情報リテラシー（第10回 後期）

データ分析と仮説検証サイクル

今日のねらい

- 仮説検証サイクルの考え方を理解できる
- 散布図と相関の関係を理解できる
- Pythonで線形回帰を実装できる
- 予測モデルの可能性と限界を知る

仮説検証サイクルとPDCA

仮説検証サイクル（科学）

仮説 → 検証 → 分析 → 結論 → 新仮説

PDCA（ビジネス）

Plan → Do → Check → Act

どちらも
一度で完璧を目指さず、繰り返し改善する

仮説検証サイクルの4ステップ

ステップ1：仮説（Hypothesis）

「〇〇なら△△になるはず」と予想を立てる

ステップ2：検証（Experiment/Test）

データで仮説が正しいか確かめる

ステップ3：分析（Analysis）

結果を詳しく調べる

ステップ4：結論（Conclusion）

仮説は正しい？修正が必要？

今日の仮説

仮説

「打率が高い選手ほど、本塁打も多い」

この仮説をデータで検証します！

検証方法

- **散布図**で視覚的に確認
- **相関係数**で数値的に確認
- **線形回帰**で予測モデルを作成

散布図と相関係数（復習）

散布図（Scatter Plot）

2つのデータの関係を見るグラフ

相関係数

相関の強さを数値化（-1 〜 +1）

今日のデータ

打率 vs 本塁打 = **約0.7前後**

→ **比較的高い正の相関**

相関の注意点

相関 = 因果ではない

相関が高くても、
「打率が本塁打を増やしている」とは断定できない
他の要因が関係する可能性がある
（体格、スイング、戦術など）

線形回帰で予測

散布図に「最も当てはまる直線」を引く方法

数式

y = a x + b

- x：説明変数（打率）
- y：目的変数（本塁打）
- a：傾き
- b：切片

回帰式と決定係数（R^2）

今回のデータでは：

相関係数：約0.7前後
決定係数（R^2）：約0.5〜0.6

→ 比較的強い相関がある
→ でも完全には説明できない（約半分程度）

※ 正確な値はColabで確認しましょう

回帰予測の例

回帰式を使って予測：

打率.300の選手

計算すると → 約15〜20本程度

打率.250の選手

計算すると → 約10〜15本程度

※ 正確な予測値はColabで確認！
※ 直線から遠い点ほど外れやすい

回帰直線とデータの関係

直線に近い点： 予測が当たりやすい
直線から遠い点： 予測が外れやすい

外れ値

- 打率低いのに本塁打多い
- 打率高いのに本塁打少ない

外れ値は「例外の発見」につながる

サイクルの結果

仮説：
「打率↑ → 本塁打↑」

検証結果：

- 散布図：右上がり
- 相関係数：**約0.7前後**
- 回帰：予測モデル作成
- R^2：**約0.5〜0.6（半分程度を説明）**

結論：
おおむね正しいが、完全ではない

次のサイクルへ

発見：外れ値がある
例：

- 打率低いのに本塁打多い
- 打率高いのに本塁打少ない

新しい仮説：

- 「打撃スタイル別に見るべき？」
- 「体格も考慮すべき？」
- 「長打率の方が関係が強い？」

ビジネスでも同じ発想

データ分析はスポーツだけではない

- **打率 = 商談成功率**
- **本塁打 = 売上**
- 外れ値 = 超成績者（化け物営業）

仮説 → 検証 → 改善
仕事でも使う思考

外れ値の影響

外れ値があると：

- 回帰直線がずれる
- 予測精度が下がる
- 一般パターンを見逃す可能性

しかし
外れ値は新しい発見につながる

データの紹介

野球選手の成績データ（架空）

- セ・リーグ：20人
- パ・リーグ：20人
- 合計：40人

項目：
選手名、リーグ、打率、本塁打、打点、安打、出塁率
教育的に外れ値も含まれている

今日のColab実習

必ず次の順番で実行する：

1. データを読み込む（pandas）
2. 散布図を描く
3. **corr()** で相関を計算
4. **train/test分割** ← 重要！
5. **LinearRegression()** で回帰
6. **predict()** で予測

訓練データとテストデータ

なぜ分ける？

訓練データ（Training Data）

- モデルの学習に使う（60%）
- このデータでパラメータを決める

テストデータ（Test Data）

- モデルの評価に使う（40%）
- **未知のデータ**として扱う

目的

モデルが未知のデータでも予測できるか確認

課題

データ分析実習

提出物：Wordファイル1つ

1. セル6の実行結果のスクリーンショットを貼る
   - 散布図
   - 予測結果
2. 外れ値が出た理由を1行で書く
   例：「打撃スタイルの違いによるもの」

提出方法： WordファイルをTeamsに提出

まとめ

- 仮説検証サイクルは改善の基本
- 今日の仮説：**打率↑ → 本塁打↑**
- 相関係数：
- R^2：
- 外れ値：
- 訓練/テスト分割の重要性

仮説 → 検証 → 改善 → 新仮説…

Teamsに感想を書いてください