

ビジネスパーソン向け 統計データ活用セミナー

跡見学園女子大学
山澤成康

スライド 1

プログラム概要

- 1. なぜ今「データ分析×政府統計」か？
- 2. データ分析は何から始めればいいのか？
- 3. Excelを使った政府統計データの分析（演習）

1

スライド 2

1. なぜ今「データ分析×政府統計」か？

- ① データ分析に政府統計を使うメリット
ー 企業内データだけでは見えない社会的傾向を補完
- ② ビジネスシーンでの活用可能性
ー 市場分析・需要予測・地域比較などへの応用
- ③ 政府統計データ活用の先行事例紹介
ー 民間企業での具体的導入例を紹介

2

スライド
3

①データ分析に政府統計を使うメリット
ー企業内データだけでは見えない社会的傾向を補完

政府統計のメリット

信頼性

公的に整備された信頼データ
経営計画・報告書

多様なデータ

市場・需要の把握
新商品開発・販路拡大

マクロ経済の把握

景気・産業の動向分析
リスクマネジメント

詳細さ

立地・市場分析
店舗展開計画

コスト削減

無料の高品質データ
分析の基本

説得力

公的権威で説得力UP
提案書・IR資料

政府統計のメリットは、主に 6 つあります。「信頼性」、「多様なデータ」、「マクロ経済の把握」、「詳細さ」、「コスト削減」、「説得力」です。

スライド
4

①データ分析に政府統計を使うメリット
ー企業内データだけでは見えない社会的傾向を補完

主な統計の紹介

	統計名称	調査項目
人口統計	国勢調査	年齢、性別、世帯構成、職業など
	人口推計	世帯数、家族構成、住宅情報
生活関連統計	家計調査	年齢別人口、男女別人口、地域別人口
	労働力調査	消費支出、収入、貯蓄・負債
産業・企業統計	経済センサス	就業形態、産業形態、失業率
	従工業務数	企業数、事業別売上高、従業員数
	商業動向統計調査	販路数や顧客の生産数量、出荷額、在庫水準
		小売業、卸売業の売上高、在庫、販売数量

ビジネスでも活用できる主な統計について、「人口統計」、「生活関連統計」、「産業・企業統計」に分け、統計名称と調査項目に整理した表を示します。

スライド
5

①データ分析に政府統計を使うメリット
ー企業内データだけでは見えない社会的傾向を補完

統計の紹介

家計調査

・毎月何にどの程度支出したか
・収入階級別、年齢別

国勢調査

・地域ごとの人口、世帯

経済センサス
活動調査

・地域ごとの事業所数、従業員数、売上高

ここでは、代表的な統計を3つ紹介します。

「家計調査」は、毎月の支出内容が分かる統計で、収入階級別、年齢別の支出額を把握できます。

「国勢調査」は、地域ごとの人口や世帯の実態を把握できます。

「経済センサス活動調査」は、地域ごとの事業所数、従業員数、売上高などが分かります。

スライド
6

①データ分析に政府統計を使うメリット ー企業内データだけでは見えない社会的傾向を補完		
ビジネスシーンでの活用例（１）		
分野	活用内容	主な利用統計例
① 市場分析	商品の特性を分析する	家計調査
② 地域分析・立地戦略	地域ごとの人口構成、所得水準、産業構造を分析して出店・投資先を決定	国勢調査、経済センサス活動調査
③ 顧客セグメンテーション	年齢・性別・職業・家族構成などの人口統計からターゲット顧客層を特定	国勢調査、就業構造基本調査
④ 景気・業界動向分析	マクロ経済・産業別動向をモニタリングし、経営判断や在庫調整に活用	景気動向指数、鉱工業生産指数、法人企業統計調査
⑤ 労働市場・人材戦略	労働需給や賃金動向を分析し、採用・給与戦略に反映	労働力調査、賃金構造基本統計調査

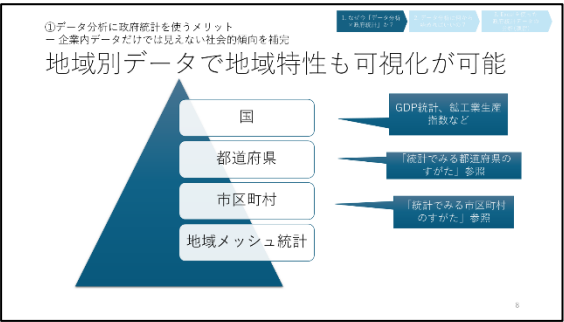
次は、企業側の視点から「どのような統計を使えば良いか」という観点で活用例を説明します。
まずは「①市場分析」、「②地域分析・立地戦略」、「③顧客セグメンテーション」、「④景気・業界動向分析」、「⑤労働市場・人材戦略」の切り口から、想定される活用内容と利用可能な主な統計は表のとおりです。

スライド
7

①データ分析に政府統計を使うメリット ー企業内データだけでは見えない社会的傾向を補完		
ビジネスシーンでの活用例（２）		
分野	活用内容	主な利用統計例
⑥サステナビリティ・ESG対応	CO ₂ 排出、エネルギー消費、廃棄物量などの統計で環境経営を支援	総合エネルギー統計
⑦観光・インバウンド分析	訪日外国人の動向や旅行支出の分析で観光ビジネスを最適化	インバウンド消費動向調査（旧 訪日外国人消費動向調査）
⑧不動産・都市開発分析	地価・住宅需要・建築動向を把握し、開発計画や投資判断に活用	住宅・土地統計調査、建築着工統計調査
⑨リスクマネジメント・事業継続計画（BCP）	災害リスク、人口減少、感染症動向などを定量的に評価	気象統計、人口動態調査
⑩イノベーション・新規事業探索	統計を組み合わせて、未来ビジネス・成長分野を特定	科学技術研究調査

次は、「⑥サステナビリティ・ESG対応」、「⑦観光・インバウンド分析」、「⑧不動産・都市開発分析」、「⑨リスクマネジメント・事業継続計画（BCP）」、「⑩イノベーション・新規事業探索」という観点についても、想定される活用内容と利用可能な主な統計は表のとおりです。

スライド
8



国全体やマクロ経済を捉える統計としては、「GDP 統計」や「鉱工業生産指数」などがあります。都道府県別の統計については「統計でみる都道府県のすがた」を見るとよくわかります。市区町村別の統計は、「統計でみる市区町村のすがた」を見るとよいでしょう。政府統計には、さらに細かい「地域メッシュ統計」もあり、250m メッシュにどの程度人口が存在するか、などが分かります。

