機械学習で知っておくべき統計学

THE POWER TO KNOW.

18PD0

機械学習で
知っておくべき統計学

コースノート

Statistics You Need to Know for Machine Learning
スライドおよびノート SAS® Viya® リリース LTS 2025.03 に基づくコンテンツ

コースコード EVST151、作成日 2025 年 11 月 4 日。http://support.sas.com/training/

本コースは米国コースノート Statistics You Need to Know for Machine Learning (コースコード EVST151)を元に作成しています。

目次

Lesson 1: 統計学と機械学習

ビッグデータと機械学習における統計学の重要性

ビッグデータと機械学習における統計学の重要性を理解する

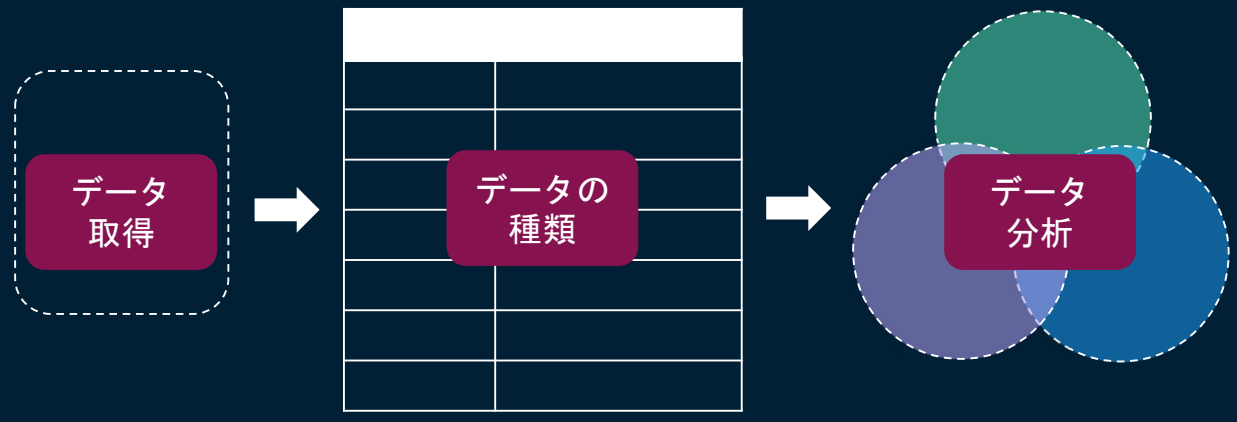
Lesson 01, Section 01

データとデジタル経済

"統計的思考は、いつの日か、読み書きの能力と同様に、有能な市民であるために不可欠なものとなるだろう。"

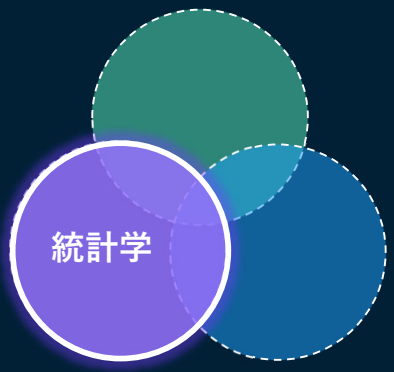
― サミュエル・S・ウィルクス（1951年）、ハーバート・G・ウェルズ（1903年）*の言葉を意訳*

データとデジタル経済



データとデジタル経済

実験データ	
目的	研究
価値	科学的
生成	能動的な制御
サイズ	小
衛生	クリーン
状態	静的
例	医薬品の臨床試験



統計学とは、実証的データを**収集、分析、解釈、提示し**、関連する不確実性を考慮に入れる科学である。

データとデジタル経済

機会的データ	
目的	運用
価値	ビジネス
生成	受動的な観察
規模	大
衛生	ノイズが多い
状態	動的
例	銀行データ

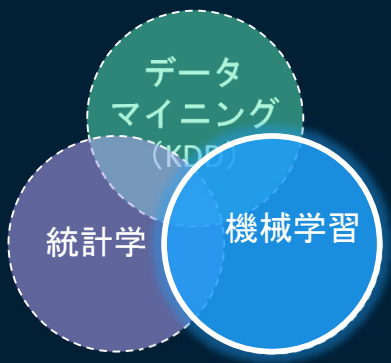
データマイニングとは、**大規模なデータセットを分析してパターンや関連性を特定し、将来の傾向を予測する**プロセスであり、データ分析を通じてビジネス上の課題解決に役立てることができます。

機会的データ（Huber 1997）

データとデジタル経済

従来の方法では処理が困難、あるいは不可能なほど、大規模、高速、または複雑なデータ

ビッグデータ	
目的	新しい形態の情報
価値	ビジネス
生成	**受動的な観察**
規模	極めて巨大
衛生	**非常にノイズが多い**
状態	ストリーミング
例	画像、音声、動画、テキスト



