

# 数字が真実を語るとは限らない

AI時代のデータ探偵術：探索的データ分析（EDA）入門

AI時代のデータ探偵術：探索的データ分析（EDA）入門

# 統計的に「同一」な、4つのデータセット

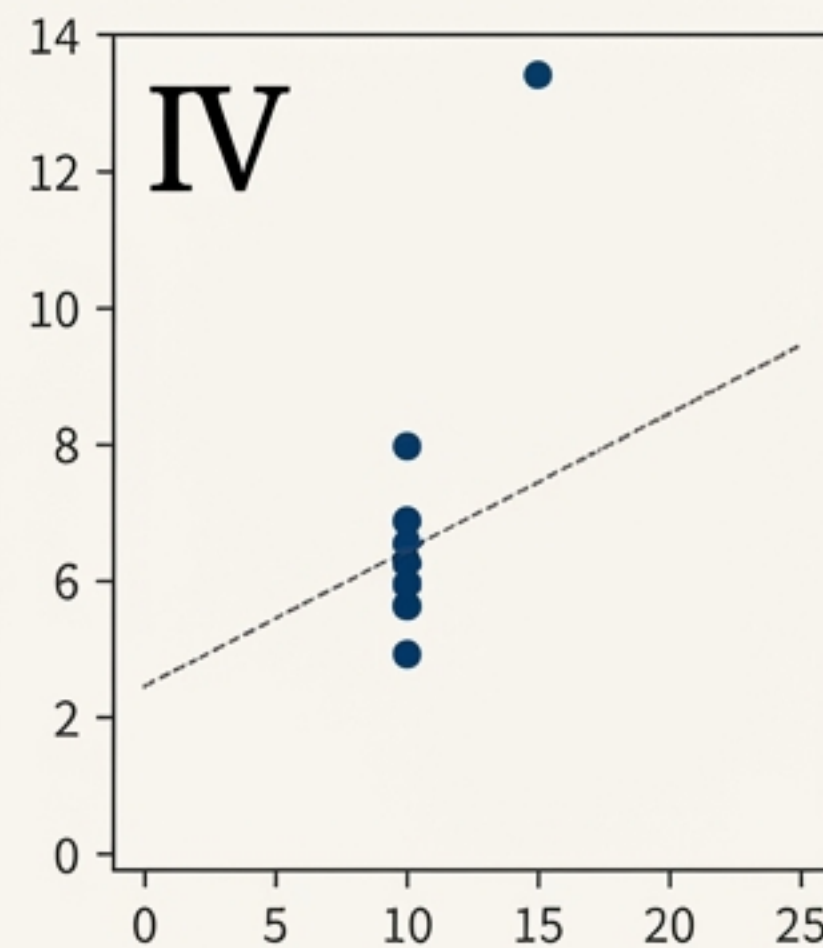
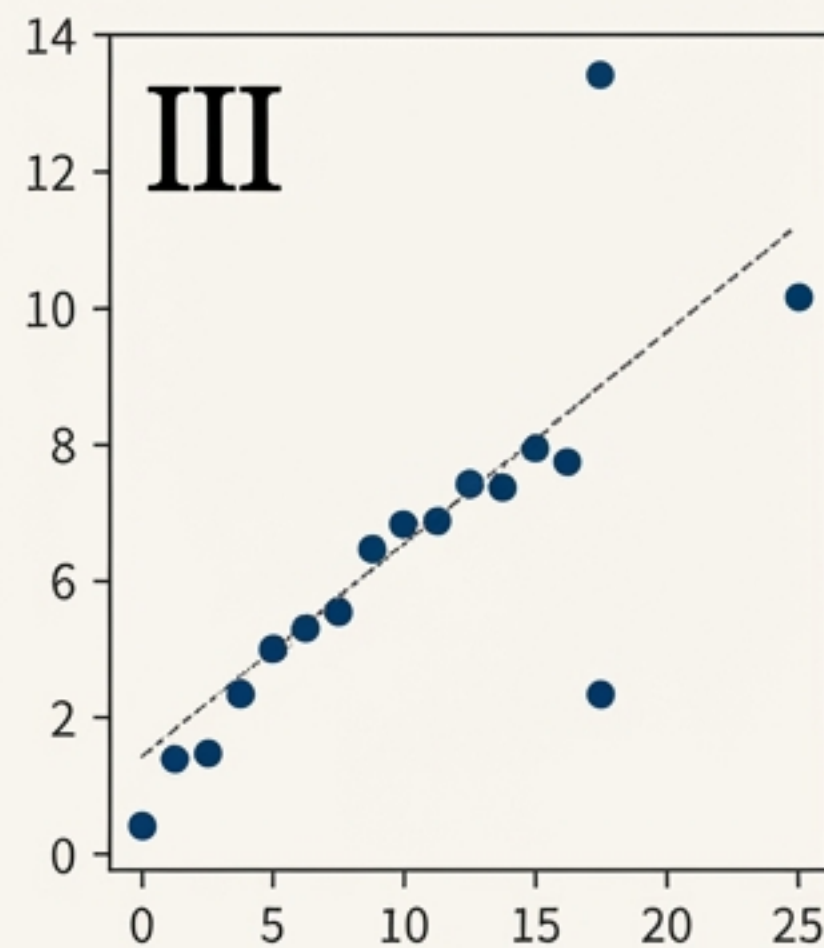
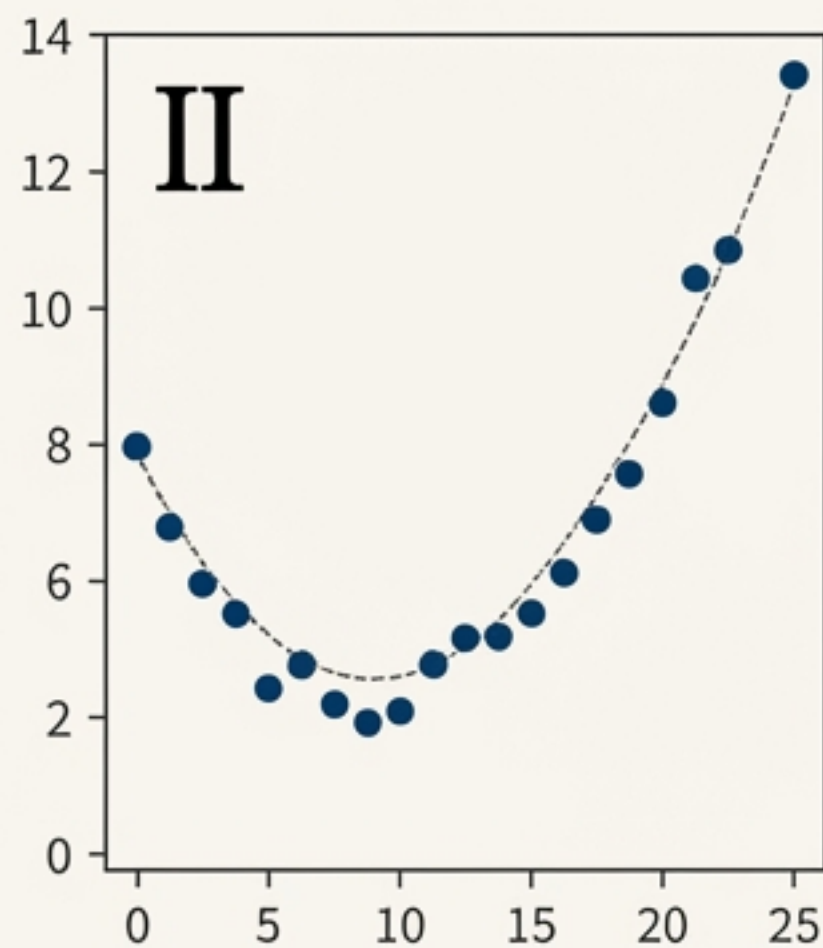
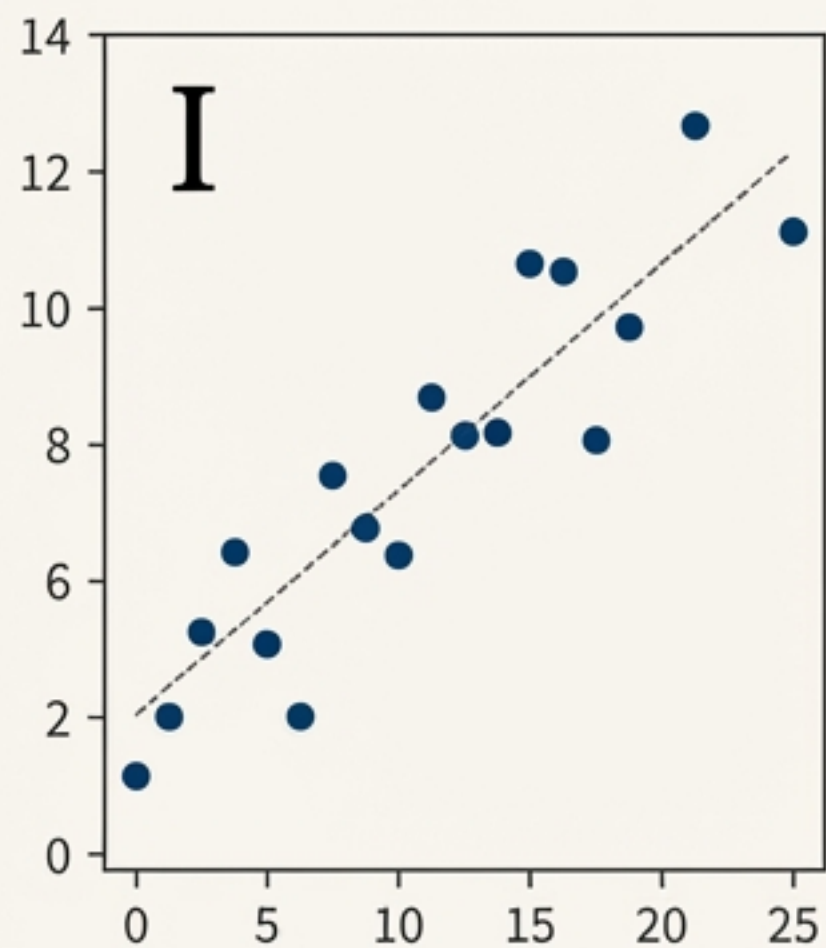
| アンスコム の例 (Anscombe's Quartet) |                  |                  |                  |                  |
|-------------------------------|------------------|------------------|------------------|------------------|
| 特性                            | データセット I         | データセット II        | データセット III       | データセット IV        |
| 平均値 (x)                       | 9.0              | 9.0              | 9.0              | 9.0              |
| 分散 (x)                        | 11.0             | 11.0             | 11.0             | 11.0             |
| 平均値 (y)                       | 7.50             | 7.50             | 7.50             | 7.50             |
| 分散 (y)                        | 4.12             | 4.12             | 4.12             | 4.12             |
| 相関係数                          | 0.816            | 0.816            | 0.816            | 0.816            |
| 回帰直線                          | $y = 3.0 + 0.5x$ | $y = 3.0 + 0.5x$ | $y = 3.0 + 0.5x$ | $y = 3.0 + 0.5x$ |

