公開データを分析してみよう（統計分析入門）

関西学院千里国際高等部　　河野光彦

統計分析の授業へようこそ！ 

統計分析の授業へようこそ！📊

高校最後の探究活動、始まりましたね！

皆さんはこれから、卒業まで長い時間をかけて、自分自身のテーマを深く探究していくことになります。その最初の３時間で、私たちは「データ分析」の基礎を学びます。

探究活動とデータ分析の関係

多くの人が「探究活動」と聞くと、「まずは面白そうなテーマを見つけること」と考えがちです。もちろんそれは大切ですが、面白いテーマを見つけたら、次にどうすればいいのでしょうか？

多くの研究は次のような流れで進みます。

1. **データ取得**：
   公開されている情報を集めたり、とりあえず測定したりする。

2. **分析方法の検討**：
   集めたデータをどうやって「**意味のある情報**」に変えるか考える。

3. **分析方法の学習**：
   分析方法を学んで自分の武器（スキル）にする。

4. **分析の実行**：
   実際にデータを分析する。

5. **結果の解釈・考察**：
   分析結果から何が言えるのかを考える。新たなスタートをする。

[3. 分析方法の学習]

分析に取り掛かる前に自分自身で学んでいかなければならない
（これ、けっこう時間がかかります。）

1. どんな分析ができるか先行研究などで調査する。
2. 分析方法を**自分で勉強**する。
3. 学んだ方法で分析する。

しかし、「分析方法を知らないから、なんか難しそう…」と感じて、せっかく集めたデータをうまく活かせないまま終わってしまうことがあります。

[5. 結果の解釈・考察]

より質の良いＲＱが生まれる
（ここが皆さんの研究のスタート地点です。）
ぼやけたRQからは ぼやけた結論しか出てきません

研究を始める前に、現実の問題を詳細に分析して、
質の良い課題研究テーマを見つけ出してください。

統計分析入門
授業のねらい

そこで、この授業では、皆さんが「分析を知らない」という壁にぶつからないように、基礎的な分析方法をたった１つだけ、みんなで一緒に学びます。

この授業を通して、「データを分析するって、こういうことか！」という感覚をつかんでほしいと思います。

1. **前処理**：
   データを分析できる形に整える作業

2. **相関係数**：
   2つの事柄の間の関係性を数値で見る方法

3. **重回帰分析**：
   いくつかのデータから未来を予測するモデルを作る方法

学ぶのは、この3つのステップです。

授業の流れ 

みんなで一緒にやってみよう！

自分で分析してみよう！

発展

去年の「理数探究」と同じことをするので、既にやった生徒は、「２．自分で分析してみよう！」から始めてください

1. みんなで一緒にやってみよう！
   まずは、公開されている気象データ（SSDSE）を使って、先生と一緒に手を動かしながら分析の基本を学びます。

1. 自分で分析してみよう！
   学んだ方法を使って、皆さんが興味のある別の公開データを分析してみます。

1. 発展
   自分で見つけた課題を解決するために、さらに高度な分析に挑戦することもできます。
   （多項式曲線フィッティングや主成分分析など）

統計分析の「壁」と「ご褒美」

正直にお伝えしておきます。データの前処理の段階は、非常に地味で、時には「なんでこんなことしてるの…？」と感じてしまうかもしれません。

データ分析は、最初が一番大変です。

しかし、この壁を乗り越えた先には、大きな「ご褒美」が待っています！

ご褒美その1：課題が明確に見える！

自分が本当に知りたいことは何なのか、何が分かれば研究が進むのかが、分析を通してはっきりします。

ご褒美その**2**：研究の質が格段に上がる！

勘や感覚ではなく、「**データ**」という確かな根拠に基づいて、説得力のある結論を導き出すことができます。

🌊✨

さあ、一緒にデータの海へ飛び込んでみましょう！

まとめ

統計分析は、数字やグラフをただ眺めることではありません。

それは、「**データから真実を見つけ出すための思考法**」です。

この授業で学んだツールを使って、皆さんの探究活動をより深く、より説得力のあるものにしていきましょう。

"In God we trust;
all others must bring data."

W. Edwards Deming

This is the most important takeaway that everyone has to remember.

Data speaks. But it's up to you to make it speak.

Thanks!

