

エキスパート活動 課題 A 「平均値（アベレージ）」

平均値は、**データの中心傾向を表す代表値の一つで、日常的に最もよく使われる統計量のひとつ**です。

□定義と計算式

平均値は、『データの値をすべて合計し、その合計をデータの個数で割った値』です。

計算式： 平均値 = (データの合計) / (データの個数)

□特徴

- データの**中心的な傾向を示す**：データ全体がどのあたりに集中しているのかを把握することができます。
- 計算が容易で理解しやすい：計算方法が単純で、誰でも簡単に理解し、計算することができます。
- **すべてのデータ**を使用：すべてのデータの値が計算に利用されるため、データ全体の情報を反映しています。

□平均値の種類

- ①算術平均：最も**一般的な平均値**で、データの値をすべて足してデータの個数で割ったものです。
- ②幾何平均：データの値をすべて掛け合わせて、その積の n 乗根を取ったものです。
主に、成長率や割合の平均を計算する際に用いられます。
- ③調和平均：データの逆数の算術平均の逆数です。主に、速度や時間の平均を計算する際に用いられます。

□使用する際の注意点

①外れ値の影響を受けやすい：

極端に大きい値や小さい値（外れ値）があると、平均値はその影響を大きく受けてしまいます。

例えば、10 人の年収の平均値を計算する際に、9 人が 300 万円前後なのに 1 人だけ 1 億円の人が含まれていると、平均値は大きく引き上げられてしまいます。

②データの分布が偏っている場合は、代表値として適切でない場合がある：

データの分布が偏っている場合、平均値はデータの中心的な傾向を正しく表さないことがあります。

例えば、右に裾が長い分布の場合、平均値は中央値よりも大きくなり、データの中心よりも右側に位置することになります。

□使用場面や分野

- ①学生のテストの平均点、②ある商品の 1 日の平均売上高、③ある国の国民の平均寿命
- ④教育分野：学生のテストの平均点や、クラス全体の平均学力を把握するために用いられます。
- ⑤ビジネス分野：売上高の平均値や、顧客満足度の平均値を分析することで、経営戦略の改善に役立てられます。
- ⑥医療分野：患者の平均寿命や、特定の病気の患者の平均治療期間を把握することで、医療水準の向上に役立てられます。

☆平均値は、データセットの**中心傾向を表す最も基本的な指標の一つ**です。しかし、外れ値やデータの分布によっては、平均値が適切でない場合もあることに注意が必要です。平均値を活用する際は、データの特性をよく理解し、必要に応じて他の指標（中央値や最頻値など）と組み合わせて分析することが重要です。