機械学習は、多くの統計学やデータマイニングの概念や手法を用いて、人間の介入を最小限に抑えながらデータから**反復的に学習**できるモデル構築を自動化します。

ビッグデータとは何ですか？

質問：ビッグデータ

次のうち、ビッグデータの例となるものはどれですか？（該当するものすべてを選択してください）

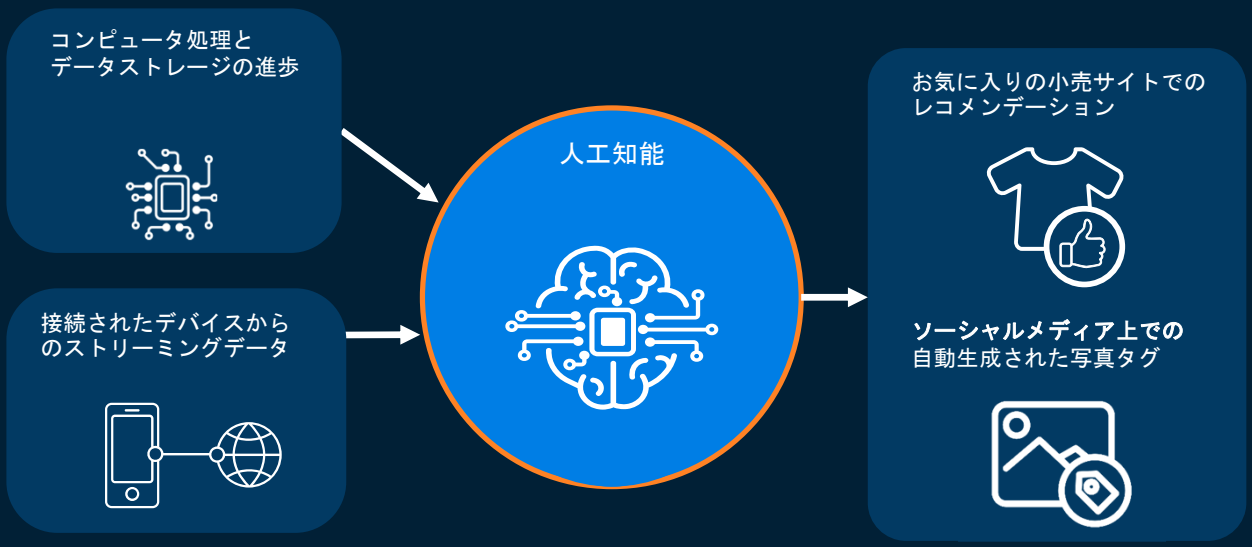
a. 管理された臨床薬物試験のサンプル
b. ソーシャルメディアのフィード
c. 市内の道路に設置されたカメラから送信される映像
d. 同じ植物から2つのサンプルを採取し、一方を日光に当て、もう一方を日光から遠ざけておく

回答：ビッグデータ

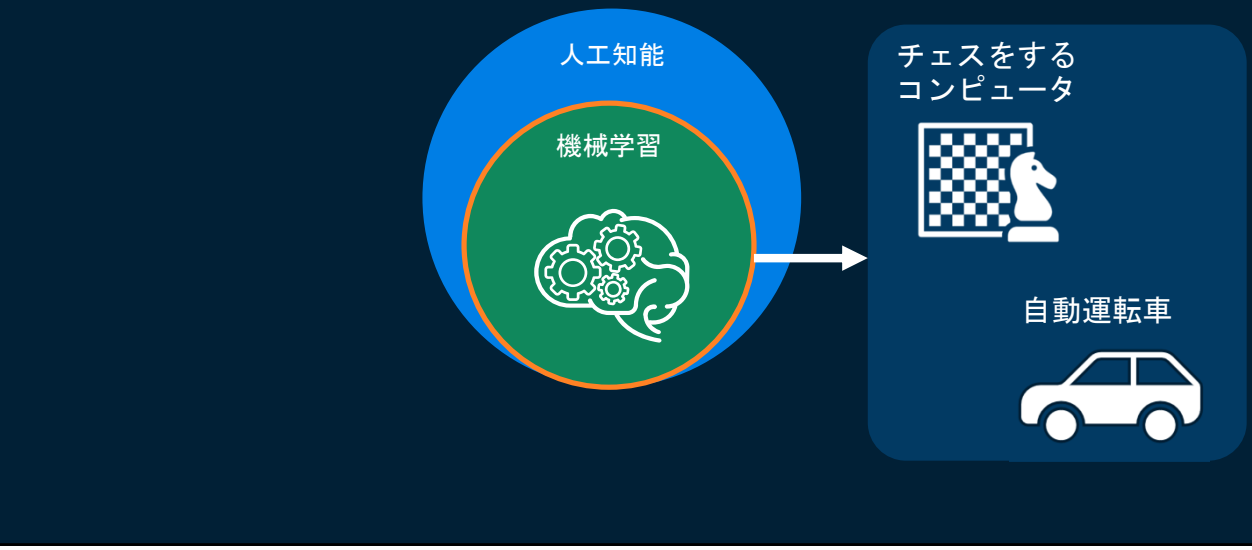
次のうち、ビッグデータの例となるものはどれですか？（該当するものすべてを選択してください。）

a. 管理された臨床薬物試験のサンプル
b. ソーシャルメディアのフィード
c. 市内の道路に設置されたカメラから送信される映像
d. 同じ植物から2つのサンプルを採取し、一方を日光に当て、もう一方を日光から遠ざけておく