4つのデータセットI, II, III, IVは、以下の点で完全に一致します。

- 平均値 (Mean)
- 分散 (Variance)
- 相関係数 (Correlation Coefficient)
- 回帰直線 (Regression Line)

数値だけを見れば、これらは「同じ性質のデータ」です。しかし、本当にそうでしょうか？

# しかし、その「真の姿」は全く異なる



I: 綺麗な右肩上がりの直線

II: 放物線を描く曲線

III: ほぼ直線だが、一つの外れ値が存在

IV: Xの値がほぼ同じで、一つの外れ値が存在

**数値要約はデータを「丸める」行為であり、  
可視化はデータの「個性を復元する」行為です。**

# データと「対話」するための哲学

探索的データ分析 (Exploratory Data Analysis: EDA)

提唱者：ジョン・テューキー (John Tukey)、米国の統計学者

## 従来の分析 (CDA)

「立てた仮説が正しいか」を確認する  
確証的データ分析。



## EDAの哲学

「そもそもデータの中に何があるのか」  
を探る、刑事の捜査のようなプロセス。



計算の前に、まず自分の目で観察すること。  
それがすべての始まりです。

# 事件の全体像を掴む：データの「かたち」を暴く

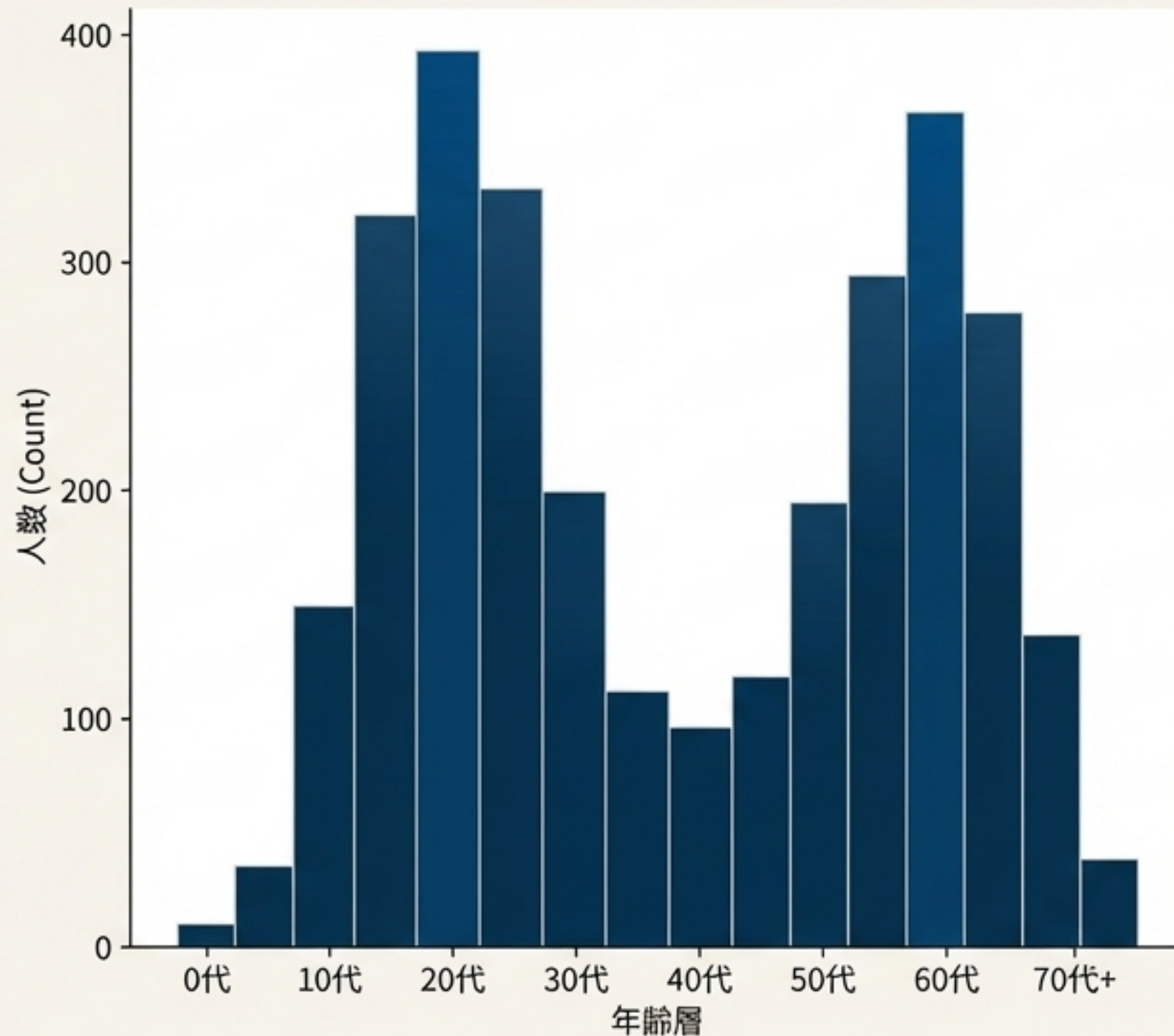
## ヒストグラム (Histogram)

