AI（人工知能）講座
第４講
データの可視化と
探索的データ分析（EDA）

日本経済大学経営学部
教授　荒木貴之

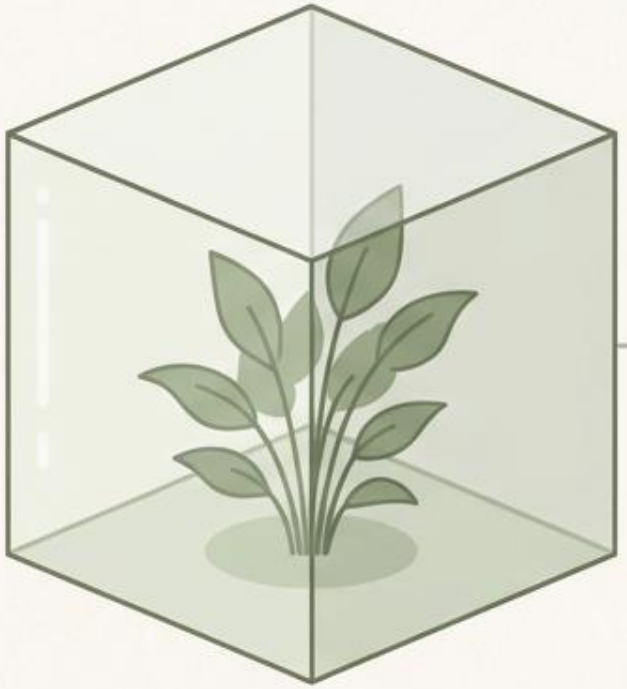
【学習到達目標】

1. 「探索的データ分析（EDA）」の概念を理解する：仮説検証型の分析に入る前に、データの構造や特徴を直感的に把握するプロセスの重要性を理解する。

2. 基本統計量の限界と可視化の効用を知る：平均値や分散などの数値要約だけでは見落としてしまうデータの「真の姿」を、グラフ化によって発見できることを学ぶ。

3. 多角的な視点でデータを解釈する能力を養う：ヒストグラム、箱ひげ図、散布図などの適切な使い分けを習得し、シンプソンのパラドックスなどの統計的な落とし穴を回避する視座を持つ。

