03 記述統計（集計・可視化・分布の把握）

1. 第 2 回：記述統計（集計・可視化・分布の把握）

本日は、手元にあるデータの特徴を整理し、全体像を読み解く「記述統計」を学びます。

- **データの要約：**大量のバラバラなデータを「代表的な値」で表現する。
- **分布の可視化：**ヒストグラムを使って、データの「形」を観察する。
- **外れ値の発見：**特徴的なデータ（外れ値）を見つけ、その背景を考える。
- **実践：**ピボットテーブルで、商品の売上傾向を明らかにする。

「分布を読む」ことが、後の授業で学ぶ確率や推論の土台となります。

講義の狙い

数式の計算手順を覚えることよりも、集計結果やグラフを見て「現場で何が起きているか」を言語化できる力を養います。

2. 記述統計の役割：全体像をひと目で捉える

記述統計（Descriptive Statistics）とは、手元にあるデータの性質を要約し、分かりやすく記述する手法です。

- **必要性：**数百、数千行の生データをそのまま眺めても、傾向を掴むことは困難です。
- **解決策：**「代表的な数値（平均など）」や「グラフ」を用いて、情報を適切なサイズに圧縮します。

情報を適切に要約（圧縮）することで、初めて「前月との比較」や「カテゴリ間の違い」を客観的に議論できるようになります。

情報の圧縮

「情報を捨てることで、本質を見えやすくする」という統計の基本的な考え方を伝えます。

3. 代表値：データ群を象徴する一つの数値

大量のデータを説明する際、その「中心（真ん中）」あたりを示す数値を「代表値」と呼びます。

主な代表値

1. **平均値：**すべてのデータを均らした（重心となる）値。
2. **中央値：**データを大きさ順に並べたとき、真ん中にくる値。

本日は、この 2 つの代表値の使い分けをマスターします。これらを正しく選ばないと、データの実態を読み間違えるリスクがあるためです。

代表値

「中心」を定義する方法は複数あり、データの性質によって使い分ける必要があることを導入として示します。

4. 平均値：最も広く使われる代表値

平均値は、全データの値を合計し、それをデータの件数（サンプルサイズ）で割って算出します。

平均値の計算式

$$\text{平均値} = \frac{\text{全データの合計}}{\text{データの件数}}$$

（数式：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ）

理由：すべてのデータが同じ価値を持っていると仮定した際の「公平な水準」を示すため、全体の規模感や収支を測るのに適しています。

平均値

全データが「重心」を支えている物理的なイメージを伝えます。

5.【問い】年収の平均は「普通の人」の実感と同じか？

例えば、10 人のグループがあり、1 人だけ年収 3000 万円、残り 9 人が年収 300 万円だったとします。

- このグループの全員の年収を足して 10 で割ると、平均年収は **570 万円** です。

問い：この「570 万円」は、グループの一般的な姿（最多層）を正しく表しているでしょうか？

（答えは次のスライドで解説します）

問い

平均値が一部の極端な値に引きずられる性質を意識させます。

6.【答え】外れ値が平均値を大きく引き上げる

前の例で、平均値が実感とズレた原因は、3000 万円という極端に大きな値が混じっていたからです。

- **外れ値（Outlier）：**他のデータから大きく離れた極端な値のこと。
- **平均値の弱点：**1 件でも巨大な外れ値があると、平均値はそちらへ強く引っ張られてしまいます。

このように分布に偏りがあるデータでは、平均値だけを見て「普通はこのくらいだ」と判断するのは危険です。

外れ値の影響

平均値は全データを計算に入れるため、1 つの異常値の影響をまともに受けることを強調します。

7. 中央値：外れ値に左右されない指標

データの偏り（外れ値）の影響を受けにくい、もう一つの代表値が「中央値」です。

- **定義：**データを大きさ順に並べたとき、ちょうど中央（50%地点）に位置する値。

具体例：年収 300 万（9 人）と 3000 万（1 人）の例では、中央値は **300 万円** です。外れ値がどんなに巨大でも、順番が入れ替わらなければ中央値は変わりません。所得や売上個数など、偏りの大きいデータを扱う際の必須指標です。

