

# 高校数学とJulia言語 Day 4

---

データの可視化と統計処理

城北中学校・高等学校 中学3年・高校1年

夏期講習会III 2025/8/24～2025/8/28

担当：清水団



## 5日間の学習予定






- **Day 1** : Google Colabの紹介・基本計算 ✓
- **Day 2** : 関数のグラフの描画 ✓
- **Day 3** : 最適化（最大・最小） ✓
- **Day 4** : データの分析 ← 今日はこちら！
- **Day 5** : 確率・シミュレーション

今日のゴール：データを読み取り、統計的に分析できるようになるう！

## データサイエンスの重要性

現代社会では、データに基づいた判断が重要！

### 実社会での応用例

-  **スポーツ**：選手のパフォーマンス分析
-  **医療**：病気の早期発見、治療効果の測定
-  **教育**：学習効果の測定、個別指導の最適化
-  **ビジネス**：売上予測、顧客分析
-  **環境**：気候変動の分析

グラフでデータを可視化し、統計量で定量的に分析する方法を学びます！

## 今日使うパッケージ

統計処理に必要なツールを準備

```
using Plots          # グラフ描画
using Statistics      # 統計関数 (mean, std など)
using StatsBase       # 統計分析 (mode, cor など)
using Random          # 乱数生成
using StatsPlots      # 統計グラフ作成 (boxplot)

# フォント設定
gr(fontfamily="ipam")
```

新しいパッケージがたくさん！でも使い方は簡単です

## 基本統計量を計算してみよう

### 問題1：テストの点数データの分析

```
# あるクラスの数学のテストの点数
test_scores = [85, 92, 78, 88, 95, 82, 90, 87, 83, 91,
               76, 89, 94, 80, 86]
```

### 基本統計量を一気に計算！

```
println("平均値:", round(mean(test_scores), digits=2))
println("中央値:", median(test_scores))
println("標準偏差:", round(std(test_scores), digits=2))
println("最大値:", maximum(test_scores))
println("最小値:", minimum(test_scores))
```

## 基本統計量の意味

### 代表値（データの中心）

- **平均値**：全データの合計÷個数
- **中央値**：データを並べた時の真ん中
- **最頻値**：最も多く現れる値

### 散らばりの指標

- **分散**：データの散らばりの大きさ
- **標準偏差**：分散の平方根
- **範囲**：最大値 - 最小値
- **四分位範囲**：上位25%～75%の範囲

これらの数値から、データの特徴を読み取ることができます！

## ヒストグラムでデータを可視化

データの分布を視覚的に表現

```
# 基本的なヒストグラム
histogram(test_scores, bins=5,
           title="数学テストの点数分布",
           xlabel="点数", ylabel="人数",
           color=:skyblue, alpha=0.7)

# 平均値の線を追加
vline!([mean(test_scores)], lw=3, color=:red, label="平均値")

# 中央値の線を追加
vline!([median(test_scores)], lw=3, color=:green,
       label="中央値", linestyle=:dash)
```

平均値と中央値の位置を比較してみましょう！

## 散布図で関係性を探る

### 問題2：学習時間と成績の関係分析

```
# 学習時間と成績のデータ
study_hours = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...]
exam_scores = [45, 55, 60, 68, 75, 80, 85, 88, 92, 95, ...]

# 相関係数を計算
correlation = cor(study_hours, exam_scores)
println("相関係数:", round(correlation, digits=3))

# 散布図の作成
scatter(study_hours, exam_scores,
        title="学習時間と試験成績の関係",
        xlabel="学習時間 (時間)", ylabel="試験成績 (点)")
```



## 相関係数の意味

### 相関係数の値と関係の強さ

相関係数の値	関係の強さ	意味
<b>+1.0</b>	完全な正の相関	一方が増えると他方も必ず増える
<b>+0.7～+0.9</b>	強い正の相関	一方が増えると他方も増える傾向
<b>+0.3～+0.7</b>	中程度の正の相関	やや関係がある
<b>0付近</b>	相関なし	関係がない
<b>-0.3～-0.7</b>	中程度の負の相関	一方が増えると他方は減る傾向
<b>-1.0</b>	完全な負の相関	一方が増えると他方は必ず減る

相関関係があっても因果関係があるとは限らない！

## 回帰直線を描いてみよう

### データの傾向を直線で表現

```
# 最小二乗法で回帰直線の係数を求める
x_mean = mean(study_hours)
y_mean = mean(exam_scores)

# 回帰直線の傾き
slope = sum((study_hours .- x_mean) .* (exam_scores .- y_mean)) /
        sum((study_hours .- x_mean).^2)

# 回帰直線の切片
intercept = y_mean - slope * x_mean

# 散布図 + 回帰直線
scatter(study_hours, exam_scores, label="データ点")
x_line = 0:0.1:10
y_line = slope .* x_line .+ intercept
plot!(x_line, y_line, color=:red, lw=3, label="回帰直線")
```

