吉岡 隆宏，紺野 剛史

本章では，統計学の中でも特にデータサイエンスにおいて重要な役割を果たす要素に焦点を当てています．平均値や標準偏差といった基本的な統計量から，共分散や相関係数といった関係性を示す指標，さらにはデータ補完や正規化といったデータの前処理技術まで，原理や数式を交えて幅広く取り上げています．

統計の計算においては，データが持つ意味や前提，構造を理解して処理を実行することが重要です．例えば，平均値や標準偏差はデータの傾向を把握するための基礎となり，共分散や相関係数は変数間の関係性を明らかにします．また，データ補完や正規化は，欠損値の補填やスケールを統一するといったデータの処理によりデータ全体の品質を保証し，より信頼性の高い分析を可能にします．

これらの統計の手法は，ビジネスにおいて市場のトレンド予測，あるいは交通の分野で渋滞予測や最適ルートの提案，エネルギー消費予測などさまざまな分野で利用されています．

7-1 統計量（平均値，分散，標準偏差，期待値）

表1
4人のゲームの得点

	得　点
A	300
B	200
C	400
D	100

得点の平均 $\mu = 250.0$
得点の分散 $V = 12500.0$
得点の標準偏差 $\sigma = 111.8$

● 概要

平均，分散，標準偏差，期待値は，データの特徴を表す値です．

● 仕組み

▶平均値

平均値は複数のデータを足し合わせ，データの数で割ることで算出します．μを平均値，x_kをデータ，nをデータの個数とすると，次の式で定義されます．

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k \cdots\cdots (1)$$

▶分散，標準偏差

分散と標準偏差はどちらもデータのばらつき具合を表すものです．Vを分散，σを標準偏差とすると，それぞれ次の式で定義されます．

$$\text{分散} \quad V = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2 \cdots\cdots (2)$$

$$\text{標準偏差} \quad \sigma = \sqrt{V} = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2} \cdots\cdots (3)$$

分散は，平均値と各データの差の2乗を元に計算するのに対し，標準偏差は分散の平方根を算出するため，標準偏差の単位はデータと同じになります．そのため，標準偏差の方がデータの特徴を直感的につかみやすくなることが多いです．

なお，標準偏差にはデータ数nではなく，$n-1$で除算する不偏標準偏差があります．目的によって使い分ける必要があります．データ数nで除算する標準偏差は，データの，平均値からのばらつきを表します．

リスト1　平均，分散，標準偏差，期待値の算出

```
import numpy as np

x = np.array([300, 200, 400, 100]) # データ
P = np.array([0.5, 0.3, 0.2, 0.1]) # 各データが得られる確率

x_ave = np.average(x) # xの平均値
x_var = np.var(x) # xの分散
x_std = np.std(x) # xの標準偏差
x_E = np.sum(P*x) # xの値がそれぞれPの確率で得られるときの期待値
```

一方，不偏標準偏差は，データが大きな母集団から取り出された標本であるとして，標本の平均値が，母集団の平均値（真の平均値）からどのくらいばらついているかを表します．ただし，本稿ではデータのばらつきを表す標準偏差のみを扱います．

▶期待値

期待値は，データが確率分布であるとみなしたときの中心を表します．値の平均的な出現傾向を示します．Eを期待値，x_kをデータ，P_kをx_kが得られる確率とすると，次のように計算できます．

$$E = \sum_{k=1}^{n} P_k x_k \cdots\cdots (4)$$

事前に確率が分かっているかどうかという違いがあります．データがあらかじめ分かっている場合は平均と期待値は同じ値になります．

● コード

リスト1に本節で説明した統計量を計算するコードを示します．NumPyというライブラリで利用できます（Pandasやstatisticsなどでも可能）．

▶平均値，分散，標準偏差の算出

表1にAさん，Bさん，Cさん，Dさんのあるゲームの得点の結果を表します．**リスト1**に示したプログラムでの計算の結果，得点の平均値は250.0点，得点の分散は12500.0，得点の標準偏差は111.8です．