第 4 講　データの可視化と探索的データ分析（EDA）

荒木貴之（日本経済大学経営学部・教授）

【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。

本講義では、AI 時代に人間が担うべき「データの読み解き」に焦点を当てます。統計学者のジョン・テューキー（John Tukey）が提唱した「探索的データ分析（EDA）」を軸に、数値の羅列であるデータをグラフ化し、そこに潜むパターンや特異点を発見するための実践的技法を学びます。

１．なぜ私たちはデータを可視化するのか

１－１　データと対話する技術

「データ分析」と聞くと、高度な数式や AI による自動予測を思い浮かべるかもしれません。しかし、最も重要なステップは、計算の前に人間が自分の目でデータをじっくりと観察することです。これを体系化したのが、米国の統計学者ジョン・テューキー（John Tukey）が 1970 年代に提唱した「探索的データ分析（Exploratory Data Analysis: EDA）」です。

従来の統計学が「立てた仮説が正しいか」を確認する「確証的データ分析（CDA）」を重視していたのに対し、EDA は「そもそもデータの中に何があるのか」を探る、いわば刑事の捜査のようなプロセスです。デジタルアーカイブに眠る膨大なメタデータやログデータも、まずは EDA によってその性質を理解しなければ、有効な活用はできません。

１－２　アンスコムの例（Anscombe's quartet）が教える「平均値の嘘」

私たちが普段業務で使う「平均値」は、時に重大な事実を隠蔽します。これを劇的に示したのが「アンスコムの例（Anscombe's quartet）」です。

いま、4 つの異なるデータセットⅠからⅣがあり、それぞれの「平均値」「分散」「相関係数」は完全に一致しています。数値だけで判断すれば、これらは「同じ性質のデータ」です。しかし、これをグラフ（散布図）にすると、その姿は全く異なります。

表１　基本統計量＜アンスコムの例（Anscombe's quartet）＞

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

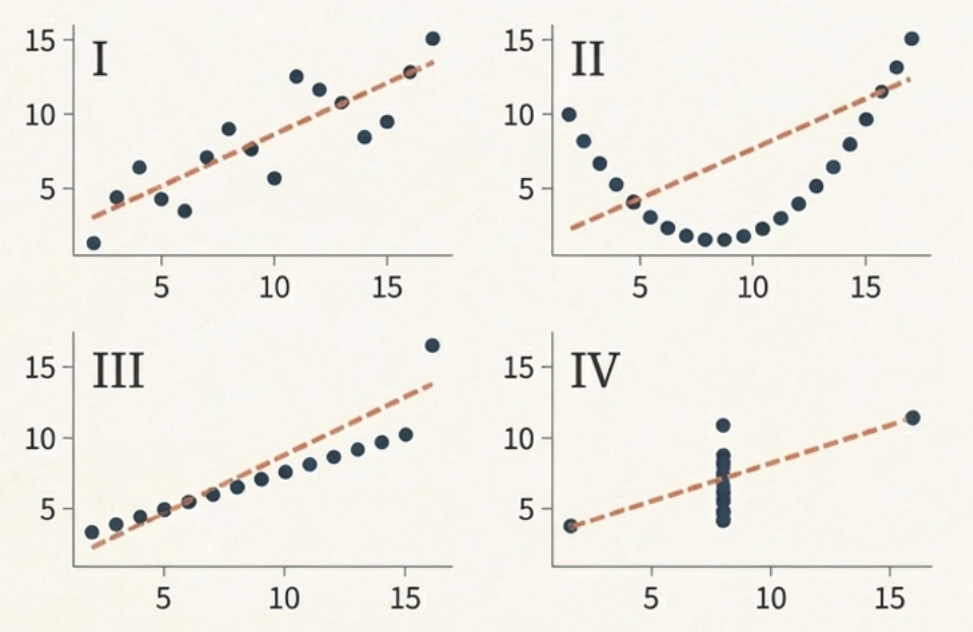

図４－１　散布図＜アンスコムの例（Anscombe's quartet）＞

Ⅰ　綺麗な右肩上がりの直線
Ⅱ　放物線を描く曲線
Ⅲ　直線上に乗っているが、一つだけ極端な外れ値があるもの
Ⅳ　X の値が全て同じで、一つだけ離れた点があるもの

このように、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。特に文化情報や地域データのような複雑な対象を扱う場合、可視化を省略した分析は極めて危険です。