スライド
9

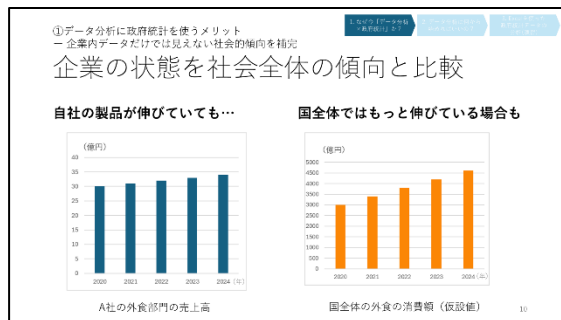
①データ分析に政府統計を使うメリット
—企業内データだけでは見えない社会的傾向を補完
参考：「統計でみる都道府県のすがた」
「統計でみる市区町村のすがた」

「統計でみる都道府県のすがた」「統計でみる市区町村のすがた」は、社会・人口統計体系の指標データの中から自治体の主要なものを按群し報告書に取りまとめたもの。数値を見やすくするための目安として、自治体別順位を付している。

URL : <https://www.stat.go.jp/data/k-sugata/index.html>

「統計でみる都道府県のすがた」
「統計でみる市区町村のすがた」
は総務省統計局のホームページに掲載されています。PDF ファイルの形式で冊子が手に入ります。さらに、e-Stat（政府統計の総合窓口）というポータルサイトを活用すると、実際にどの程度の数の統計が公表されているのかが分かります。

スライド
10



自社のデータだけを見て、例えば「外食部門の売上が上がっているので順調だ」と判断していても、国全体のデータを確認すると、外食産業全体はそれ以上に大きく成長しているケースもあり得ます。このように、自社の実績を相対的に把握するためにも、社会全体の動向を注視することが重要になります。

スライド
11

②ビジネスシーンでの活用可能性
—市場分析・需要予測・地域比較などへの応用
マーケティングツールとしての家計調査

- ・「ビッグデータ」
- ・さまざまな商品の特性を捉えられる
- ・日本全体の傾向がわかる
- ・総額、収入階級別、年齢階級別で分析
- ・金額、数量、価格面からも分析可能（今回は金額中心）
- ・高い信頼性

次はマーケティングツールとしての「家計調査」について少し詳しく説明します。いわゆる「ビッグデータ」と呼んでもいいような、膨大なデータです。これを活用することで、様々な商品やサービスの特性を捉えることができます。総額、収入階級別、年齢階級別の分析のほか、日本全体の傾向に加え、地域別の傾向も分析することができます。

スライド
13

[illegible]

2 番目は「レンタカー・カーシェアリング料金」、3 番目が（意外にも）「寄付金」です。これにはふるさと納税の普及などが影響しているものと考えられます。また、全体を見ていくと、果物加工品やキウイフルーツ、乳飲料など、健康志向を反映した品目の支出が増えている傾向もうかがえます。

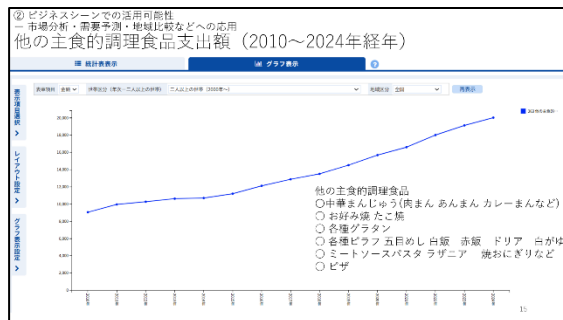
2019 年まで増加傾向ですが、新型コロナウイルス感染症の拡大により一度減少し、2023 年から 2024 年にかけて再び増加に転じていることが見て取れます。

スライド
14



次は「チョコレート菓子」の支出額です。こちらも順調に増加しています。これは「チョコレート」そのものではなく、チョコレートを使ったお菓子を指します。

スライド
15

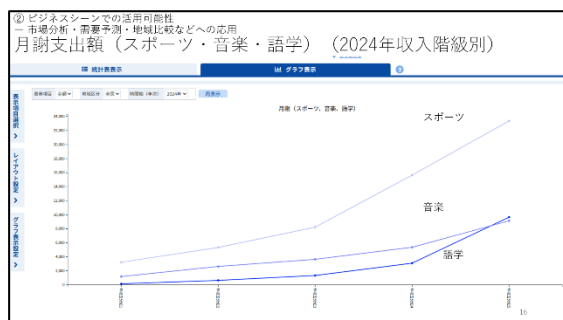


次は「他の主食的調理食品」の支出額です。こちらも右肩上がりで上昇しています。

「主食的調理食品」の代表的なものはお弁当や調理パンで、「他の主食的調理食品」は、それ以外のものを指します。

具体的には、中華まん、肉まん、あんまん、お好み焼き、たこ焼き、グラタンといった食品が該当し、支出が増えていることが分かります。

スライド
16

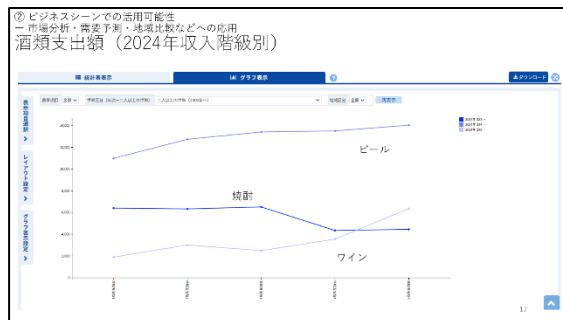


次はスポーツ、音楽、語学の「月謝」に対する支出額を収入階級別に見たグラフです。「年収五分位」とは、全世帯を収入の低い順から 5 つのグループに分けたもので、グラフの右側に向かって年収が高くなっています。

グラフを見ると、収入が増えるにつれて「スポーツ」への支出額が顕著に増えていることが分かります。一方、「音楽」と「語学」を比較すると、第 4 分位までは音楽への支

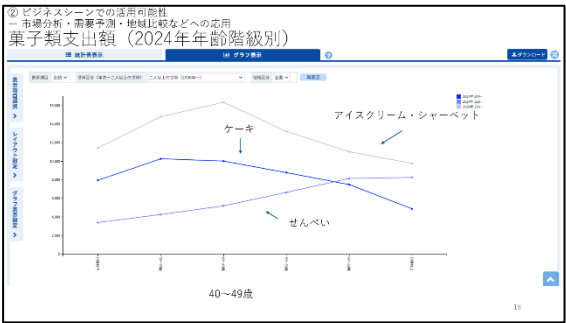
出が上回っていますが、最も年収が高い第 5 分位になると、語学への支出が音楽を逆転して上回るという特徴が見て取れます。