データ全体の「分布の形」を把握する。特に「山」がいくつあるかに注目する。

ある観光地の来訪者平均年齢が「40歳」だったとします。

しかし、ヒストグラムを描くと、20代の山と60代の山の「二峰性」かもしれません。

**The Flaw:** 「平均40歳向けの施策」は、誰にも響かない可能性があります。



# 異質な目撃者を探す：「外れ値」という宝の山

**Tool: 箱ひげ図 (Box Plot)**

**Purpose:** 複数のグループの分布を比較し、「外れ値」を発見する。

## A Detective's Mindset Shift

**統計学では (In statistics):** 外れ値はしばしば「ノイズ」として扱われる。

**EDAでは (In EDA):** 外れ値こそが「宝の山」。その背後には必ず固有のストーリーが存在する。

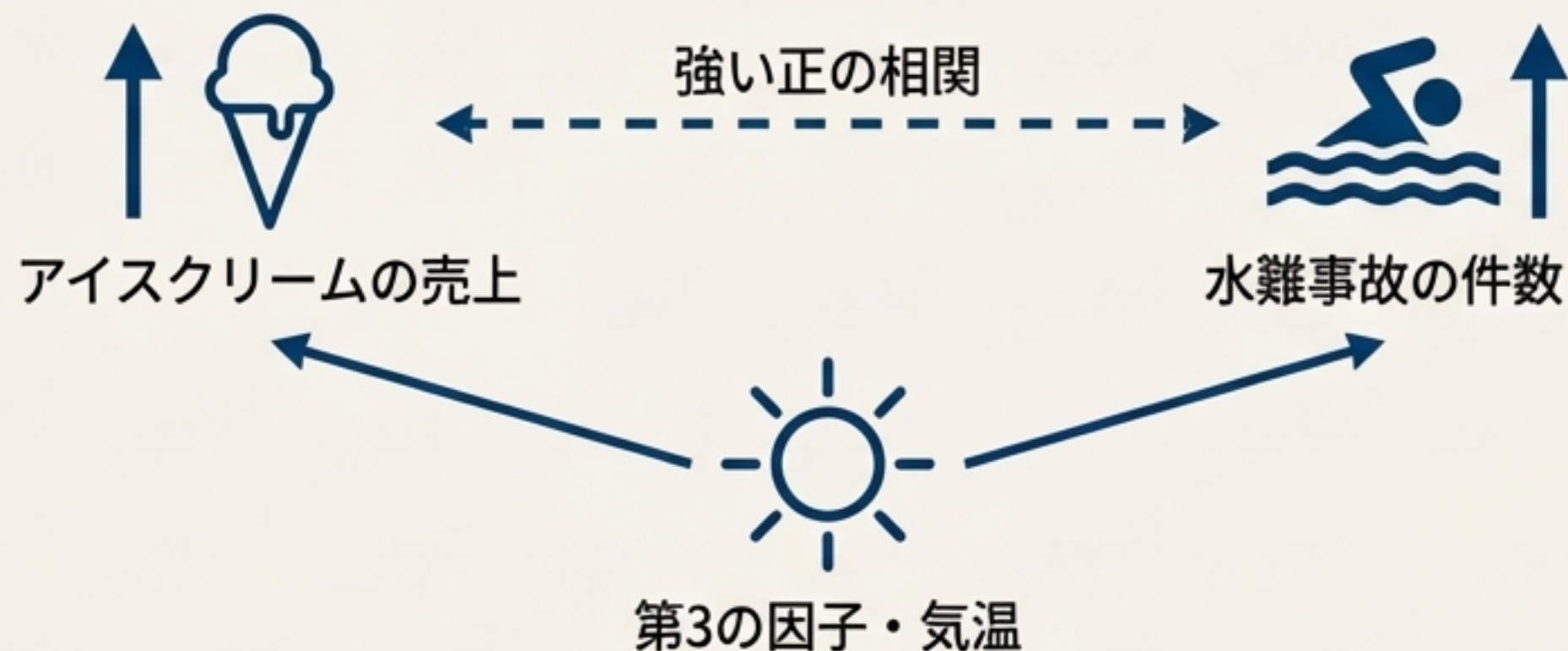
## Investigative Questions

- 「なぜこの日だけアクセスが急増したのか？」
- 「なぜこの地域だけ数値が突出しているのか？」



# 関係性の罠 ①：第3の容疑者を見抜く

The Pitfall: 相関関係と因果関係の混同



Classic Case: 「アイスクリームの売上」と「水難事故の件数」には強い正の相関がある。