２．分布の形状を読む（１変量データの可視化）

２－１　ヒストグラムと「ビンの魔術」

データの全体像をつかむ基本は「ヒストグラム（度数分布図）」です。横軸に階級（区間）、縦軸に度数（件数）をとります。ここで重要なのは、分布の「山」がいくつあるかです。

例えば、ある観光地の来訪者年齢層の平均が「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（bimodal）」になっているかもしれません。この場合、「平均 40 歳向けの施策」は、若者にもシニアにも響かない残念な手となります。

また、ヒストグラムは階級の幅（ビンサイズ）を変えるだけで印象が激変します。意図的に印象操作を行わないためにも、複数のビン設定を試す姿勢が必要です。

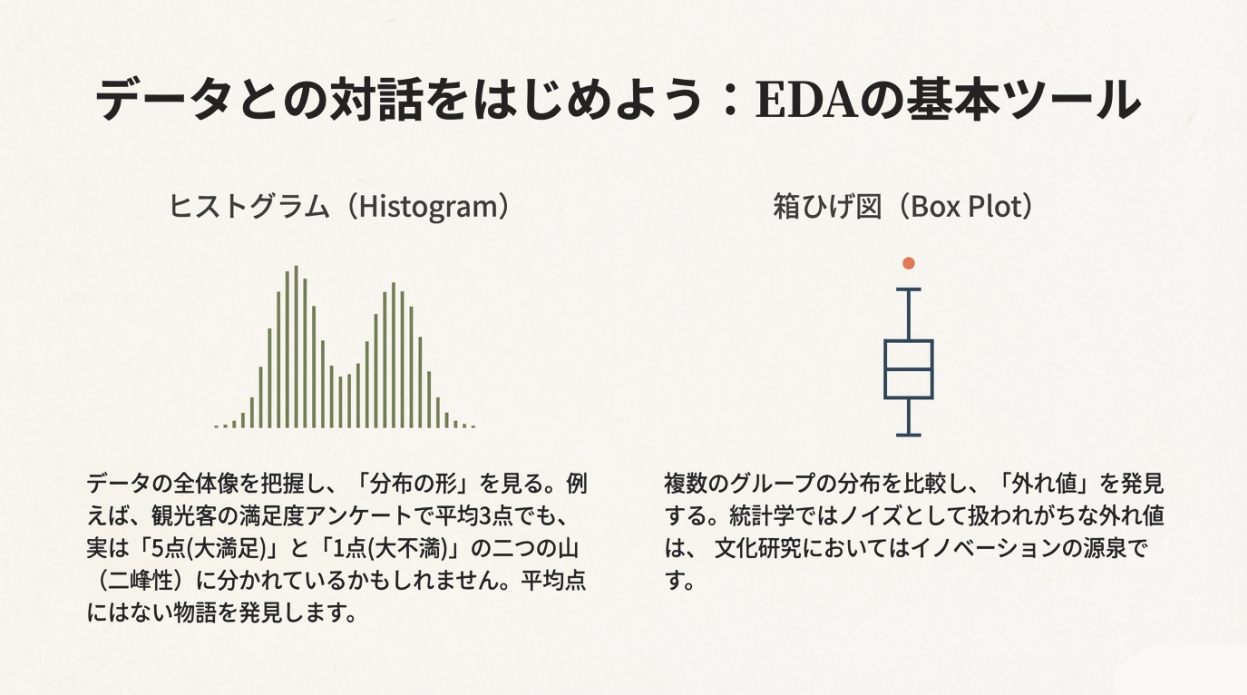

図４－２　EDA の基本ツール＜ヒストグラムと箱ひげ図＞

２―２　箱ひげ図で「外れ値」を愛する

複数のグループ（例：月ごとの来館者数）を比較する際に強力な武器となるのが

「箱ひげ図（Box Plot）」です。

- 箱（Box）：データの中央 50%（第 1 四分位数から第 3 四分位数まで）が含まれ、データの「実力」を示します。
- 中央値（Median）：箱の中の線。平均値より外れ値の影響を受けにくい指標です。
- ひげ（Whiskers）：通常のデータの範囲を示します。
- 外れ値（Outliers）：ひげの外にある点。

統計処理において外れ値は「ノイズ」として削除されることもありますが、EDA ではこの外れ値こそが「宝の山」です。「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？」。外れ値の背後には、必ず固有の文脈（コンテキスト）やストーリーが存在します。