スライド
17



次に、「酒類」についてです。経済学では、所得が上がるにつれて消費が増えるものを「上級財」、逆に所得が増えると消費が減るものを「下級財」と呼びます。グラフを見ると、「ビール」は概ね収入に比例して支出が増える傾向にあります。「ワイン」は、最も収入が高い層（第5分位）でかなり支出が増えていることが分かります。逆に「焼酎」は収入が増えるにつれて支出が減っており、「下級財」に近い動きを示しています。

スライド
18



今回は、「菓子類」への支出額を世帯主の年齢階級別に見たグラフです。年齢によって消費支出がどのように変化するかを確認します。

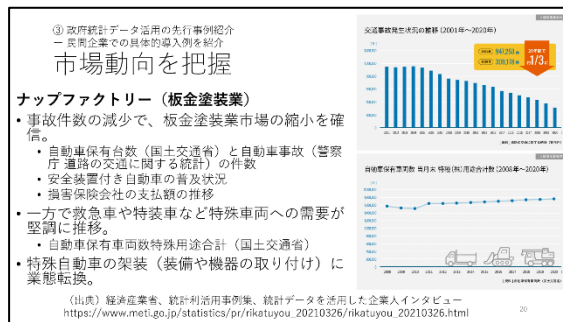
「アイスクリーム・シャーベット」は40～49 歳をピークに、年齢が高くなるほど減少していく傾向にあります。また、ケーキは 30～39 歳が最も高く、その後は年齢とともに減少していきます。一方、「せんべい」に関しては、年齢が高まるとともに消費支出額が増えていることが分かります。

スライド
19

政府統計データ活用の先行事例紹介

次に政府統計データ活用の先行事例を紹介します。

スライド
20



板金塗装業を営んでいたナップファクトリーは、市場縮小の予兆を検証するため政府統計を活用しました。事故件数の推移や安全装置の普及状況、損害保険の支払額といった多角的なデータから、従来の板金需要の減少を客観的に裏付けました。一方で、統計から特殊用途車両の保有台数が堅調であることに着目。業界構造を分析し、特殊車両の架装へと業態転換の成功に役立てた事例です。

スライド
21



次は、コンビニの立地戦略の事例です。これは「地域メッシュ統計」という、地域を網の目（メッシュ）状の区画に区切り、それぞれの区域にどの程度人口が存在しているかを把握できる統計を利用しています。国勢調査の地域メッシュ統計を活用すれば、総人口（夜中人口）だけでなく、昼間人口や世帯数も把握することができます。これを活用すれば、採算が取れる人口や世帯数などの条件に合致する場所を可視化（視覚的に特定）することができます。

スライド
22

③ 政府統計データ活用の先行事例紹介
ー 民間企業での具体的な導入例を紹介

不動産取引価格を推定

株式会社おたに

- GEEQでは、総務省統計局や国土交通省などが提供している路線価、国勢調査、住宅・土地統計調査などの官公庁系オープンデータをベースに、独自のアルゴリズムで物件の販売価格を予想。
- 価格予想にオープンデータを取り入れることで、解析対象となるデータ量は格段に増加し、アルゴリズムの工夫次第では、今後データから解釈できる内容を広げることが可能。

（出典）デジタル庁 民間事業者による利活用事例、オープンデータ100民間事業者による事例
https://www.digital.go.jp/resources/data_case_study_private

次は、不動産取引価格の予測に政府統計を活用した事例です。総務省統計局や国土交通省などが提供している路線価、国勢調査、住宅・土地統計調査などを用いて、物件の販売価格を予測するサービスです。

スライド
23

③ 政府統計データ活用の先行事例紹介
ー 民間企業での具体的な導入例を紹介

揺れやすさ指標の作成

大和ハウス工業株式会社

- 建設業者や住宅購入者向けに提供される地震リスク評価サービス。
- 所在地を入力すると、J-SHISの地震動予測値や地盤データを基に揺れやすさを算出する。

（出典）デジタル庁 民間事業者による利活用事例、オープンデータ100民間事業者による事例
https://www.digital.go.jp/resources/data_case_study_private

→国勢調査データを組み合わせると、人口密度を確認したうえで災害時の危険度を評価することも可能。

次は、大和ハウス工業の「ココゆれ」というサービスの事例です。これは、建設業者や住宅購入者向けに地震リスクを評価するものです。防災科学技術研究所が公開している J-SHIS（地震ハザードステーション）の地震予測値データを基にしています。さらに、このようなサービスに国勢調査のデータを組み合わせれば、人口密度を確認したうえで、「揺れやすさ」だけでなく、災害時や火災時の危険度などのリスク評価が可能になります。

スライド
24

2. データ分析は何から始めればいいのか？

- ①データの「見え方」を知る
 - 平均値・中央値・最頻値・標準偏差などの基礎
 - 相関分析：2つの指標の関係を捉える
 - 回帰分析：2つの変数の関係を捉える
- ②データ分析の基本姿勢
 - 思い込みで左右されない判断
 - グラフの目的に応じた使い方を学ぶ

24

続いて、政府統計から少し離れて、統計を使った分析手法について紹介します。

最初に、「平均値」、「相関分析」、「回帰分析」といった内容を扱います。これらは Excel を使って自分で分析することもできます。

加えて、データ分析の基本的な姿勢として、何に気を付ける必要があるのか、について整理します。

スライド
25

- ①データの「見え方」を知る
 - 平均値・中央値・最頻値・標準偏差などの基礎
- ### 見え方の基礎
- 平均値、中央値、最頻値
 - 標準偏差
 - 相関

25

まず、「平均値」、「中央値」、「最頻値」、「標準偏差」、「相関」といった内容を扱います。

スライド
26

- ①データの「見え方」を知る
 - 平均値・中央値・最頻値・標準偏差などの基礎
- ### 多くのデータを代表値で表す
- 平均値：すべてのデータを足して、データの数で割る
 - 中央値（メジアン）：データを値の大きさの順に並べた場合の真ん中の値
 - 最頻値（モード）：データセットの中で最も頻繁に出現したデータ