Contact us:

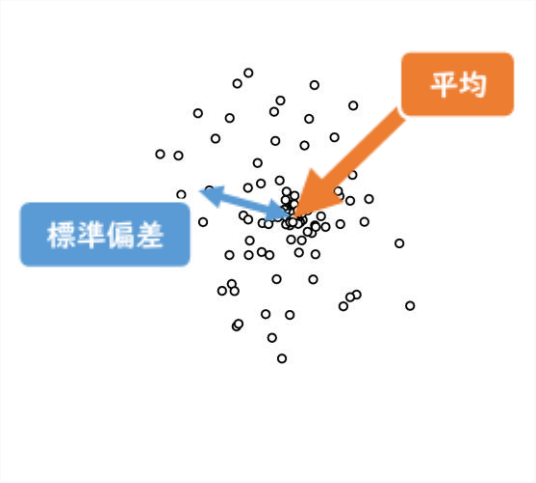
Mitsuhiko Kono
Yasutaka Kikuchi
Fumi Nagao
Hisashi Munemasa

じゃぁ早速はじめますか

SSDSE（教育用標準データセット）
https://www.nstac.go.jp/use/literacy/ssdse/

データを統計分析しよう

平均値・標準偏差・**相関係数・回帰分析**（最小二乗法）



平均

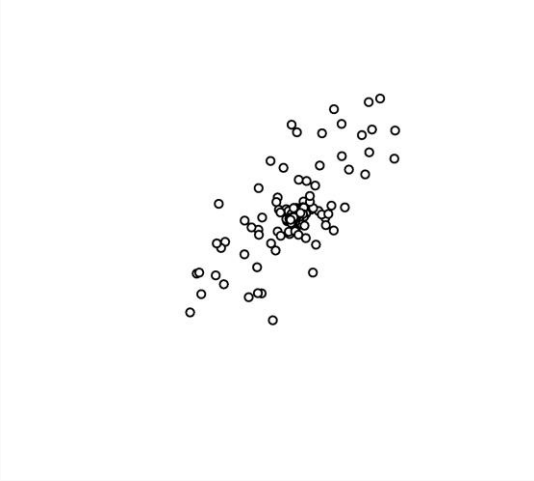
だいたいどこが真ん中か？

標準偏差

どこまで広がっているのか？

データを統計分析しよう

平均値・標準偏差・**相関係数**・**回帰分析**（最小二乗法）



相関係数

右上に伸びてるっぽい
どれぐらいそれが明らかなのか？

データを統計分析しよう

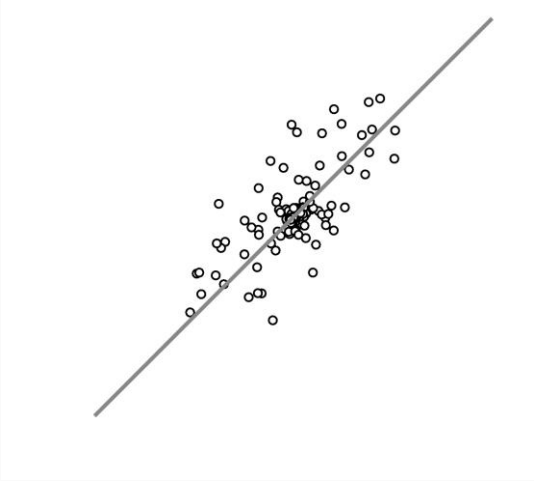
平均値・標準偏差・**相関係数**・**回帰分析**（最小二乗法）

相関係数

確実に右上に伸びてる
どれぐらいそれが明らかか？

データを統計分析しよう

平均値・標準偏差・**相関係数・回帰分析**（最小二乗法）



回帰分析（最小二乗法）

直線上に並んでいると考えると
だいたいどんな直線なのか？

この線がわかると色々なことが
予想できる

（２変数の場合がは単回帰分析）

前処理

実際に分析を行う前にデータを加工する工程

分析の前処理とは、分析を行う前に、生データを加工し、分析しやすい形に整える作業全般を指します。このエ程は、後続の分析の精度と品質を大きく左右する重要なステップです。