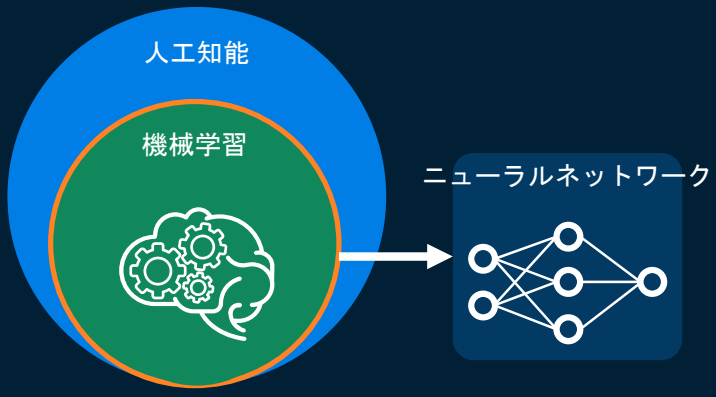
ビッグデータが生み出したスマートアプリケーション



ビッグデータが生み出したスマートアプリケーション



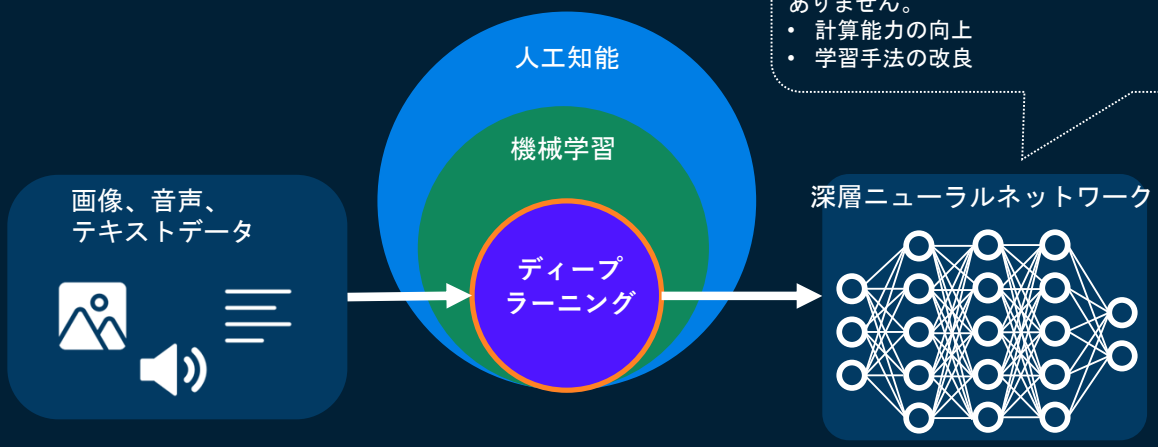
ビッグデータが生み出したスマートアプリケーション



機械学習は、分析モデルの
構築を自動化します。

ビッグデータが生み出したスマートアプリケーション

すべての多層ニューラルネットワーク
がディープラーニングであるわけでは
ありません。
- 計算能力の向上
- 学習手法の改良



ビッグデータが生み出したスマートアプリケーション

AIの一般的な用途

コンピュータビジョン

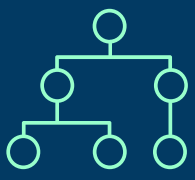
IoT

自然言語処理

ビッグデータが生み出したスマートアプリケーション

AIを実現・支援する技術

グラフィックス処理ユニット（GPU）



高度な
アルゴリズム

アプリケーション・プログラミング・インターフェース（API）

質問：コンピュータビジョン

次のうち、コンピュータビジョンの例となるものはどれですか？

a. スマートフォンの顔認証によるセキュリティアクセス
b. スマートフォンと連携したスマートウォッチ
c. スマートフォンのSiriやGoogleアシスタント
d. スマートフォンの自動修正機能

回答：コンピュータビジョン

次のうち、コンピュータビジョンの例となるものはどれですか？

a. スマートフォンの顔認証によるセキュリティアクセス
b. スマートフォンと連携したスマートウォッチ
c. スマートフォンのSiriやGoogleアシスタント
d. スマートフォンの自動修正機能