それぞれ、エクセルで計算できる。

26

人間は大量の数値を一度に理解することが難しいため、データの全体像をつかむには以下の 2 つのアプローチが有効です。

1. 視覚化（グラフ）

数値をグラフに置き換えることで、データ全体の動向や変化を直感的に把握できます。

2. 代表値による集約

「代表値」とは、ひとつの数値でデータ全体の傾向を示す指標です。主に以下の 3 種類がありま

す。

平均値：すべての値を合計し、データの数で割った値。

中央値：データを大きさ順に並べた際、ちょうど真ん中に位置する値。

最頻値：データの中で最も出現回数が多い値。

スライド
27

①データの「見え方」を知る
・平均値・中央値・最頻値・標準偏差などの基礎

平均値の計算（その1）

- 平均値を求めてください。
- 平均値は全部足して、個数で割ったもの。
- $55 \div 11 = 5$

5 7 5 6 3 4 4 5 6 5 5

27

具体的な数値データを用いて、平均値を算出する手順を確認します。

1.合計を求める

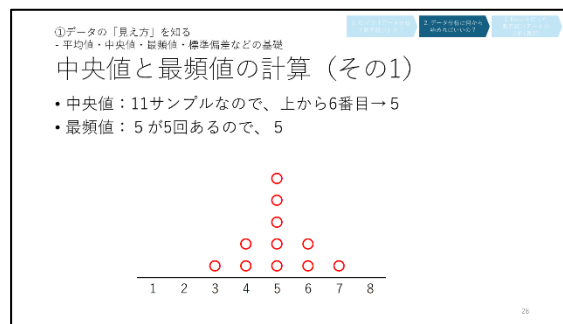
データ（11 個）をすべて足し合わせると、合計は 55 になります。

2.データ数で割る

合計値をデータの個数（11）で割ります。

$55 \div 11 = 5$ で、平均値は5となります。

スライド
28

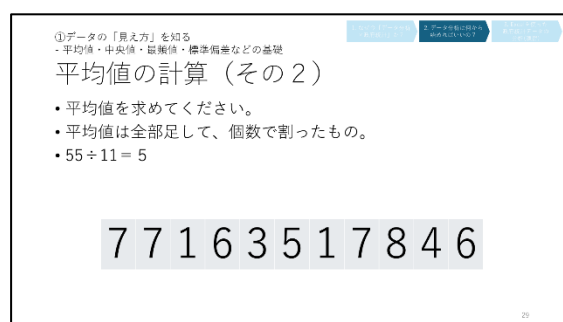


下の図は、先ほどのデータを「ヒストグラム」（各値の出現回数を可視化したグラフ）にまとめたものです。

中央値は、データを大きさ順に並べた際、真ん中に位置する値です。今回は全 11 サンプルあるため、「6 番目」にあたる 5 が中央値となります。

最頻値はデータの中で、最も出現回数が多い値のことです。今回の場合、「5」の出現回数が最も多いので、最頻値は 5 になります。

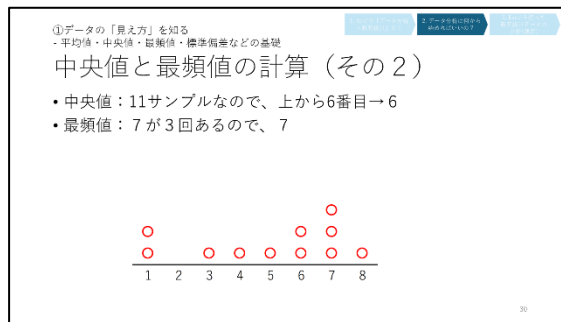
スライド
29



常に「平均値・中央値・最頻値」が一致するわけではありません。今度は異なるデータの平均値を求めてみます。

これも平均値を求めると、 $55 \div 11$ で 5 となります。

スライド
30

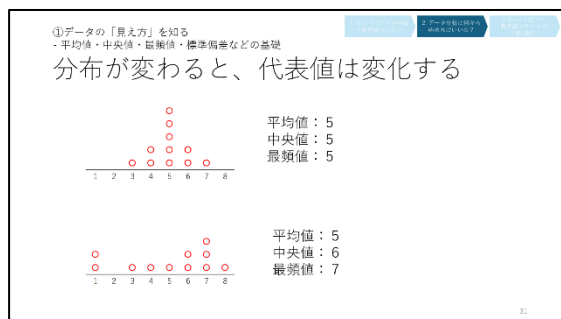


次に、同じようにヒストグラムを書いていきます。すると、先ほどの左右対称とは全く違う形になります。

中央値は、上から6番目の値です。6です。

最頻値は、7が最も多く出現していますので、7になります。

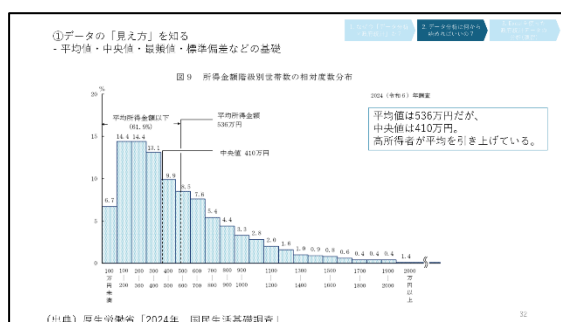
スライド
31



上の図のような左右対称な、「正規分布」やそれに近い分布の場合、平均値、中央値、最頻値は全て一致します。

一方で、下の図のように分布が少し歪んでいる（偏っている）データでは、これら3つの値にズレが生じます。

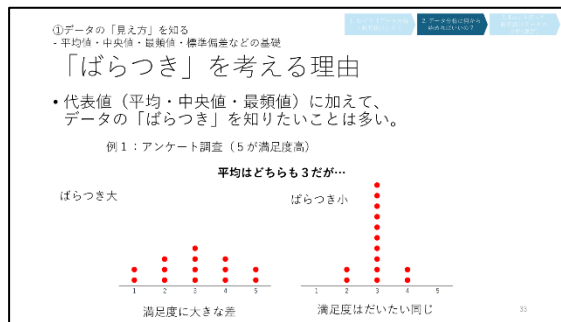
スライド
32



各世帯の所得分布を確認してみましょう。平均値は536万円ですが、実際には約6割の世帯が平均所得以下です。これは、少数の高所得者が全体の平均値を大きく引き上げているためです。

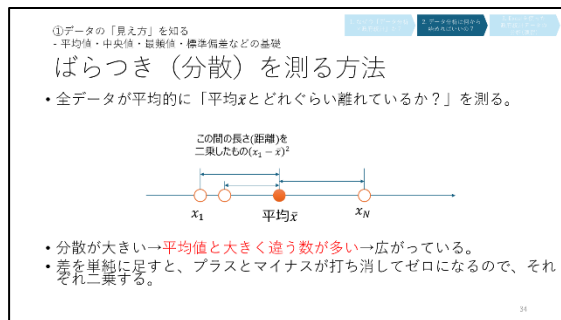
このように分布に偏りがある場合、所得の実態を見る時に平均値を用いるのはふさわしくない可能性があります。代わりに中央値（410万円）を確認すると、より実態を表していると言えるでしょう。

スライド
33



次に、ばらつきを確認することの重要性についてです。平均が同じでも、左のアンケート調査の結果は、ばらつきが大きく、満足している人もいれば不満に思っている人もいます。一方で 右の例では、「普通」という回答が最も多くなっています。同じ平均値の場合でも、ばらつきを把握することで、より丁寧な分析ができます。