中央値

Excel では MEDIAN 関数を使用することを補足します。

8. 分布：数値の背後にある「広がり」を知る

数値（平均や中央値）だけでは、データの全体像は完全には分かりません。そこで「分布」を確認します。

- **定義：**どのような値が、どの程度の頻度（件数）で存在しているかを示したもの。

データの集まり方や広がり方（分布の形）を見ることで、「どこにボリュームゾーンがあるか」「データがどのように散らばっているか」を視覚的に捉えることができます。

分布

「分布を読む ＝ データの個性（形）を知る」という感覚を養わせます。

9. ヒストグラム：分布を映し出すグラフ

分布を把握するために最もよく使われるのが「ヒストグラム」です。

- **階級（ビン）：**横軸。数値を一定の幅（例：10 点ごと）で区切った区間。
- **度数：**縦軸。その区間に含まれるデータの件数。

ヒストグラムの形を見ることで、データの山がどこにあるか、左右にどう広がっているかが一目で判明します。棒グラフと違い、横軸は「連続した数値」を表します。

ヒストグラム

実習で実際に作成するグラフなので、構成要素を丁寧に説明します。

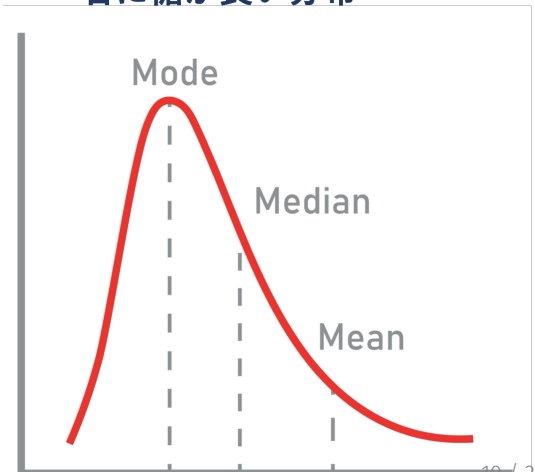
10. 分布の「歪み」：左右対称とは限らない

現実世界のデータ（特に所得や売上）は、右側（大きい方）に裾が長く伸びた形をしています。

右に裾が長い分布（正の歪み）

- **特徴：**大多数は低い値に集中しているが、一部に非常に高い値（外れ値）が存在する。
- **指標のズレ：**平均値が外れ値に引っ張られ、以下の関係になりやすい。
  - **中央値 < 平均値**

右に裾が長い分布



歪みの解説

「Positively Skewed」という用語と、なぜ平均値が中央値より大きくなるのかというメカニズムを解説します。

11. 実習 1：代表値を計算して観察する

Excel シート「chap2-1」の売上データ（682 件）を用いて、以下の数値を求めてください。

1. **関数を使用：**AVERAGE（平均）、MEDIAN（中央値）、MAX（最大値）、MIN（最小値）。
2. **分析ツールを使用：**「データ」タブ →「データ分析」→「基本統計量」を実行し、「要約統計量」を出力。

考察：平均値と中央値はどの程度離れていますか？ 最大値は平均から見て極端に離れていませんか？

実習1

数値を出すだけでなく、自分の目でその差を確認させます。

12. ピボットテーブル：大量データを一瞬で要約する

生データが何百行あっても、ピボットテーブルを使えばカテゴリごとの特徴を数秒で集計できます。

今回の実習での設定

- **行：**「タイプ（商品カテゴリ）」を指定。
- **値 1：**「商品 ID」を入れ、集計方法を「個数」にする。
- **値 2：**「売上個数」を入れ、集計方法を「平均」にする。

これにより、「どのカテゴリの商品が多く、どれがよく売れているか」という比較が可能になります。

ピボットテーブル

記述統計の実践において、ピボットは情報の「要約」を最も効率的に行うツールです。

13. 実習 2：商品タイプ別の売上傾向を掴む

シート「chap2-2」を用い、ピボットテーブルを作成してください。

操作手順

1. 行に「タイプ」を配置。
2. 値に「商品 ID（個数）」と「売上個数（平均）」を配置。
3. 売上の平均が高い順に並べ替えてください（降順ソート）。

集計後、平均値が最も高いカテゴリ（アルコール類など）の数値に注目してください。

実習 2

操作を支援しつつ、集計結果を「見る」準備をさせます。

14. 実習結果：カテゴリ別の要約表

以下のような集計表が完成したはずです。

商品タイプ（行ラベル）	商品数（個数）	売上個数の平均
アルコール類	23	33.13
スナック食品	135	30.83
青果物	136	28.07
（中略）		
総計	**682**	**27.73**