統計学の知識（1）

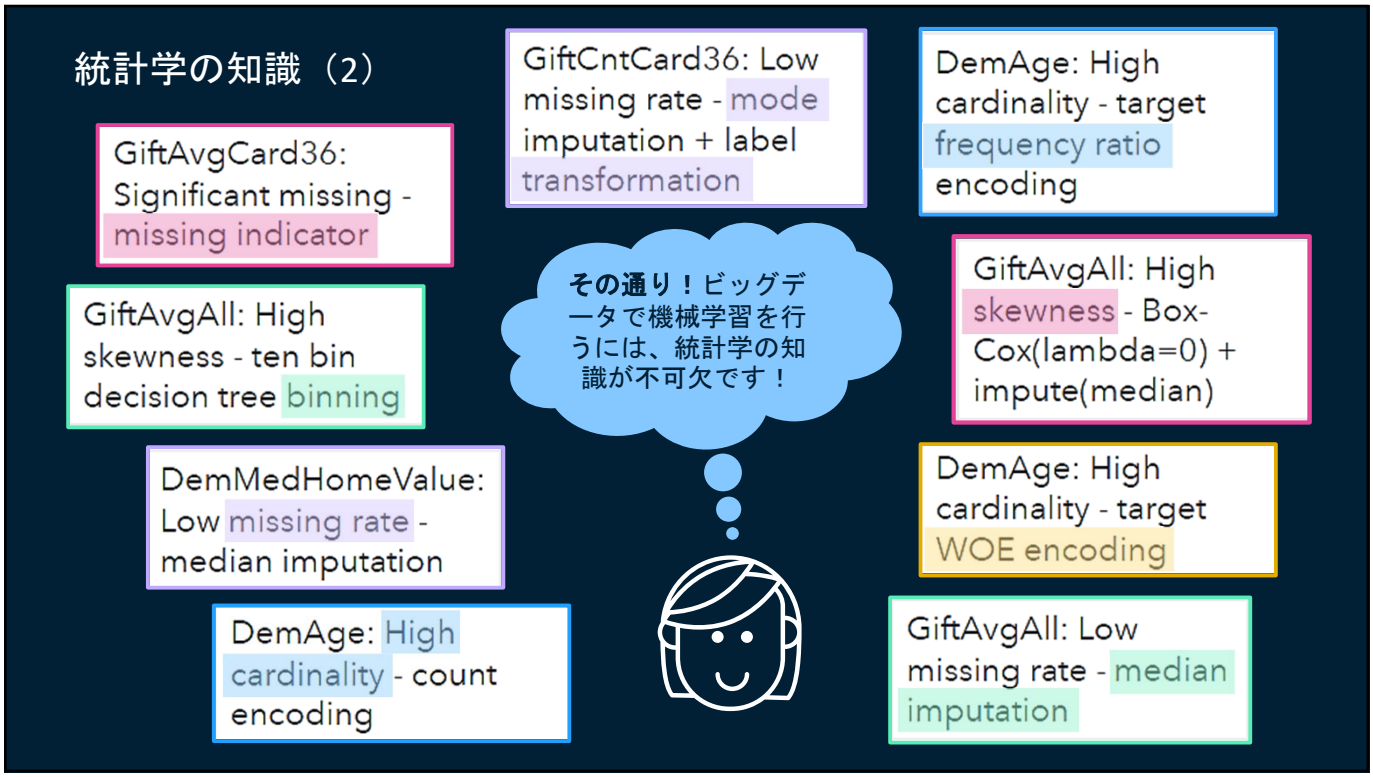
どう思いますか……ビッグデータで機械学習を行うには、統計学の知識が必要だと思いますか？

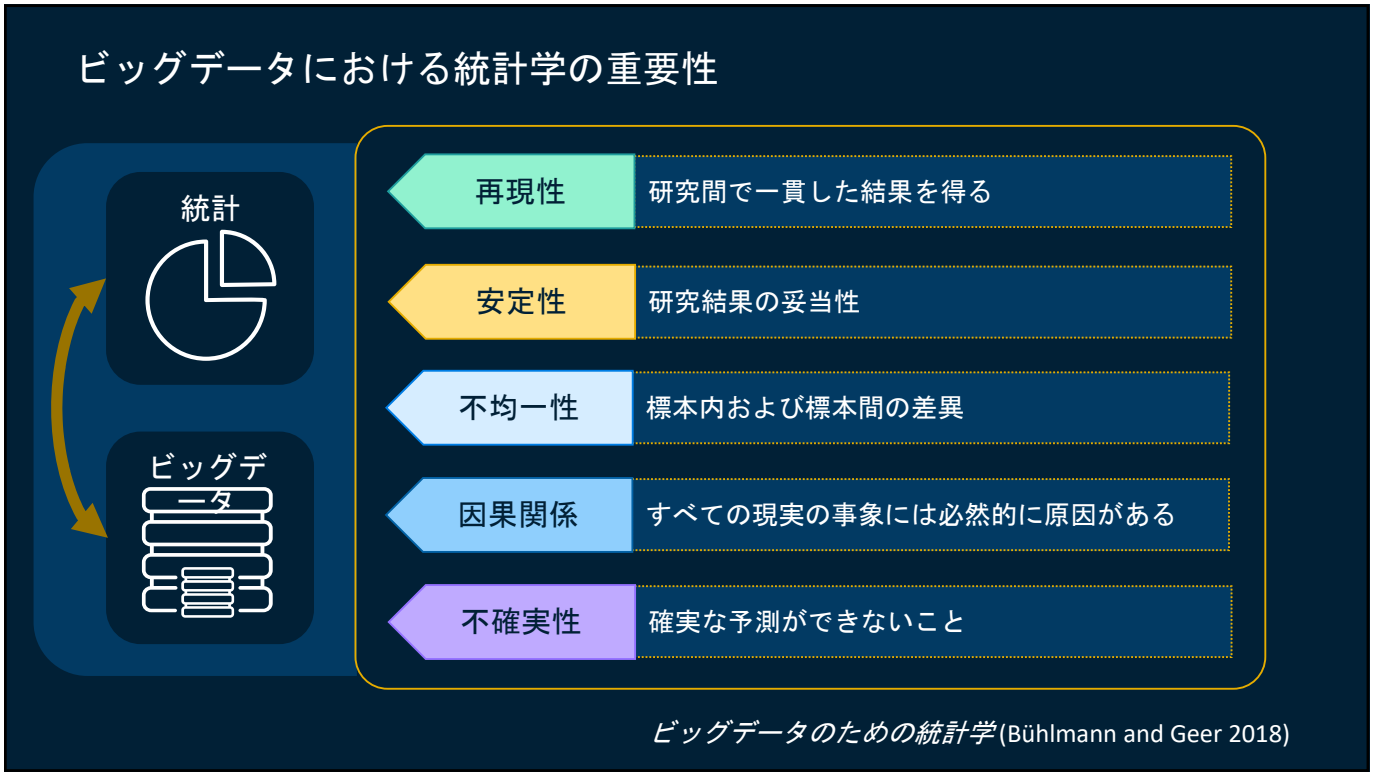
- ❑ いいえ、全く必要ありません。
- ❑ もしかすると、よく分からない。
- ❑ はい、間違いなく。