スライド
34



データのばらつきを数値で表したものを「分散」と呼び、「平均からどの程度離れているか」を表します。「分散が大きい」ということは、「平均値から大きくかけ離れたデータが多く存在している」ということです。

スライド
35

①データの「見え方」を知る
・平均値・中央値・最頻値・標準偏差などの基礎

分散の公式

- データ $x_1, x_2, \dots, x_N$ の分散
 - 「平均値との差の二乗」の平均

平均値

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N}$$

分散を計算する際は、各データと平均値との差を二乗し、二乗した値の合計をデータ数で割ります。

スライド
36

①データの「見え方」を知る
- 平均値・中央値・最頻値・標準偏差などの基礎

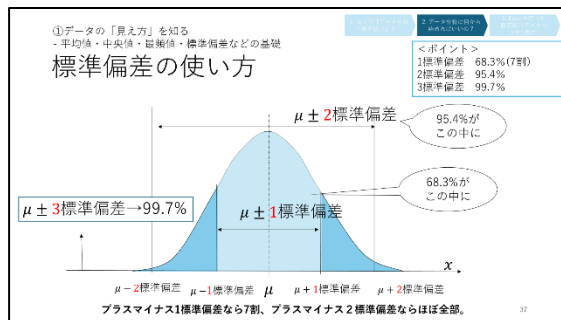
標準偏差もばらつき具合を表す

- 例
 - 分散が100なら、標準偏差は10
 - 分散が2なら、標準偏差は $\sqrt{2}$
 - 分散が0なら、標準偏差も0
- なので、標準偏差を二乗したら分散
 - 標準偏差が10なら、分散は100
 - 標準偏差が $\sqrt{2}$ なら、分散は2

・ばらつき具合を表すという意味では分散と同じ。分散の平方根をとったもの。
 ・あるデータがどのくらい平均から離れているかを表すときの物差しとなる。
 ・ばらつきの1単位。
 (例) 標準偏差の0.5倍ならあまり平均からはなれていない。2倍なら結構離れている…など。

「標準偏差」もデータのばらつきの指標です。分散の平方根（ルート）を取ったものが「標準偏差」で、「元のデータと同じ単位」になるので、標準偏差の方が、ばらつき具合が把握しやすいという利点があります。分散が 100 の場合は標準偏差が 10、分散が 2 の場合は標準偏差が $\sqrt{2}$ となります。

スライド
37



標準偏差の使い方ですが、図のようにデータが分布している場合、標準偏差を「ばらつきの単位」として計算すると、平均から±1標準偏差の範囲内には、全データの7割が含まれることが分かっています。さらに、平均から標準偏差の2倍ずつ（±2標準偏差）の範囲内には95.4%が含まれ、平均から標準偏差の3倍（±3標準偏差）の範囲内になると99.7%と、ほぼ全てのデータが含まれることになります。

スライド
38

①データの「見え方」を知る
・平均値・中央値・最頻値・標準偏差などの基礎

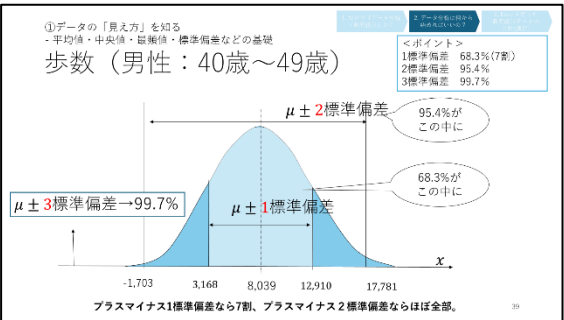
標準偏差

・（出典）厚生労働省「2023年 国民健康・栄養調査」40歳～49歳

	性別	平均値	標準偏差	平均-2 標準偏差	平均+2 標準偏差
体重 (kg)	男性	70.4	12.1	46.2	94.6
	女性	55.8	10.3	35.2	76.4
歩数 (歩)	男性	8,039	4,871	-1,703	17,781
	女性	7,172	4,218	-1,264	15,608

この図は、厚生労働省の 2023 年国民健康・栄養調査の 40 歳～49 歳の男女の数字を表したものです。40 歳～49 歳の男性の平均体重は 70.4kg です。女性は 55.8kg です。歩数を見ると、男性の平均は 8,039 歩、女性は 7,172 歩です。「男女どちらのばらつきが大きいか」に着目すると、標準偏差が大きい男性の方が体重・歩数ともにばらつきが大きい、ということが分かります。

スライド
39



平均から「2 標準偏差」分離れた範囲でデータを見ることもできます。例えば、男性体重の平均値 70.4kg から標準偏差 12.1kg の 2 倍 = 24.2 kg を引いた値、24.2kg を足した値の範囲に、約 95% のサンプルが含まれます。歩数についても同様に、「2 標準偏差」離れた範囲を見ると、標準偏差は約 5,000 歩（2 倍で 1 万歩程度）あります。このことから、ほとんど歩かない人から 2 万歩近く歩く人まで、非常に幅広い分布になっていることが分かります。

スライド
40

①データの「見え方」を知る
・ 平均値・中央値・最頻値・分散・標準偏差などの基礎

エクセルの関数を使って計算できます

代表値	関数
平均値	=AVERAGE(セルの範囲)
中央値	=MEDIAN(セルの範囲)
最頻値	=MODE(セルの範囲)
分散	=VAR.P(セルの範囲) =VARP(セルの範囲)
標準偏差	=STDEV.P(セルの範囲) =STDEVP(セルの範囲)

セルの範囲の示し方→

ここまで説明してきた平均値、中央値、最頻値、分散、標準偏差は、Excelを使えば、簡単に計算できます。

例えば平均値なら、「AVERAGE 関数」を使います。具体的には =AVERAGE (B2 : B5) のように、計算したいセルの範囲を指定すれば、計算が可能です。

スライド
41

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

散布図

- 2つの変数の関係を表す。1つの変数が増えた時、もう1つの変数が増えるかわかる。

姓	数学 (A)	英語 (B)
A	100	95
B	90	85
C	85	80
D	75	70
E	65	60
F	55	50
G	45	40
H	35	30
I	25	20
J	15	10
K	5	5
L	100	90
M	95	85
N	80	75
O	70	65

次に「散布図」について説明します。散布図は、2つのデータを1つのグラフに表すもので、例えばAさんの成績が数学100点・英語95点であれば、数学100点、英語95点が変わる場所にAさんの点をプロットします。このように、全員分のデータの組を全て点で表したものが右の図です。散布図にすると、2つのデータの間にどのような関係があるのかよく分かります。