３．関係性の罠を見抜く（２変量・多変量の可視化）

３－１　散布図と相関関係

２つの変数の関係を見るには「散布図」が最適です。右上がりなら「正の相関」、右下がりなら「負の相関」です。しかし、ここで陥りやすいのが「相関関係を因果関係と混同する」ことです。有名な例として「アイスクリームの売上と水難事故の件数には正の相関がある」という話があります。アイスクリームが事故の原因ではありません。ここには「気温」という第３の変数（交絡因子）が隠れています。可視化によって相関を見つけた後は、必ず人間のドメイン知識（現場の知見）でその背景を解釈しなければなりません。

図４－３　因果関係を見抜く（擬似相関）

３－２　シンプソンのパラドックス

データを分析する際、全体で見た傾向と、グループ別に分けた傾向が逆転する現象を「シンプソンのパラドックス」と呼びます。

例えば、デジタルアーカイブの利用調査で、「グループＡ：観光客（一般）」と「グループＢ：研究者（専門家）」の２つのグループにおいて、全体では「（デジタルアーカイブの）閲覧数が多いほど幸せではない」という結果が出たとします。しかし、これを「観光客」と「研究者」とに層

別化（ドリルダウン）して散布図を描くと、それぞれの層の中では「閲覧数が高いほど満足度が高い」という逆の傾向が見えることがあります。

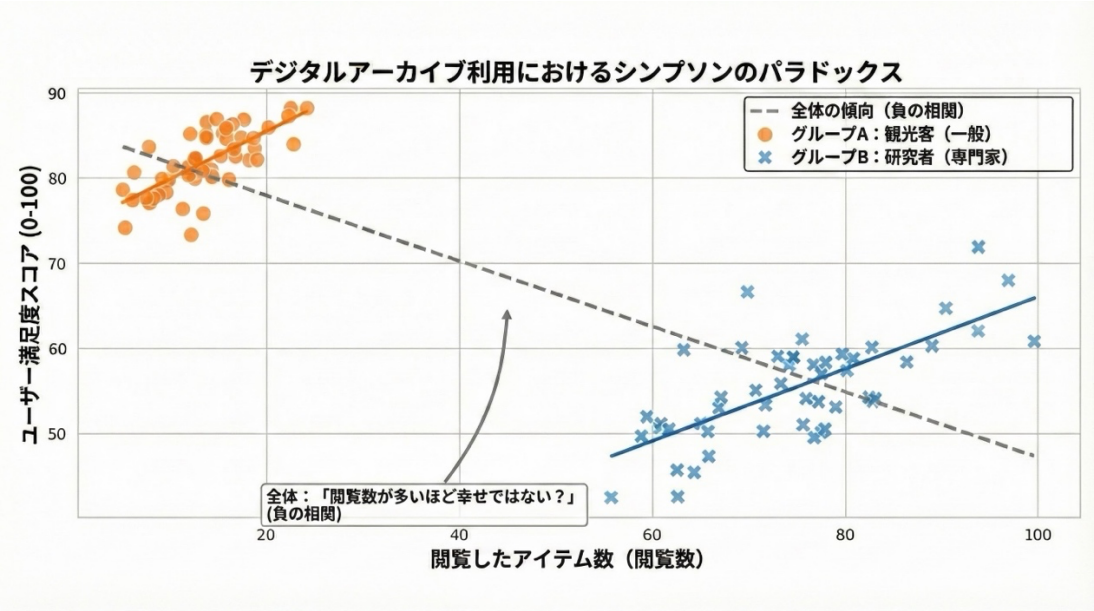

図４－４　シンプソンのパラドックス

これは、観光客（一般）の多くがユーザー満足度の出やすい利用者であり、研究者（専門家）は観光客に比べてユーザー満足度が出にくい傾向があるというような場合に起こる現象です。データを「混ぜるな、分けろ」は、EDA の鉄則です。

４．まとめ

本講義では、AI や高度な分析手法を使う前の基礎体力として、探索的データ分析（EDA）の重要性を学びました。

1. 数値は嘘をつく: 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こと。
2. 分布の形を見る: 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こと。
3. 層別化して見る: 全体の相関だけで判断せず、属性ごとにデータを分割（層別化）して、隠れた構造（シンプソンのパラドックスなど）をあぶり出すこと。

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「レンズ」なのです。本講の発展としての第 13 講「データの可視化の高度な技術」では、より複雑な「つながり」や「空間」を可視化する技術へと進みます。

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）。