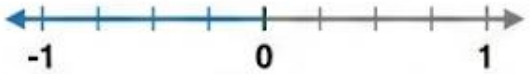
正の相関
(例：勉強時間↑ 得点↑)
一方が増えると他方も増える



負の相関
(例：ガソリン代↑ 車の利用↓)
一方が増えると他方は減る



相関なし
(例：足のサイズ、勉強時間)
関係がない

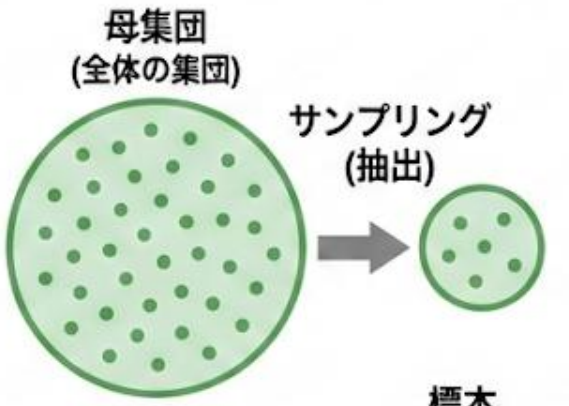


相関係数: -1 (負) ← 0 (なし) → 1 (正)
※「強弱」の尺度。「傾き」ではない。

Pythonでの算出:
```
import numpy as np
np.corrcoef(x, y)
```

2. 母集団と標本

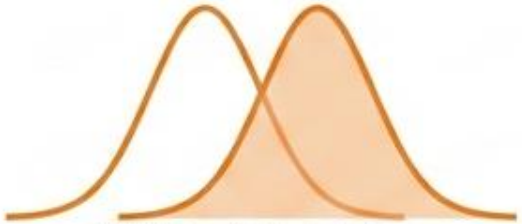
調査・研究の対象



- 注意点
  - 標本サイズ小 → 信頼性低下
  - 十分なサイズが必要
  - ランダムな選択が重要

3. t検定

2標本平均の有意差判断



平均値の差は有意か？

- 前提条件
  - 正規分布
  - 外れ値処理
  - 十分なサイズ

p値 (偶然の確率)

小さい (目安 < 0.05) → 有意な可能性高
大きい → 有意・有意でないと言えない

Pythonでの実施:
```
from scipy import stats
stats.ttest_ind(group1, group2)
```


図解は Nano Banana Pro を用いて作成