スライド
42

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

散布図の作り方

- Microsoft Excelのグラフ機能に標準で搭載されている。

データの範囲を選択→挿入タブ→グラフ挿入画面で散布図を選択。

散布図もExcelで簡単に作成することができます。

後の演習でも扱いますが、Excelで2つのデータの範囲を選択し、左上の「挿入」タブから「グラフ挿入」画面を表示させて、散布図を選択します。

スライド
43

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

「相関」とは

2つの量に一定の関係があること

- 2つの量の関係性
 - 経済分析では因果関係を知りたいことが多い
 - 因果関係とまではいなくても、関係性があるかどうかをまず調べる
- 例
 - 気温とアイスの売り上げ高の関係性
 - 広告費と売上高
 - 価格と需要量
 - 運動量と休脂防率
 - 降水量と傘の売上高

相関とは、2つの量に関係があることを指します。例えば、「気温とアイスの売上高は関係があるのか」、「広告費を使うと売上が増えるのか」、「価格を上げると需要量が減るのか」といったことが挙げられます。

2つの変数の間に関係（相関）があるかどうかは、先ほどの「散布図」の形を見たり、「相関係数」という数値を算出したりすることで、比較的簡単に分かります。

スライド
44

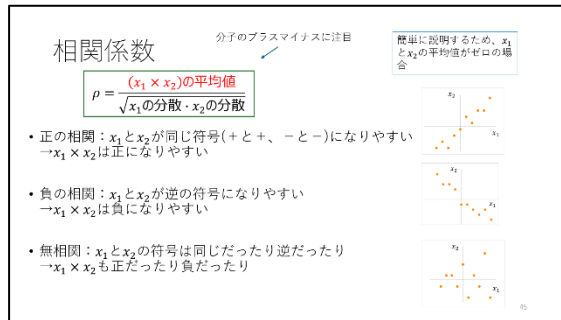
3種類の相関

- 正の相関
 - $x_1 \rightarrow \text{大}$, $x_2 \rightarrow \text{大}$
 - 例：身長と体重
- 負の相関
 - $x_1 \rightarrow \text{大}$, $x_2 \rightarrow \text{小}$
 - 例：身長と100m走のタイム
- 無相関
 - $x_1 \rightarrow \text{大}$, $x_2 \rightarrow \text{特段の傾向無し}$
 - 要するに、 x_1 と x_2 は無関係
 - 例：身長と年収

先ほどの数学と英語の点数の例のように、一方が増えると、もう一方も増える状態を「正の相関」と呼びます。散布図では、点が右上がりに分布するのが特徴です。逆に、一方が増えるともう一方が減少する状態を「負の相関」と呼びます。こちらは、散布図の点が右下がりに分布します。

また、一方が増えた時にもう一方が増えることも減ることもある（両者に関係がない）ような場合は、「無相関」と言います。

スライド
45



相関係数は以下の式で求められます。

$$\rho = \frac{(x_1 \times x_2) \text{の平均値}}{\sqrt{x_1 \text{の分散} \cdot x_2 \text{の分散}}}$$

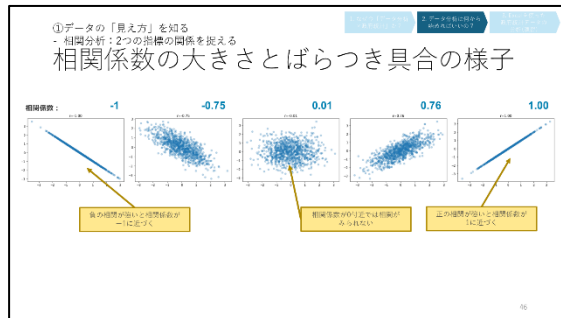
分子の±に注目してみると、

正の相関: $x_1 \times x_2$ は正になりやすい (x_1 と x_2 が同じ符号 (+と+、-と-))

負の相関: $x_1 \times x_2$ は負になりやすい (x_1 と x_2 が逆の符号)

無相関: 正の場合も負の場合もあります。

スライド
46



相関係数は、2 変数の関係の強さを示す指標です。値が「1」や「-1」に近いほど相関が強く、「0」に近いほど相関が弱いといえます。

左側の図: 一方が増えるともう一方が減る関係にあり、相関係数はマイナスの値 (負の相関) をとります。

右側の図: 一方が増えると、もう一方の変数も増える場合、相関係数が「1」になります。

真ん中の図: 両者に関係がない (無相関) 場合で、相関係数は「0」に近くなります。

スライド
47

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

相関を確認する

- 相関係数の関数はCORREL
- 2つの変数のセルを選ぶ
- =CORREL(セルの範囲,セルの範囲)

相関係数も Excel を使うと簡単に求められます。

使用するのは「CORREL 関数」です。

2 つのデータ（変数）を使うので、図のように「1 つ目のセル範囲」を左側に入力し、「2 つ目のセル範囲」を右側に入力すると計算することができます。

CORREL という名称は correlation（相関）の略称です。

スライド
48

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

コーヒーチェーン店出店数と総人口の相関係数

データの収集

- コーヒーチェーン店出店数
 - ホームページの店舗一覧から入力
- 総人口（e-Statを利用）
 - 統計データを活用する・地域
 - 市区町村データ・データ表示
 - 東京都・特別区
 - 基礎データ・A人口・世帯
 - A1101 総人口

例えば、「総人口とコーヒーチェーンの出店数には、何か関係がありそうだ」という仮説が立てられます。

そこで今回は東京都 23 区を例に、各区の「人口」と「コーヒーチェーン店の出店数」がどのような関係にあるのか、データを調べてみましょう。

スライド
49

①データの「見え方」を知る
・ 相関分析：2つの指標の関係を捉える

コーヒーチェーン店出店数と総人口の相関係数

- 散布図を描く
- 相関係数 =CORREL()
- = -0.40

コーヒーチェーン店の出店数と「総人口」の相関係数を、先ほどの「CORREL」関数で計算してみましょう。

コーヒーチェーン店の出店数のセル範囲を左側に入力し、総人口のセル範囲を右側に入力すると計算することができます。

計算すると、結果は-0.40 となります。マイナスというのは「人口が増えるとコーヒーチェーン店が減る」という負の相関を示していま

す。そこで散布図を見ると、千代田区・港区・新宿区といった区のデータが他の区とは異なる傾向にあることが分かります。これらの区はオフィス街が多く、昼間は多くの人が集まりますが、夜間の人口（総人口）は少ないのが特徴です。そのため総人口（夜間人口）を使った分析は適切ではないということが分かります。