自動化された機械学習フローの実行について探ってみましょう。

デモ： SAS Studio自動機械学習

このデモでは、SAS Viya内のSAS Studioインターフェースへのアクセスと操作、コースデータのメモリへの読み込み、および自動機械学習フローの実行について解説します。


統計学の知識（2）
GiftCntCard36: Low missing rate - mode imputation + label transformation
DemAge: High cardinality - target frequency ratio encoding
GiftAvgCard36: Significant missing - missing indicator
その通り！ビッグデータで機械学習を行うには、統計学の知識が不可欠です！
GiftAvgAll: High skewness - Box-Cox(lambda=0) + impute(median)
GiftAvgAll: High skewness - ten bin decision tree binning
DemMedHomeValue: Low missing rate - median imputation
DemAge: High cardinality - target WOE encoding
DemAge: High cardinality - count encoding
GiftAvgAll: Low missing rate - median imputation


ビッグデータにおける統計学の重要性
統計
ビッグデータ
再現性
研究間で一貫した結果を得る
安定性
研究結果の妥当性
不均一性
標本内および標本間の差異
因果関係
すべての現実の事象には必然的に原因がある
不確実性
確実な予測ができないこと
ビッグデータのための統計学 (Bühlmann and Geer 2018)

ビッグデータにおける統計の重要性について

機械学習の世界への移行

"データから結論を導き出すための統計モデリングの活用には、2つのアプローチがある。1つは、データが特定の確率的データモデルによって生成されたと仮定するものである。もう1つは、アルゴリズムモデルを用い、データの生成メカニズムを未知のものとして扱うものである。"

– レオ・ブライマン (2001)

データの分析

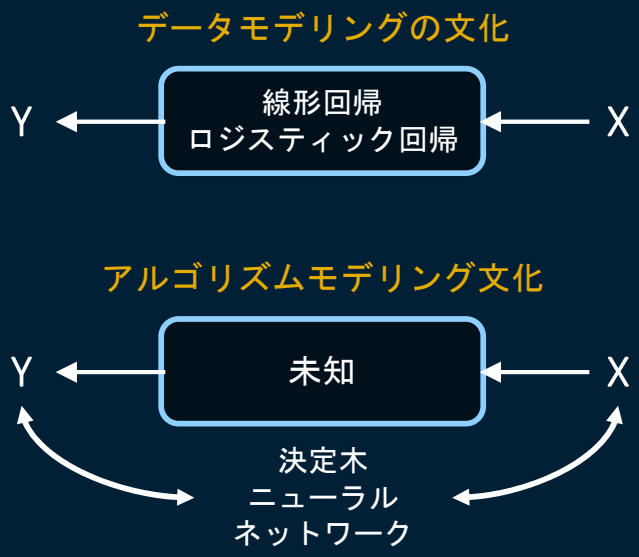
説明

自然は応答変数と入力変数はどのように関連付けられているのでしょうか？

予測

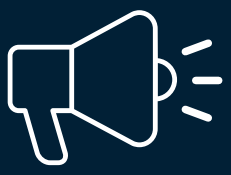
将来の入力変数に対する応答はどのようなものになるでしょうか？

統計モデリング：二つの文化



統計モデリング：二つの文化

マーケティング
キャンペーンの例



データモデリングの文化

反応行動の心理とは何か？
あるいは、特定の特性を持つ人がなぜ
反応しないのか？

アルゴリズム的モデリングの文化

ある人がマーケティングキャンペーン
に反応するかどうか？

質問：モデリングの文化

機械学習の世界では、次の2つのモデリング文化のうち、どちらがより頻繁に使われていると思いますか？

a. データモデリング文化
b. アルゴリズム的モデリング文化

回答：モデリング文化

機械学習の世界では、次の2つのモデリング文化のうち、どちらがより頻繁に使われていると思いますか？

a. データモデリング文化

b. アルゴリズムモデリング文化

モデリングの手法

従来のアプローチ

(SEMMA)