The Truth: アイスが事故の原因ではない。隠れた第3の因子「気温」が存在する。（このような因子を交絡因子と呼ぶ）

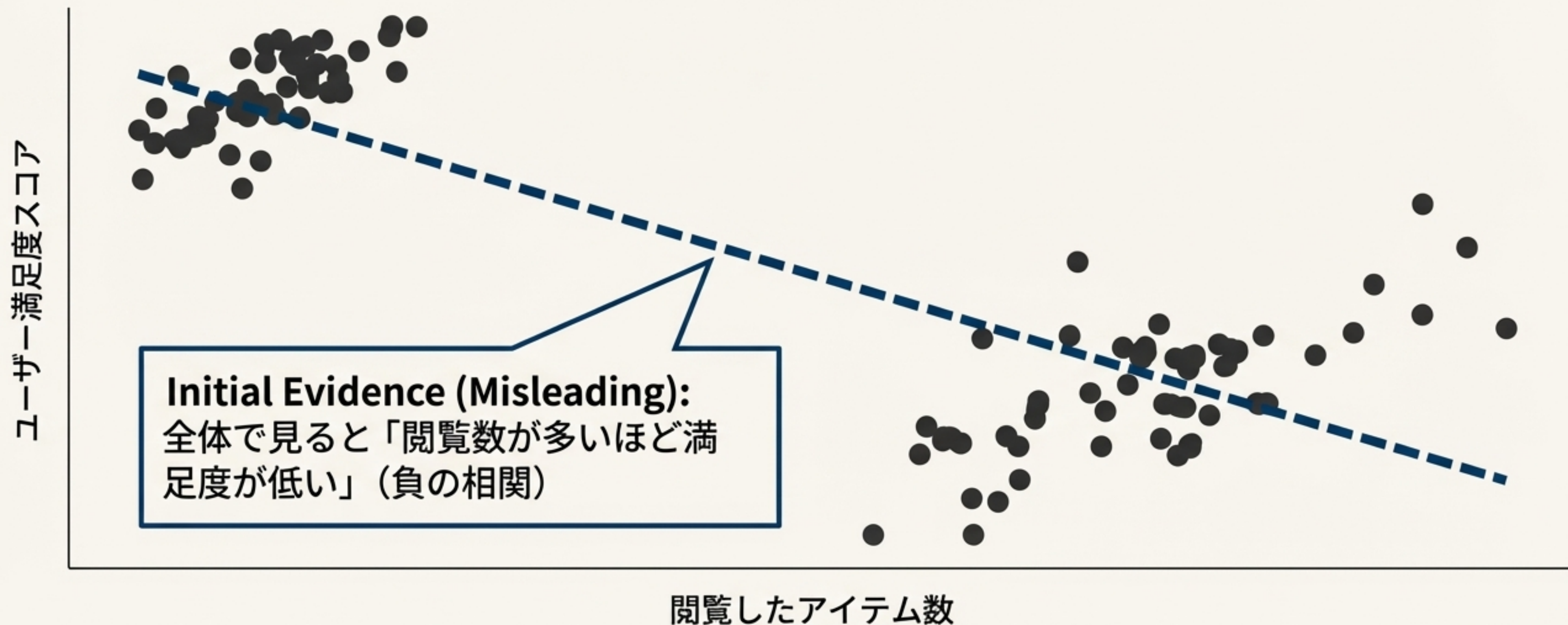
## Detective's Rule

AI は相関を見つけるのが得意だが、因果関係を正しく見抜くのは、ドメイン知識を持つ人間の仕事である。

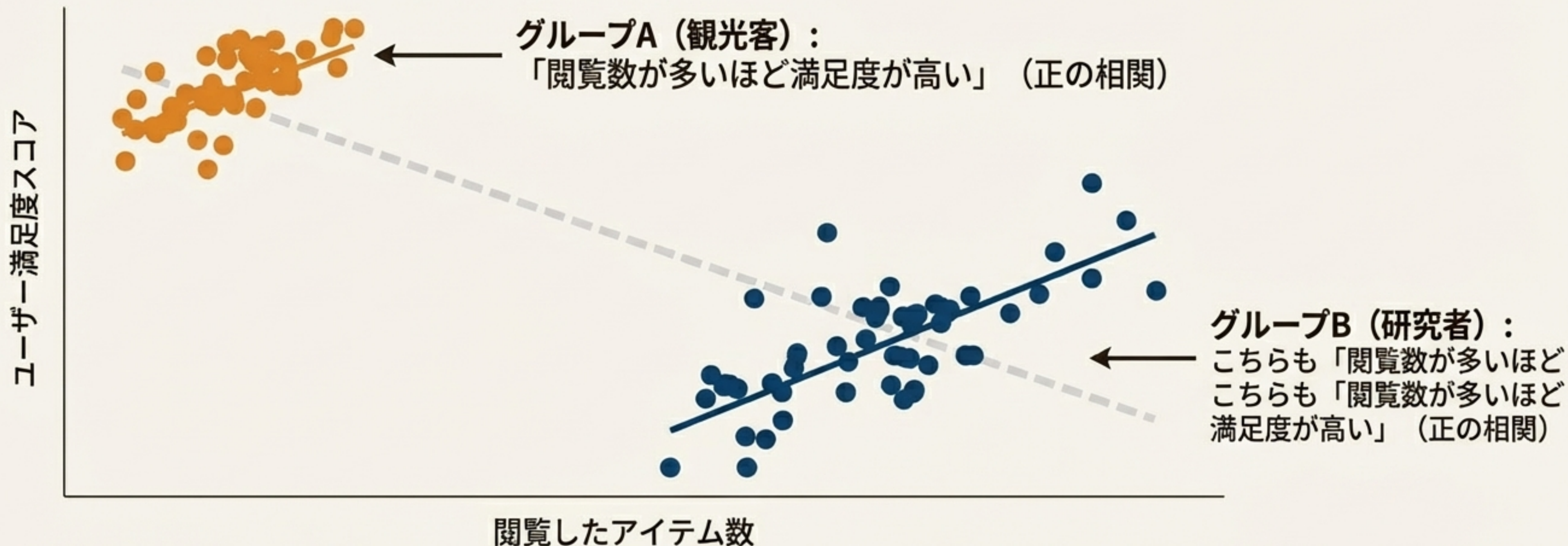
## 関係性の罠②：逆転する真実のパラドックス

The Phenomenon: シンプソンのパラドックス (Simpson's Paradox)