結果確認

総計（全体平均 27.7）と各カテゴリの差に着目させます。

15.【問い】集計結果の数値をどう解釈するか

集計の結果、アルコール類（平均 33.1）は全体平均（27.7）より明らかに高い数値です。

ここで問いを立てる：

- アルコールは「全商品が満遍なく」売れているのでしょうか？
- それとも、一部の「爆発的なヒット商品」が平均を押し上げているのでしょうか？

平均という「たった 1 つの数字」を見た後は、必ずその「背後の分布（グラフ）」を確認する必要があります。

考察

平均値だけを見て「アルコールなら何でも売れる」と結論づける危うさを指摘します。

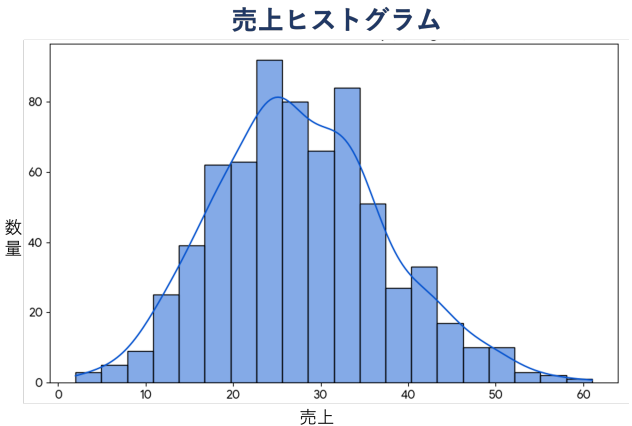
16. 実習 3：Excel でヒストグラムを作成する

売上個数のデータ全体を使って、グラフで「形」を確かめます。

操作手順：

1. 「売上個数」の列を選択。
2. 「挿入」タブ → ヒストグラムを選択。

観察ポイント：実際のデータも「右に裾が長い」形をしていますか？ 平均値 27.7 はどのあたりにありますか？



図：売上個数のヒストグラム（実データ）

実習の狙い

計算した「平均値」という点と、グラフで見た「分布」という面を一致させ、外れ値の存在を視覚的に理解させます。

17. 外れ値：異常か、それとも重要なヒントか

分布の端にある極端な値（外れ値）は、単なるノイズとは限りません。

- **理由 A：ミス**（入力間違い、測定不良）。これは削除が必要です。
- **理由 B：特殊な状況**（キャンペーン、季節特需、まとめ買い）。これはビジネスの成功要因を知るための「宝の山」です。

グラフで外れ値を見つけたら、まず「なぜこの値が出たのか？」という現場の状況と照らし合わせることが大切です。

外れ値

機械的に消すのではなく、データが生成された背景を想像させます。

18. 分布の「形」が教える現場の状況

分布の形を観察することで、データが生成された「背景」が見えてきます。

- **山が 2 つある：**性質の違う 2 つの集団が混ざっている可能性（例：ランチ客とディナー客）。
- **右に裾が長い：**多くの商品は平凡だが、一部に圧倒的な人気商品がある。

「分布」を意識することは、単なる計算作業を超えて、現場で起きている事象を正しく把握し、次の予測を立てるための土台となります。

理解の土台

この「形に意味がある」という感覚が、後の確率分布の理解へと繋がります。

19. 本日のまとめ：記述統計の流れ

データから客観的な判断を下すための、本日の実習の流れを整理します。

1. **集計：**ピボットテーブル等で情報を「要約」する。
2. **代表値：**平均値・中央値で「中心」を探る。
3. **観察：**ヒストグラムで「形（分布）」を見る。
4. **洞察：**外れ値や偏りの「原因」を考える。

数値（代表値）とグラフ（分布）を必ずセットで見る習慣をつけましょう。

まとめ

全体のポイントと、次回の「確率」への繋がりを明示します。

20. 実習課題の提出と次回の準備

作成した Excel ファイル（`chap2.xlsx`）に、以下の内容が含まれているか確認し、LMS から提出してください。

提出物のチェックリスト

- `AVERAGE`, `MEDIAN` 関数による計算結果。
- 基本統計量（分析ツール）の出力表。
- 商品タイプ別のピボットテーブル集計表。
- 全体の売上個数に関するヒストグラム。

次回（第 3 回）は、本日の「分布」の背後にある「起こりやすさ」を考えるための「確率の基礎」を学びます。

終了

課題の要件を整理し、授業を締めくくります。