サンプル	Sample
探索	Explore
修正	Modify
モデル	Model
評価	Assess

標本データに対する完全な分析

モデリングの手法

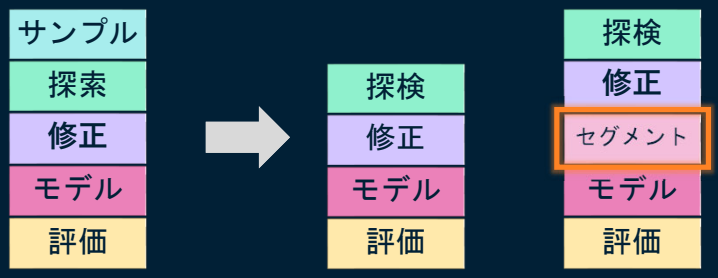
従来のアプローチ

(SEMMA)

ビッグデータへのアプローチ

(EMMA)

(EMSMA)



サンプリングされたデータに対する完全な分析

全データに対する完全な分析

全データに対する完全な分析、各セグメントでのモデリング

モデリングの手法

従来のアプローチ

(SEMMA)

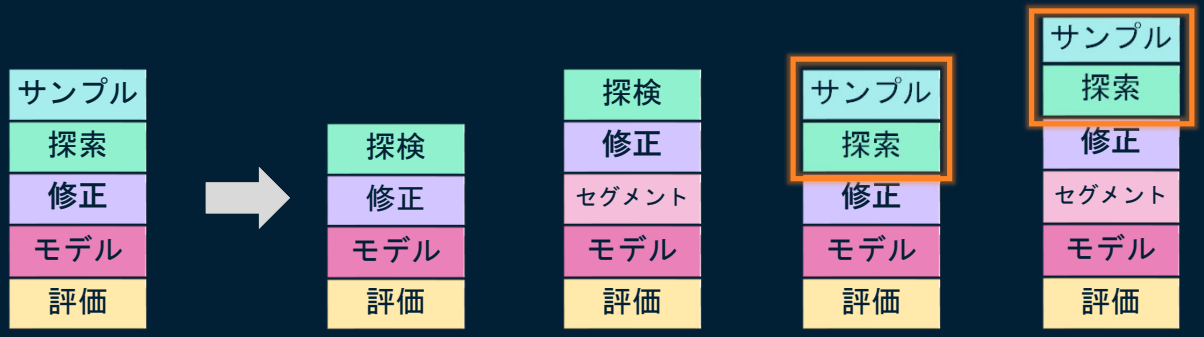
ビッグデータへのアプローチ

(EMMA)

(EMSMA)

(sEMMA)

(sEMSMA)



サンプリングデータに対する完全な分析

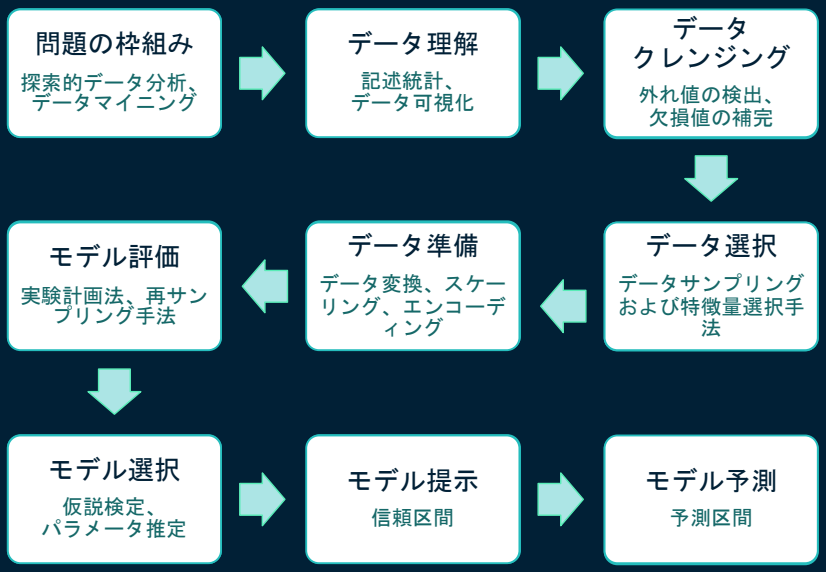
全データに対する完全な分析

全データに対する完全な分析、各セグメントでのモデリング

標本データでの探索、全データでの分析

サンプルデータでの探索、全データでの分析、各セグメントでのモデリング

機械学習において統計学はなぜ重要なのか？



統計学から機械学習への道のり

従来の統計学	機械学習
• 変数間の関係について推論を行うように設計されている。	• 可能な限り正確な予測を行うように設計されている。
• データモデリングに対する数学的な視点であり、データモデルと適合度に重点を置いている。	• データモデリングに対するコンピュータサイエンスの視点、**アルゴリズム的手法とモデルの性能に焦点を当てている。**
• 予測は最後に来る。	• 予測が最も重要である。
• 少数のサンプルや、データおよび分布に関する仮定に大きく依存している。	• 従来、計算能力に大きく依存してきた。