スライド 50

①データの「見え方」を知る

- 相関分析：2つの指標の関係を捉える

②データの「見え方」を知る

- 相関分析：2つの指標の関係を捉える

コーヒーチェーン店出店数と昼間人口の相関係数

データの収集

- コーヒーチェーン店出店数
 - ホームページの店舗一覧から入力
- 昼間人口（e-Statを利用）
 - 統計データを活用する・地域
 - 市区町村データ・データ表示
 - 東京都・特別区
 - 基礎データ・A人口・世帯
 - A6107 昼間人口

そこで、今度は指標を切り替えて、コーヒーチェーン店の出店数と「昼間人口」の相関係数を計算してみることにします。

スライド 51

①データの「見え方」を知る

- 相関分析：2つの指標の関係を捉える

②データの「見え方」を知る

- 相関分析：2つの指標の関係を捉える

コーヒーチェーン店出店数と昼間人口の相関係数

- 散布図を描く
- 相関係数 = CORREL()
- = 0.71

先ほどと同様に「CORREL 関数」を使って、今度はコーヒーチェーン店の出店数と「昼間人口」の相関係数を計算してみましょう。コーヒーチェーン店出店数のセル範囲と昼間人口のセル範囲をそれぞれ入力すると、今度は相関係数が 0.71 となりました。先ほどの総人口（-0.40）とは対照的に、コーヒーチェーン店出店数と昼間人口には「正の相関」があることが分かります。昼間人口が増え

るとコーヒーチェーン店が増えるという関係性が明らかになりました。

スライド
52

①データの「見え方」を知る
・ 回帰分析：2つの変数の関係を捉える

回帰分析とは

- データ同士を結び付ける「魔法」

①原因の解明

- ある変数がほかの変数の原因になっているかどうかを判断することができる。

②予測

- ある変数を使って、ほかの変数を予測することができる。

52

次に回帰分析について説明します。先ほどの相関分析では、「2つの数に関係があるか」を相関係数やグラフで見ることで分析することができました。

回帰分析では、「ある変数が、ほかの変数の原因になっているか」を判断することができます。

また、ある変数を使って、ほかの変数を予測することが大きなメリットです。

スライド
53

①データの「見え方」を知る
・ 回帰分析：2つの変数の関係を捉える

回帰分析は、2つの変数の関係を式で表す

木の数が増えたら、収穫量はどうなるだろうか？

収穫量 Y_i ← 木の数 X_i

結果 ← 原因

原因に当たる変数を説明変数、結果に当たる変数を被説明変数と呼ぶ。

山澤成康『回帰分析から学ぶ計量経済学』 53

例えば、「木の本数が何本の時、収穫量が何個になるか」を予測する際、それを数式で示せるのが回帰分析の大きな特徴です。

回帰分析では、以下の2つの変数が重要です。

説明変数：結果となる変数に影響を与える「原因」となる変数（例：木の本数）。

被説明変数（目的変数）：分析の対象となる「結果」の変数（例：収穫量）。

スライド
54

①データの「見え方」を知る
- 回帰分析：2つの変数の関係を捉える

2つの変数を式で関係づけるのが回帰分析

これが回帰分析という呪文

回帰分析

$$Y = \alpha + \beta X + u$$

結果 原因

収穫量 木の数

Xが原因
Yが結果を表す

木の本数が○の本の時、収穫量は何個になるかを予測することができる

山澤成康『回帰分析から学ぶ計量経済学』

2 変数（原因と結果）の関係は、以下のような数式（回帰式）で表されます。

$$Y = \alpha + \beta X + \mu$$

X は木の本数（原因）、Y がリンゴの収穫量（結果）を指し、 α や β はその関係性を決める「係数」です。

スライド
55

①データの「見え方」を知る
- 回帰分析：2つの変数の関係を捉える

係数と誤差項

$$Y_i = \alpha + \beta X_i + u_i$$

係数 誤差項 (かく乱項)

回帰式を完成するためには、 α や β を知る必要がある。
X から計算した答えはぴったり Y にならないため、誤差項も必要。

山澤成康『回帰分析から学ぶ計量経済学』

数式から、 X_i を β 倍して α を加えたものがおおよそ Y_i になるということがわかりますが、回帰式を完成するためには、 α や β （係数）を知る必要があります

スライド
56

①データの「見え方」を知る
- 回帰分析：2つの変数の関係を捉える

単回帰分析

$$Y_i = \alpha + \beta X_i + u_i$$

タンカイキ!

$$Y_i = .99 + 0.52X_i$$

α の推定値 $\hat{\alpha}$ β の推定値 $\hat{\beta}$

α や β は「最小二乗法」という方法で計算できる。

山澤成康『回帰分析から学ぶ計量経済学』

α や β は「最小二乗法」という方法で計算できます。

これは、データから得られる実測値とモデルによる予測値の間の誤差の二乗和を最小にするようにモデルを決定する方法です。

スライド
57

①データの「見え方」を知る
・ 回帰分析：2つの変数の関係を捉える

具体的な例で考える

- 大学生の毎月のアルバイト収入と毎月の支出金額の例

	X アルバイト収入 (万円)	Y 支出額 (万円)
Aさん	4.0	3.1
Bさん	8.0	4.8
Cさん	12.0	6.9
Dさん	2.0	2.0
Eさん	6.0	3.9

回帰分析を使うと…

- 「収入が多い人は支出が多い。」ということを確認できる。
- 5万円の収入の人はどのくらい支出するかを予測することができる。

具体的な例として、アルバイトの「収入」と「支出額」を見てみましょう。AさんからEさんまでの収入と支出額を見てみると、収入が多いほど支出も多くなる傾向が分かります。

回帰分析を使うと、その関係がはっきり示されるだけでなく、データにはない、「5万円の収入の人はどのくらい支出するか」なども予測可能です。

スライド
58

①データの「見え方」を知る
・ 回帰分析：2つの変数の関係を捉える

散布図で描くと

直線が引けそうなので、1次関数の式で表せよう。

最小二乗法を使えば、2つの関係を図示できる。
また、切片と傾きが推定できて予測ができる。

左の図が、収入額と支出額をプロットした散布図です。

回帰分析とは、これらのプロットされた点を1本の直線で結ぶようなイメージになります。

スライド
59

①データの「見え方」を知る
・ 回帰分析：2つの変数の関係を捉える

実際のデータに当てはめて推定値を求める

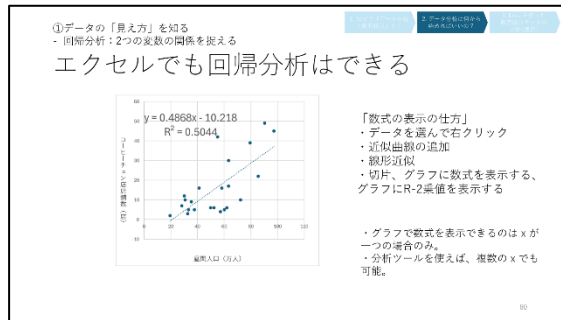
$$Y_i = \alpha + \beta X_i + u_i$$

切片 傾き 誤差項

山澤成康「回帰分析から学ぶ計量経済学」

誤差項 $\mu = \text{実測値} - \text{予測値}$ となり、データから得られる実測値とモデルによる予測値の間の誤差の二乗和を最小にするような式が「回帰式」となります。