皆さんは普段、業務で「平均値」をどれくらい信じていますか？



常時 15°C

平均
15°C

「平均気温15°Cの部屋は
快適でしょうか？」

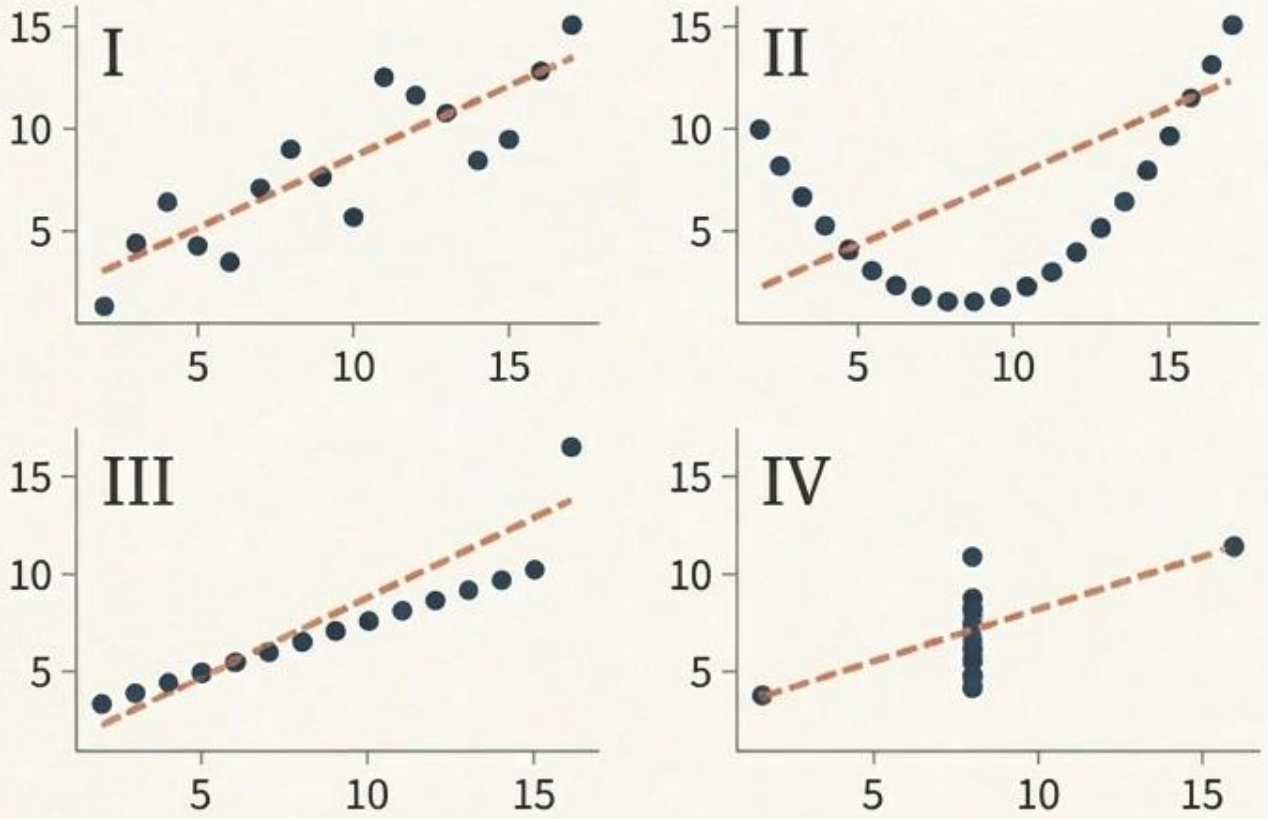
一見すると快適そうに思えます。しかし、もし昼が30°Cで、夜が0°Cだったら？
平均値はデータの「表情」を消し去り、実態を見えなくしてしまう危険性を孕んでいます。
本講座は、数字のトリックに騙されず、データの「素顔」を見るための技術を学びます。