全体で見た傾向と、グループ別に分けて見た傾向が「逆転」する現象。



# 解決策：「混ぜるな、分けろ」



**The Golden Rule of EDA: データを「混ぜるな、分けろ」。EDAの鉄則です。**

# データ探偵の心得

1

**数値は嘘をつく。まず、描け。**

平均値などの要約統計量を過信せず、必ずグラフを描いて分布を確認する。

2

**分布の形と外れ値に語らせる。**

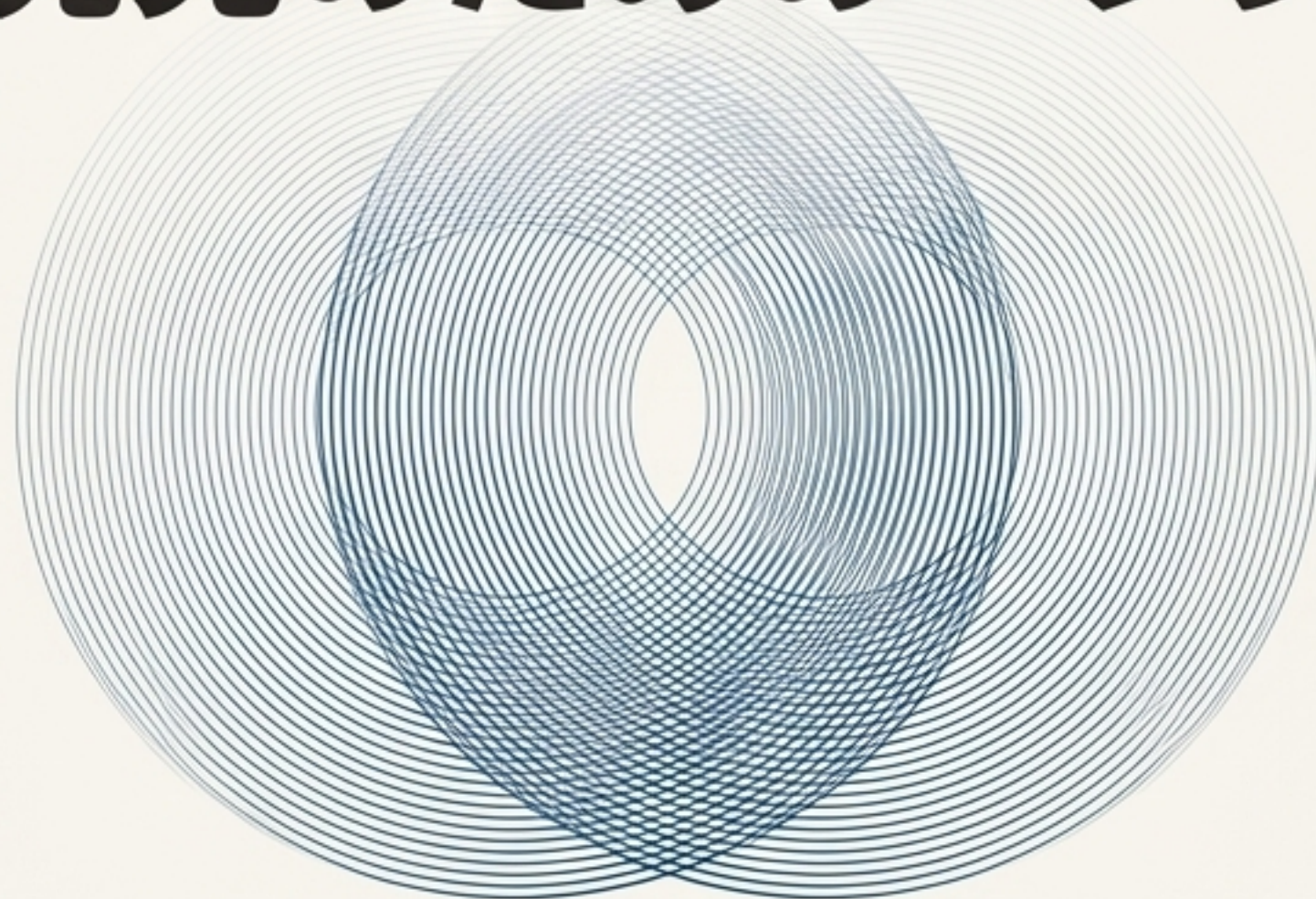
ヒストグラムで山の数を、箱ひげ図で外れ値を確認し、データの多様性を把握する。

3

**全体で判断するな。まず、分ける。**

属性ごとにデータを分割（層別化）し、隠れた構造をあぶり出す。

# 可視化は、発見のための「レンズ」である



データ可視化は、単なる発表資料作成のスキルではありません。  
それは、データという無機質な信号の中から、人間臭い事実や、  
地域固有の課題を発見するための知的な営みなのです。