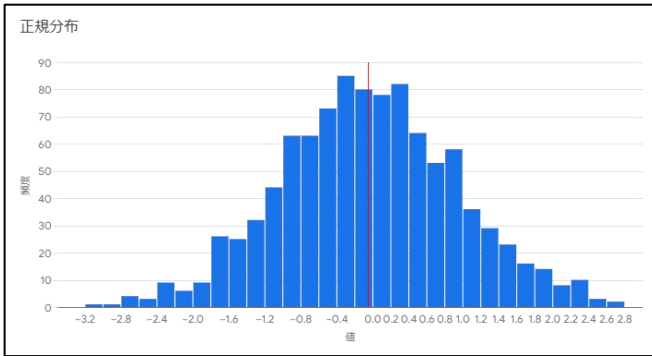
□平均値と分布の関係

平均値は、データの分布によってその意味合いが変わってきます。主な分布と平均値の関係を見ていきましょう。

1. 正規分布

正規分布は、左右対称の釣鐘型の分布です。

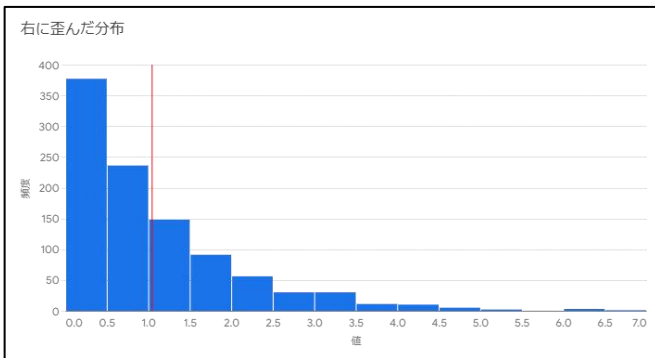
平均値は分布の中央に位置し、データの中心傾向をよく表しています。



2. 右に歪んだ分布

右に歪んだ分布は、右側に長い裾を持つ分布です。

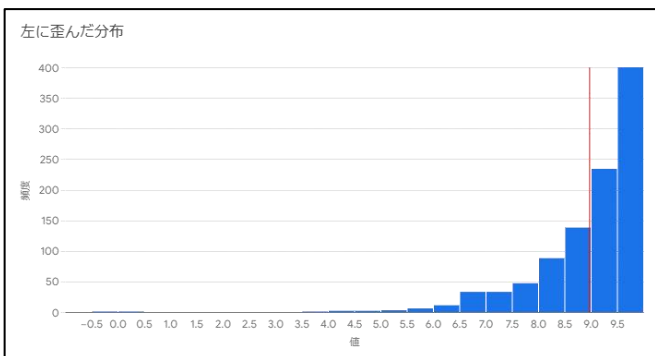
平均値は、中央値よりも右側に位置し、外れ値の影響を受けて大きくなる傾向があります。



3. 左に歪んだ分布

左に歪んだ分布は、左側に長い裾を持つ分布です。

平均値は、中央値よりも左側に位置し、外れ値の影響を受けて小さくなる傾向があります。



発問 国税庁の「令和5年分民間給与実態統計調査」によると、1年を通じて勤務した給与所得者1人あたりの平均年収は460万円です。この場合、「日本の年収ランキングでちょうど真ん中の順位（中央値に該当）」の人は、「460万円」より低い年収になるはずですが、この理由を考えなさい。

エキスパート活動 課題B 「中央値（メディアン）」

中央値とは、データセットをソートしたときに、**ちょうど真ん中に位置する値**のことです。データセットの代表値の一つとして、平均値と並んでよく用いられます。

□中央値の求め方

データセットをソートする：データを小さい順または大きい順に並べます。

中央値を求める：

データ数が**奇数**の場合：ソートされたデータの中央にある値が中央値です。

例 データセット：{3, 1, 4, 1, 5, 9, 2} // ソート：{1, 1, 2, 3, 4, 5, 9} // 中央値：3

データ数が**偶数**の場合：ソートされたデータの中央にある2つの値の平均値が中央値です。

例 データセット：{3, 1, 4, 1, 5, 9} // ソート：{1, 1, 3, 4, 5, 9} // 中央値： $(3 + 4) / 2 = 3.5$

□中央値の主な特徴

外れ値の影響を受けにくい：極端に大きな値や小さな値（外れ値）がデータセットに含まれていても、中央値はそれらの影響をほとんど受けません。

データの分布に依存しない：データがどのような分布をしていても、中央値は常にデータの中央を表します。

□中央値の使用例

年収：年収のように、外れ値の影響を受けやすいデータの場合、平均値よりも中央値の方が**実態に近い値**を示すことがあります。

住宅価格：住宅価格も、地域や物件によって大きな差があるため、平均値よりも中央値の方が一般的な価格帯を表すのに適しています。

テストの成績：テストの成績分布が歪んでいる場合、平均値よりも中央値の方が生徒の平均的な学力を示すのに役立ちます。

□中央値の注意点

データ全体の情報を反映しない：中央値は、データセットの中央の値しか考慮しないため、データ全体の分布やばらつき具合を把握することはできません。

データ数が少ない場合は不安定：データ数が少ない場合、中央値は偶然に左右されやすく、信頼性が低いことがあります。

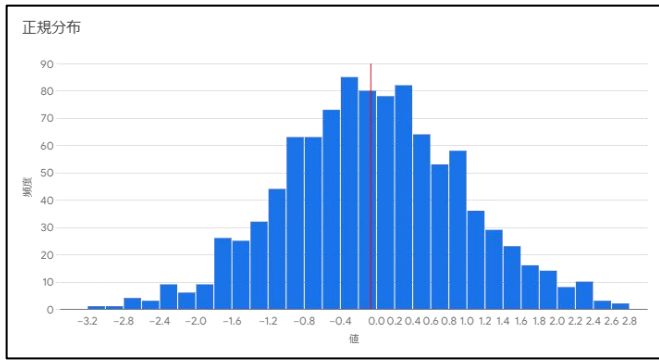
☆中央値は、外れ値の影響を受けにくく、データの分布に依存しないという特徴を持つため、様々な分野で活用されています。しかし、データ全体の情報を反映しないという点には注意が必要です。平均値と併用する、または他の統計量と組み合わせて分析することで、より深くデータの特徴を理解することができます。