これらのデータセットは、同じ性質を持つと言えるでしょうか？

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

AIや自動計算プログラムは「同一の性質を持つ」と判定するかもしれません。

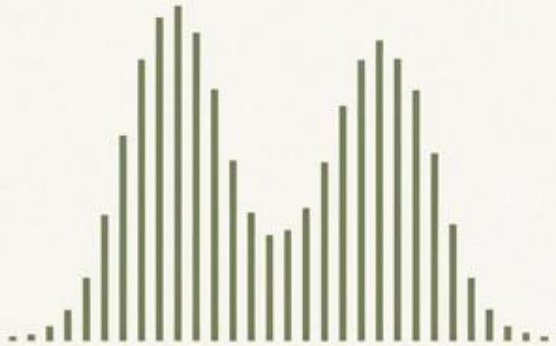
数値だけでは見えない。形にすれば、0.1秒で真実がわかる。



これが「探索的データ分析（EDA）」の力です。計算機に任せる前に、人間がその目で
データの「形」を確認する。それは、アーカイブに眠る新たな発見への第一歩です。

データとの対話をはじめよう：EDAの基本ツール

ヒストグラム（Histogram）

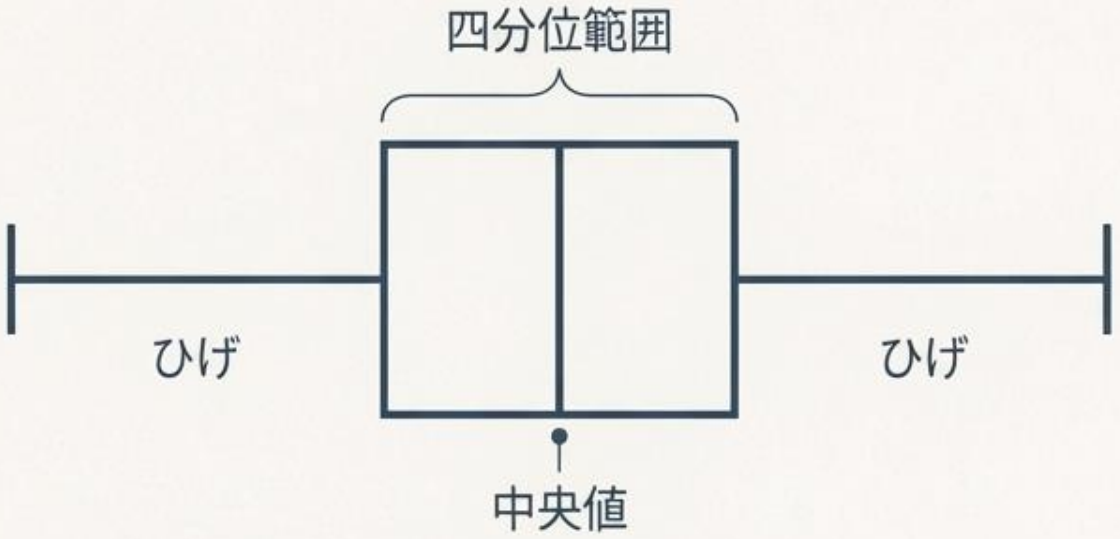


データの全体像を把握し、「分布の形」を見る。例えば、観光客の満足度アンケートで平均3点でも、実は「5点(大満足)」と「1点(大不満)」の二つの山（二峰性）に分かれているかもしれません。平均点にはない物語を発見します。

箱ひげ図（Box Plot）

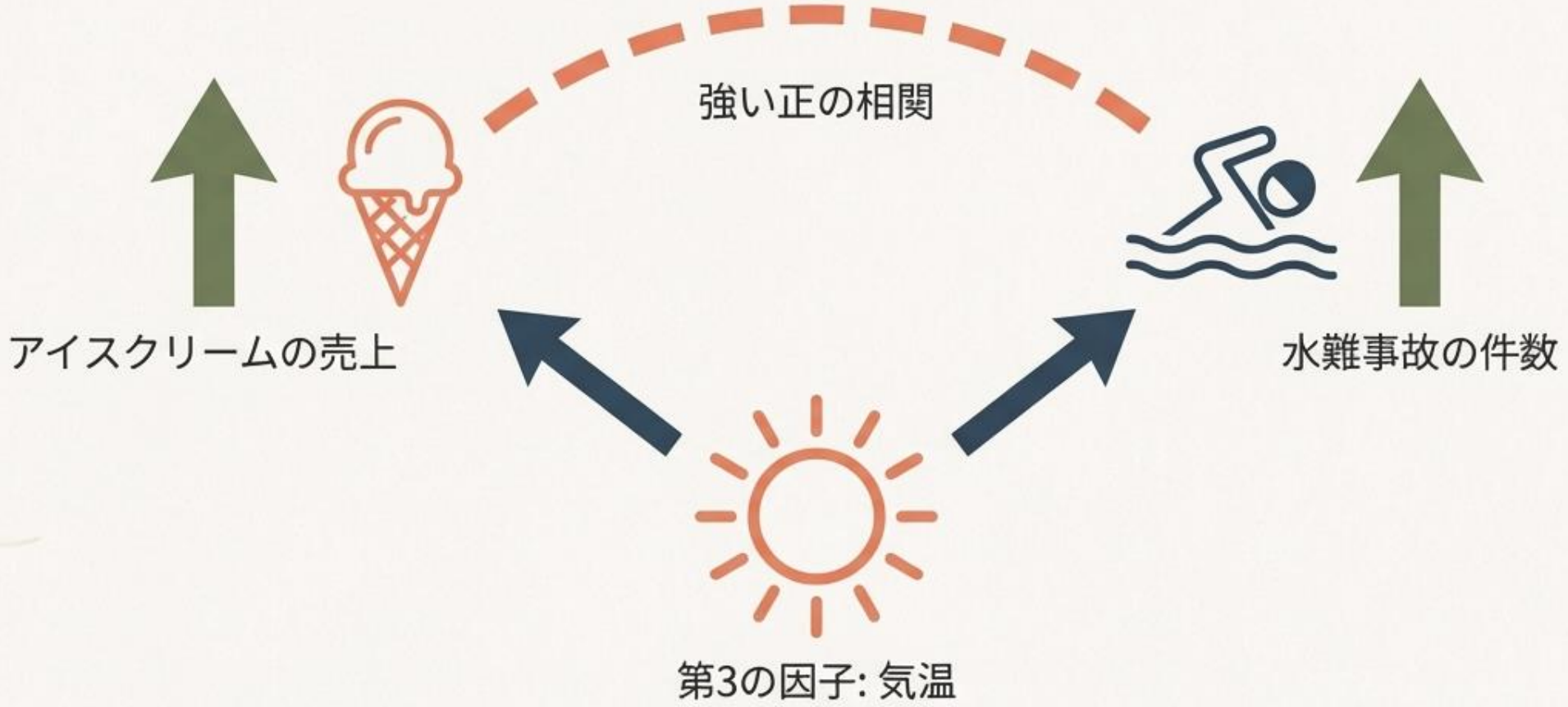
複数のグループの分布を比較し、「外れ値」を発見する。統計学ではノイズとして扱われがちな外れ値は、 文化研究においてはイノベーションの源泉です。