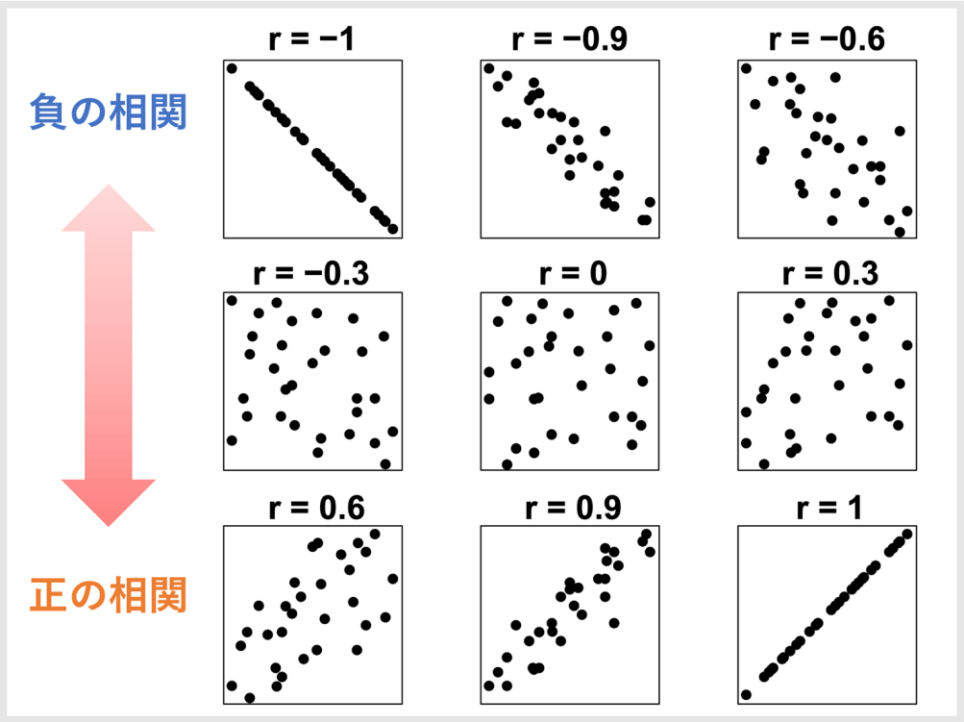
Googleスプレッドシートを使って，データの前処理をする。
（具体的には「**=QUERY()**」という関数を使う）

https://docs.google.com/spreadsheets/d/1bfZQBX4eXVHIdyq8QGGsj0dVy6LXfhhAAKTUS7L_7YQ/edit?usp=sharing

※Googleスプレッドシートが開けない方は、別添Excel「演習データ」をダウンロードしてください。

相関係数

どれぐらい明らかに右上がりか？



相関係数

どれぐらい明らかに右上がりか？

- 相関係数が「－1」に近づくほど「右下がり」が明らかになる
  （負の相関）

- 相関係数が「＋1」に近づくほど「右上がり」が明らかになる
  （正の相関）

Googleスプレッドシートを使って，データの相関係数を計算する。
（具体的には「=CORREL()」という関数を使う）

https://docs.google.com/spreadsheets/d/1bfZQBX4eXVHIdyq8QGGsj0dVy6LXfhhAAKTUS7L_7YQ/edit?usp=sharing

※Googleスプレッドシートが開けない方は、別添Excel「演習データ」をダウンロードしてください。

相関係数

どれぐらい明らかに右上がりか？

【利点】

- データ分析をするときに、まずやっておくべき手法
- 手っ取り早く分析できる

相関係数

どれぐらい明らかに右上がりか？

【注意事項】

- 抽出した標本のデータなら検定しなければならない
  (母集団の相関係数ではない)

- 因果関係ではない
  (同じ原因が他にあって相関関係が表れている可能性がある)

- 相関係数が高すぎると同じものを測っている可能性がある

相関係数

どれぐらい明らかに右上がりか？

因果関係ではない

意図的に相関関係を持ち出して強引に納得させようとする人がいるので要注意です。

健康に関する疑似相関で騙されやすい例には、サプリメントなどの民間療法で体調が改善したと感じるケース、特定の食品を摂るとガンが治ると誤解するケース、そして運動や食生活が健康に良いと関連付けて、それ以外の要素を無視するケースなどがあります。これらは、体調の変化が他の要因によるものである可能性や、がんの自然治癒が特定の療法と関連しているという誤解に基づいています。