□中央値と分布の関係

分布と中央値の関係について説明します。

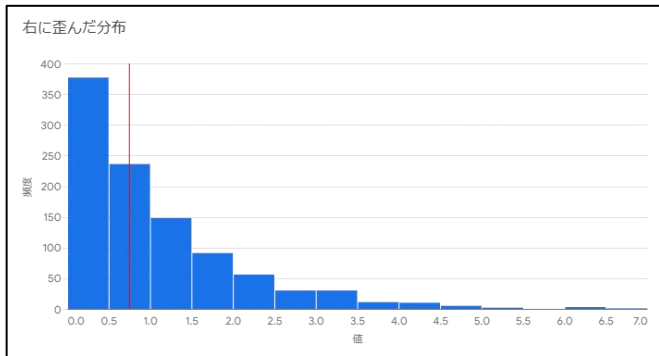
1. 正規分布

正規分布は、左右対称の釣鐘型の分布です。中央値は分布の中央に位置し、平均値と一致します。



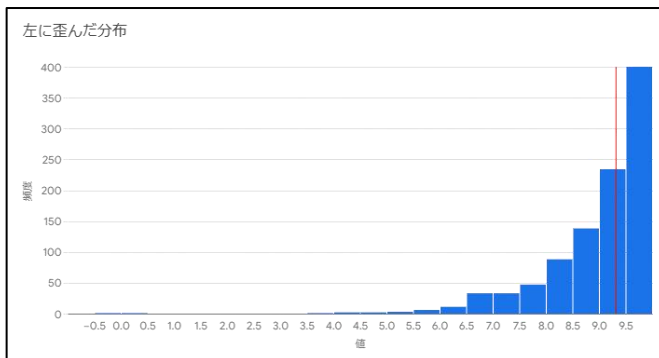
2. 右に歪んだ分布

右に歪んだ分布は、右側に長い裾を持つ分布です。中央値は平均値よりも左側に位置します。



3. 左に歪んだ分布

左に歪んだ分布は、左側に長い裾を持つ分布です。中央値は平均値よりも右側に位置します。



◎中央値は、分布の形に影響されずに、常にデータの中央に位置するという特徴があります。

これは、平均値が外れ値の影響を受けやすいのに対して、中央値は外れ値の影響を受けにくいという点で大きな違いです。

発問 データの特徴をみるために「平均値」を使うことはよくある。ただし、「平均値」よりも「中央値」の利用が望ましいときがある。それは、どのようなデータを扱うときか答えなさい。また、そのように考える理由を答えなさい。

エキスパート活動 課題C 「最頻値（モード）」

最頻値とは、データセットの中で**最も頻繁に出現する値**のことです。データセットの代表値の一つとして、平均値や中央値と並んでよく用いられます。

□最頻値の求め方

- ・データセットを整理する：データを小さい順または大きい順に並べたり、度数分布表を作成したりして、各値の出現回数を数えます。
- ・最頻値を求める：最も多く出現する値が最頻値です。複数の値が同じ頻度で最も多く出現する場合は、それらの値すべてが最頻値となります。

例

データセット：{3, 1, 4, 1, 5, 9, 2, 1} // 整理：{1, 1, 1, 2, 3, 4, 5, 9} // 最頻値：1

データセット：{3, 1, 4, 1, 5, 9, 2, 3} // 整理：{1, 1, 2, 3, 3, 4, 5, 9} // 最頻値：1と3

□最頻値の主な特徴

- ・**離散データに有効**：最頻値は、特にカテゴリカルデータや離散データ（例：性別、商品の種類、テストの点数）の代表値を求めるのに適しています。
- ・**外れ値の影響を受けにくい**：極端に大きな値や小さな値（外れ値）がデータセットに含まれていても、最頻値はそれらの影響をほとんど受けません。
- ・複数の値が存在する場合があります：データセットによっては、複数の最頻値が存在することがあります。

□最頻値の使用例

- ・売れ筋商品：ある商品の売上データにおいて、最も売れている商品の種類を把握するのに役立ちます。
- ・人気投票：人気投票の結果において、最も多くの票を集めた候補者を特定するのに使われます。
- ・顧客層の分析：顧客の年齢層や居住地などのデータから、最も多い層を把握するのに役立ちます。

□最頻値の注意点

- ・連続データには不向き：連続データ（例：身長、体重）の場合、値が重複する可能性が低いため、最頻値はあまり意味を持たないことがあります。
- ・データ全体の情報を反映しない：最頻値は、データセットの中で最も頻繁に出現する値しか考慮しないため、データ全体の分布やばらつき具合を把握することはできません。

☆最頻値は、**離散データの代表値**を求めるのに適しており、**外れ値の影響を受けにくい**という特徴を持ちます。しかし、**連続データには不向き**であり、**データ全体の情報を反映しない**という点には注意が必要です。平均値や中央値と併用したり、他の統計量と組み合わせて分析することで、より深くデータの特徴を理解することができます。

発問 平均値と中央値が同じような値を示している場合に最頻値の利用が有効になることがあります。それはどうしてでしょうか。答えなさい。