## 複数データの比較

### 問題3：2つのクラスの成績比較

異なるグループを比較してみよう

```
# 2つのクラスのテスト結果
class_a = [75, 80, 85, 78, 82, 88, 76, 84, 79, 87, ...]
class_b = [70, 85, 90, 72, 88, 92, 74, 86, 73, 91, ...]

# 統計量を比較
println("Aクラス平均:", round(mean(class_a), digits=2))
println("Bクラス平均:", round(mean(class_b), digits=2))
println("Aクラス標準偏差:", round(std(class_a), digits=2))
println("Bクラス標準偏差:", round(std(class_b), digits=2))
```

平均だけでなく散らばりも比較が重要！

## 箱ひげ図で比較

データの分布を一目で比較

```
# 箱ひげ図で2つのクラスを比較
boxplot(class_a, label="Aクラス", color=:blue)
boxplot!(class_b, label="Bクラス", color=:green,
          title="2クラスの箱ひげ図", ylabel="点数")
```

### 箱ひげ図の読み方

- 箱の中央線：中央値
- 箱の上下：第1四分位数と第3四分位数
- ひげ：最大値・最小値（外れ値除く）
- 点：外れ値

## 実際のデータ分析例

### アイスクリーム店の売上分析

#### 設定：気温と売上の関係を調べる

- 15日間の気温と売上データ
- 気温が高いほど売上が上がると予想
- 相関係数と回帰直線で関係を分析

```
# データ生成（実際の分析では実データを使用）
temperatures = 20 .+ 8 * rand(15) # 20-28℃
sales = 1000 .+ 150 * temperatures .+ 300 * randn(15)

# 相関分析
println("気温と売上の相関係数：", round(cor(temperatures, sales), digits=3))
```

## データ分析の手順

### 標準的なアプローチ

1. データの概要把握：基本統計量を計算
2. 分布の可視化：ヒストグラムで分布を確認
3. 関係性の探索：散布図で変数間の関係を調査
4. 定量的分析：相関係数や回帰分析で関係を数値化
5. 結果の解釈：統計的な結果から実用的な結論を導出

### 重要なポイント

- 可視化が第一：まずグラフで全体像を把握
- 数値で確認：主観的な印象を客観的な数値で検証
- 複数の指標：一つの統計量だけでなく複数の角度から分析

## 実習：データ分析を体験しよう

実際に**Google Colab**で以下を試してみましょう

1. 基本統計量の計算：平均、標準偏差、四分位数
2. ヒストグラムの作成：データの分布を可視化
3. 散布図の作成：2変数の関係を調査
4. 相関分析：相関係数と回帰直線の計算
5. 複数グループの比較：箱ひげ図による比較

データから新しい発見をしてみましょう！

## 本日の演習問題

### 問題1: 基本統計量の計算

高校生20人の身長データから統計量を計算し、ヒストグラムと箱ひげ図を作成

### 問題2: 相関分析

数学と物理の成績データから相関係数を計算し、散布図と回帰直線を作成

各問題で以下を実施： - 基本統計量の計算 - 適切なグラフの作成 - 結果の解釈と考察

データから何が読み取れるか考察してみましょう！



## 演習問題を解いてみよう！

**Google Colab**を開いて、実際にデータ分析してみましょう

### 取り組み方

1. データを確認する
2. 基本統計量を計算する
3. グラフを作成する
4. 関係性を分析する
5. 結果を解釈する
6. 実用的な提案を考える

データ分析は探偵の仕事に似ています。データという証拠から真実を見つけましょう！

## 🌟 今日のまとめ

### 学んだ統計の基本

- 基本統計量：平均値、中央値、標準偏差、四分位数
- データの分布：ヒストグラムによる可視化
- 関係性の分析：相関係数と回帰直線

### データ可視化の技術

- ヒストグラム：データの分布を表現
- 散布図：2変数の関係性を可視化
- 箱ひげ図：複数グループの比較

### 重要な考え方

- ✓ 可視化による直感的理解
- ✓ 統計量による定量的分析
- ✓ 相関関係と因果関係の区別

## 次回予告：Day 5

### 確率・シミュレーション

- ランダムな現象のモデル化
- 確率的な予測とシミュレーション
- モンテカルロ法による数値計算
- 実際の問題への応用

今日学んだデータ分析技術を使って、不確実な現象を分析します！

### 宿題

今日の演習問題を完成させて、Google Classroomに提出してください。

## 💡 発展的な内容（時間がある人へ）

### より高度な統計分析

```
# 多変数の相関分析
using LinearAlgebra
data_matrix = [math_scores physics_scores chemistry_scores]
correlation_matrix = cor(data_matrix)

# ヒートマップで相関関係を可視化
heatmap(correlation_matrix,
        xlabel="科目", ylabel="科目",
        title="科目間相関ヒートマップ")
```

複数の変数の関係を同時に分析できます！

## ? 質問タイム

何か分からないことはありませんか？

- 統計量の意味や解釈
- グラフの作成方法
- 相関分析の手法
- 演習問題について
- その他、何でも！

データ分析は実用的なスキルです。疑問があれば積極的に質問しましょう！

お疲れさまでした！