相関係数

どれぐらい明らかに右上がりか？

因果関係ではない

同じ原因が他にあって相関関係が表れている可能性がある

隠れた変数を見つけ出すために、たくさんの項目についてデータを集める必要がある。

因果関係を調べたいなら、統制群（特定の要因の影響を評価するため）や対照群（特定の条件が結果に与える影響を評価するため）を作り、実験群と比較する必要があります。

重回帰分析

いくつかのデータから未来を予測するモデルを作る方法

たくさんの項目についてデータを集めたら

ある「目的変数」に対して、複数の「説明変数」がそれぞれどの程度影響を与えているかを数式で表し、その関係性を分析・予測する統計手法（重回帰分析）で分析をします。

単一の要因で結果を説明する単回帰分析とは異なり、複数の要因が複合的に影響し合う複雑な関係を分析できます。

重回帰分析

いくつかのデータから未来を予測するモデルを作る方法

既知の［y の範囲］と［x_1 の範囲］，［x_2 の範囲］，… をもとに、回帰直線

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots\cdots$$

求め、係数（直線の傾き：a_1，a_2，…）や定数項（切片：a_0）を求めます。

Googleスプレッドシートを使って，データの重回帰分析をする。
（具体的には「=LINEST()」「=TREND()」という関数を使う）

https://docs.google.com/spreadsheets/d/1bfZQBX4eXVHIdyq8QGGsj0dVy6LXfhhAAKTUS7L_7YQ/edit?usp=sharing

※Googleスプレッドシートが開けない方は、別添Excel「演習データ」をダウンロードしてください。

自分で分析してみよう！

データのありか：「SSDSE」「e-Stat」「RESAS」など

- https://www.nstac.go.jp/use/literacy/ssdse/
  まずは SSDSE の興味あるデータで練習してから始めてください
- https://www.stat.go.jp/info/guide/public/kouhou/index.html
- https://www.stat.go.jp/info/guide/public/kouhou/edu/index.html
- https://www.e-stat.go.jp/
- https://resas.go.jp/

例えば、ChatGPT や Gemini で「○○に関するデータは RESAS や e-Stat にありますか」などというように聞いてみてください。

発展

自分で見つけた課題を解決するために
（多項式曲線フィッティング）

既知の［yの範囲］と［xの範囲］をもとに、回帰直線

$$y = a_0 + a_1 x + a_2 x^2 + \cdots\cdots$$

求め、係数（直線の傾き：a_1, a_2, …）や定数項（切片：a_0）を求めます。

Googleスプレッドシートを使って，データの前処理をして
多項式曲線フィッティングをする。
（具体的には同じ「=LINEST()」という関数を使う）

発展

自分で見つけた課題を解決するために
（主成分分析）

主成分分析とは、多数の変数の持つ情報を、より少ない「主成分」と呼ばれる新たな合成変数に要約する統計的データ解析手法です。例えば、ブランドの分類などに使うことができます。マーケティングなどでよく使われます。

Pythonを使って，プログラミングをする。

発展

自分で見つけた課題を解決するために
（検定）

観察や実験から得られたデータに基づき、ある仮説が統計的に正しいか、あるいはその差が偶然によるものか否かを確率的に判断する方法です。具体的には、仮説を設定し、標本データから計算される統計量（検定統計量）と、あらかじめ定められた基準（有意水準）を比較することで、仮説を棄却するか否かを決定します。このプロセスを通じて、データ間の関係性が偶然なのか、それとも統計的に意味のあるものなのかを客観的に判断し、データに基づいた意思決定が可能になります。

Googleスプレッドシートを使って，P値（帰無仮説が正しいとした場合に、観測されたデータ以上の極端な結果が得られる確率）を計算する。

まとめ

統計分析は、数字やグラフをただ眺めることではありません。

それは、「**データから真実を見つけ出すための思考法**」です。

この授業で学んだツールを使って、皆さんの探究活動をより深く、より説得力のあるものにしていきましょう。

This is the most important takeaway that everyone has to remember.

Data speaks. But it's up to you to make it speak.

Thanks!

Contact us:

Mitsuhiko Kono
Yasutaka Kikuchi
Fumi Nagao
Hisashi Munemasa