ノイズか、イノベーションの種か。
「愛すべき外れ値」を見つけ出す。

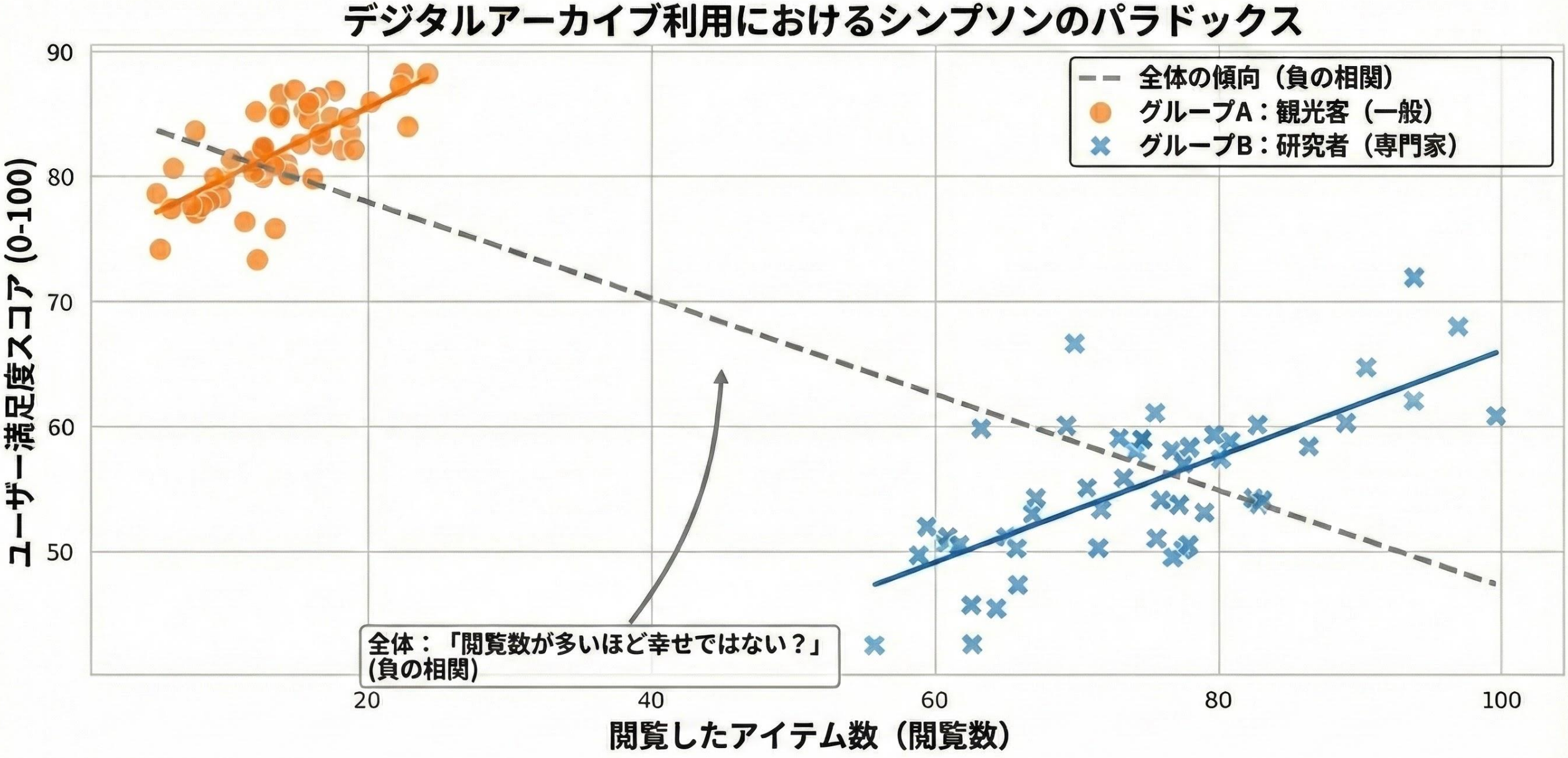


「なぜ、この日だけ来館者が爆発的に増えたのか？」
「なぜ、この浮世絵画像だけ、海外からのアクセスが異常に多いのか？」
そこには必ず、固有の文脈やストーリーがあります。
外れ値はエラーではありません。
新しい発見の種であり、イノベーションの源泉となります。

相関は因果ではない：データ解釈における最大の鉄則



「アイスクリームの売上」と「水難事故の件数」には強い相関があります。しかし、アイス販売を禁止しても事故は減りません。背後にある真の原因は「気温」です。AIは相関を見つけるのは得意ですが、その背景を解釈し、因果関係を正しく見抜くのは、　ドメイン知識を持つ人間にしかできません。


デジタルアーカイブ利用におけるシンプソンのパラドックス
ユーザー満足度スコア (0-100)
閲覧したアイテム数（閲覧数）
90
80
70
60
50
20
40
60
80
100
全体の傾向（負の相関）
グループA：観光客（一般）
グループB：研究者（専門家）
全体：「閲覧数が多いほど幸せではない？」
(負の相関)

【学習到達目標】

1. 「探索的データ分析（EDA）」の概念を理解する：仮説検証型の分析に入る前に、データの構造や特徴を直感的に把握するプロセスの重要性を理解する。

2. 基本統計量の限界と可視化の効用を知る：平均値や分散などの数値要約だけでは見落としてしまうデータの「真の姿」を、グラフ化によって発見できることを学ぶ。

3. 多角的な視点でデータを解釈する能力を養う：ヒストグラム、箱ひげ図、散布図などの適切な使い分けを習得し、シンプソンのパラドックスなどの統計的な落とし穴を回避する視座を持つ。

【課題】

1. 外れ値のケーススタディ
   ご自身の職場や身近なデータ（なければ公開されているオープンデータ）において、「外れ値」と思われるデータを探してください。そして、その外れ値が「単なるエラー（ノイズ）」なのか、それとも「重要な意味を持つ特異点（インサイト）」なのか、その背景を調査して記述してください。

2. 「平均値」の再考
   ニュースや業務報告で使われている「平均値」を一つ取り上げ、それが実態をミスリードしている可能性がないか考察してください。「もしヒストグラムを描いたら、どのような形になっていると推測されるか」を図示して説明してください。

3. シンプソンのパラドックスの構築
   「全体で見るとAの傾向があるが、層別化すると逆の傾向になる」という架空の、あるいは実際のシナリオを一つ作成してください。

   （例：病院の手術成功率、学校のテストの平均点など、身近な例で構いません）。

AI（人工知能）講座
第４講
データの可視化と
探索的データ分析（EDA）

日本経済大学経営学部
教授　荒木貴之