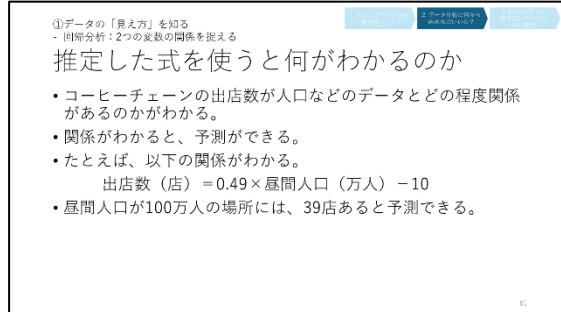
スライド
60



では、実際に回帰式を求めてみましょう。散布図を作成した後に、表示された各点を右クリックして、「近似曲線の追加」選びます。

表示された設定画面から、「線形近似」を選択し、「切片」、「グラフに数式を表示する」、「グラフにR-2乗値を表示する」にチェックを入れると、このように $y = 0.4868x - 10.218$ という回帰式がグラフの上に表示されます。 R^2 は決定係数と呼ばれるもので、0 から 1 までの値を取ります。1 に近いほど「当てはまりが良い」ことを示し、実績値をよく説明できている回帰式であるといえます。

スライド
61



推定した式を使うと、複数の変数の関係がわかるだけでなく、将来の予測もできるようになります。先ほどの回帰式を見ると、出店数（店）＝ $0.49 \times$ 昼間人口－10 ということが分かります。すると、例えば、昼間人口が 100 万人の場所では「39 店」と予測できます。新しく出店を検討する場合は、この予測値よりも既存の出店数が少なければ出店すべきであると判断できます。逆に、既に予測値以上の店舗が存在していれば、新たに出店するべきではない、といったデータに基づいた意思決定に活用できます。

スライド
62

②データ分析の基本姿勢
— 思い込みで左右されない判断

データ分析の落とし穴

- 「原因と結果」の関係ではないが、そのような関係に見えるパターンがある。

偽相関	原因と結果の関係にあるように見える二つの要素が、たまたまそのように見えるだけのパターン
第三の変数	原因と結果の両方にあるように見える二つの要素が、それらとは別の要素によって変動しているパターン
逆の因果関係	原因と結果の関係にあるように見える二つの要素が、逆になっているものと原因と結果が逆であるパターン

データ分析の落とし穴についてです。

データ分析を行う際に注意すべきなのは、「相関（数値の連動）」があるからといって、必ずしも因果関係があるとは限らない」という点です。数値がつながって見えても、実際には直接的な因果関係が存在しないケースが多々あります。

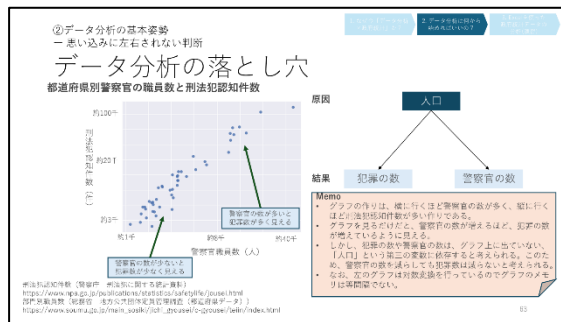
「偶然」による一致というのは、因果関係はないものの、データの動きが偶然一致するというケースです。例えば、海賊の数の減少と海水温の上昇に負の相関があるというデータがありますが、これらは全く別の要因で動いており、単なる偶然の一致に過ぎません。

「第三の変数」というのは、2 つの事象が原因と結果に関係があるように見えるものの、別の共通要素が双方に影響を与えているケースです。例えば、収入と体重だけを見ると比例して増加しているように見えますが、年齢が別の変数として存在しています。

また、「逆の因果関係」というのも存在します。

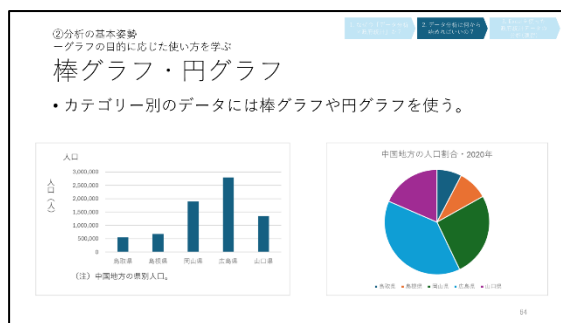
次のスライドで具体的な例を説明します。

スライド
63



例えば、警察官の数が多いほど犯罪件数が多いというデータがあったとします。しかしながら、実際は、「警察官を増やすと犯罪が増える」というわけではなく、犯罪の数や警察官の数は、グラフ上に出ていない、「人口」という第三の変数に依存すると考えられます。

スライド
64



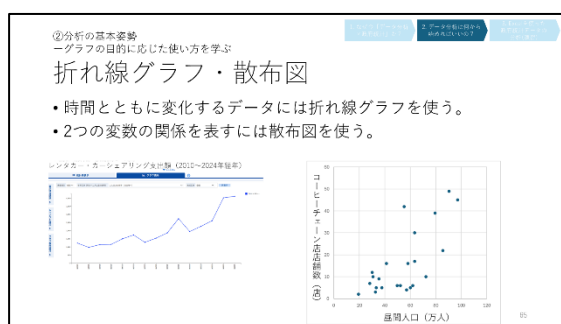
分析の基本姿勢として、グラフの種類とその使い分けについて簡単に整理します。

右と左は同じデータです。

棒グラフはデータの大小（量）を比較したい場合に使いやすいグラフです。

円グラフは、それぞれが全体のうちのどの程度を占めているのか、構成比を把握したい場合に使います。

スライド
65



折れ線グラフは、時系列など、時間とともに数値がどう変化するかを確認する場合に使い、変化の傾向（トレンド）などを把握するのに有効です。

散布図は、「コーヒーチェーン店」と「昼間人口」のように、2つの変数の間に関係性があるかを分析する際に役立ちます。

このように、「何を可視化して、何を伝えたいのか」という目的によってグラフを使い分けことが、データの正しい理解につながります。