統計学から機械学習への道のりについてさらに詳しく

用語と語彙

Lesson 01, Section 02

基本的な用語

*これらの**新しい**、分かりにくい用語は何なのか？*

統計学者

*これらの**古くて**分かりにくい用語は一体何なのか？*

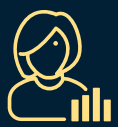
データサイエンティスト

基本的な用語

統計用語

機械学習の用語

*これらの**新しい**、分かりにくい用語は何なのか？*



統計学者

名前	年齢	性別	週の労働時間	収入
ジョン	24	男性	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

- データセット
- データテーブル

これらの古い、分かりにくい用語は何なの？

データサイエンティスト

基本的な用語

名前	年齢	性別	週の労働時間	収入
ジョン	24	男性	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

- 要素
- エンティティ

- ケース
- オブジェクト

基本的な用語

名前	年齢	性別	週の労働時間	収入
ジョン	24	男性	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

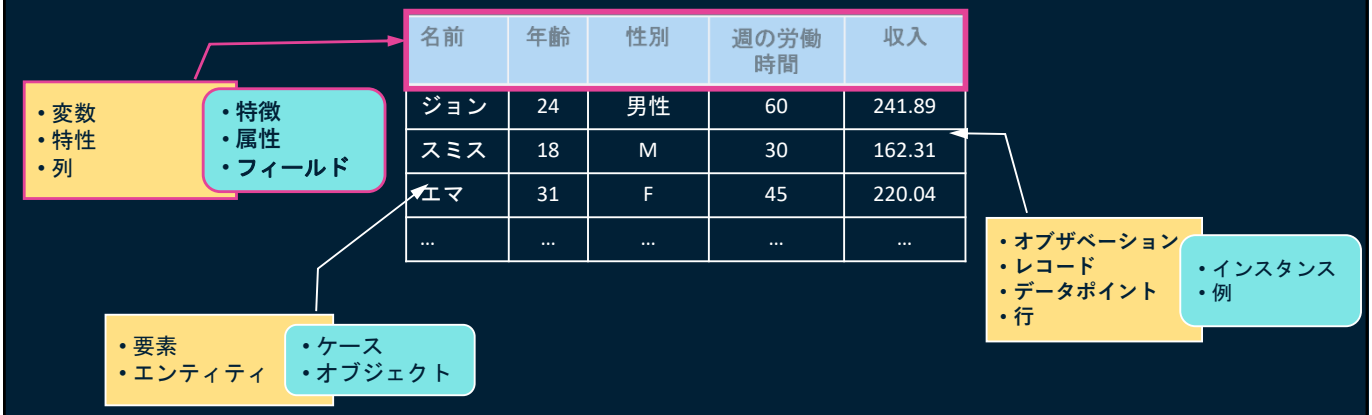
- 要素
- エンティティ

- ケース
- オブジェクト

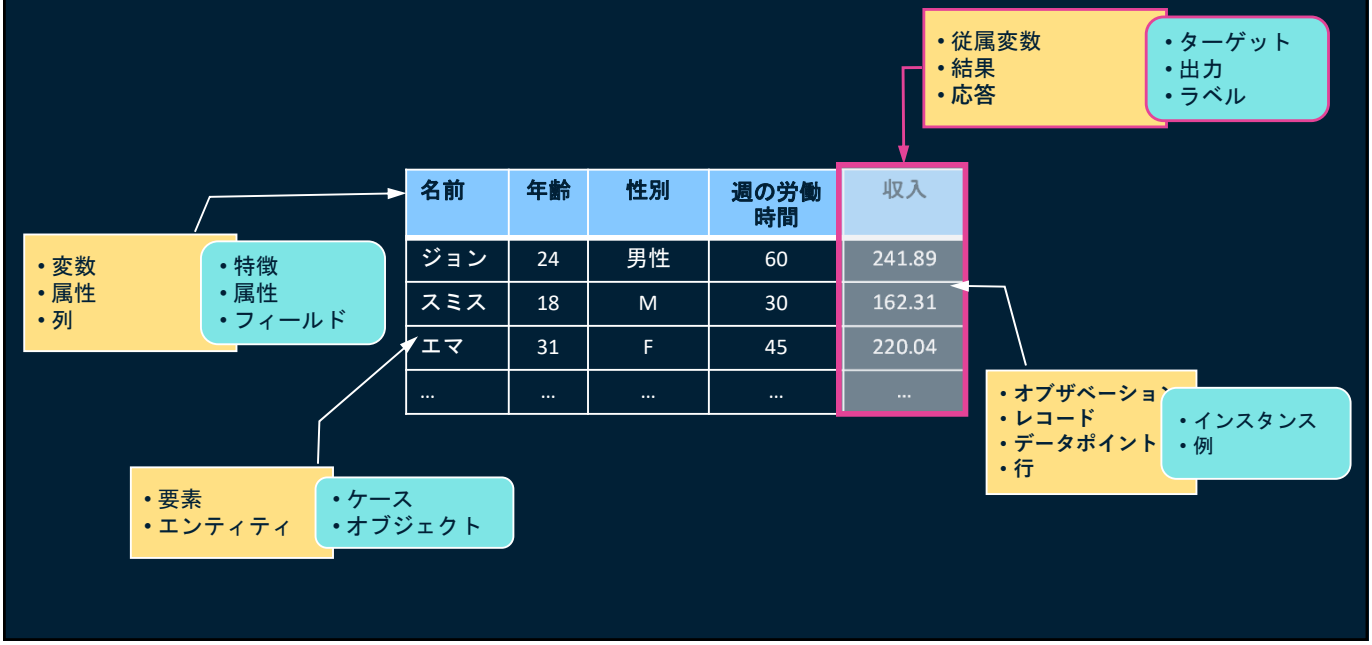
- オブザベーション
- レコード
- データポイント
- 行

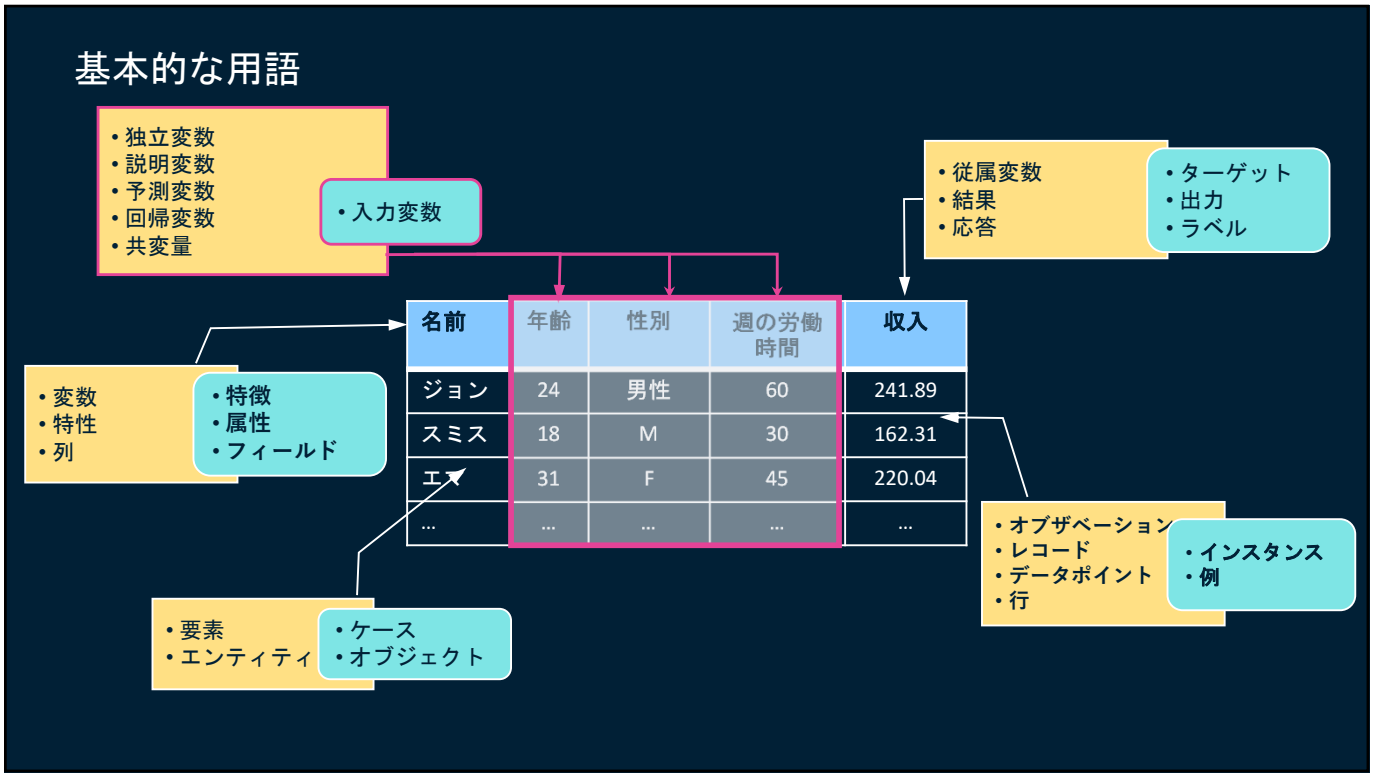
- インスタンス
- 例

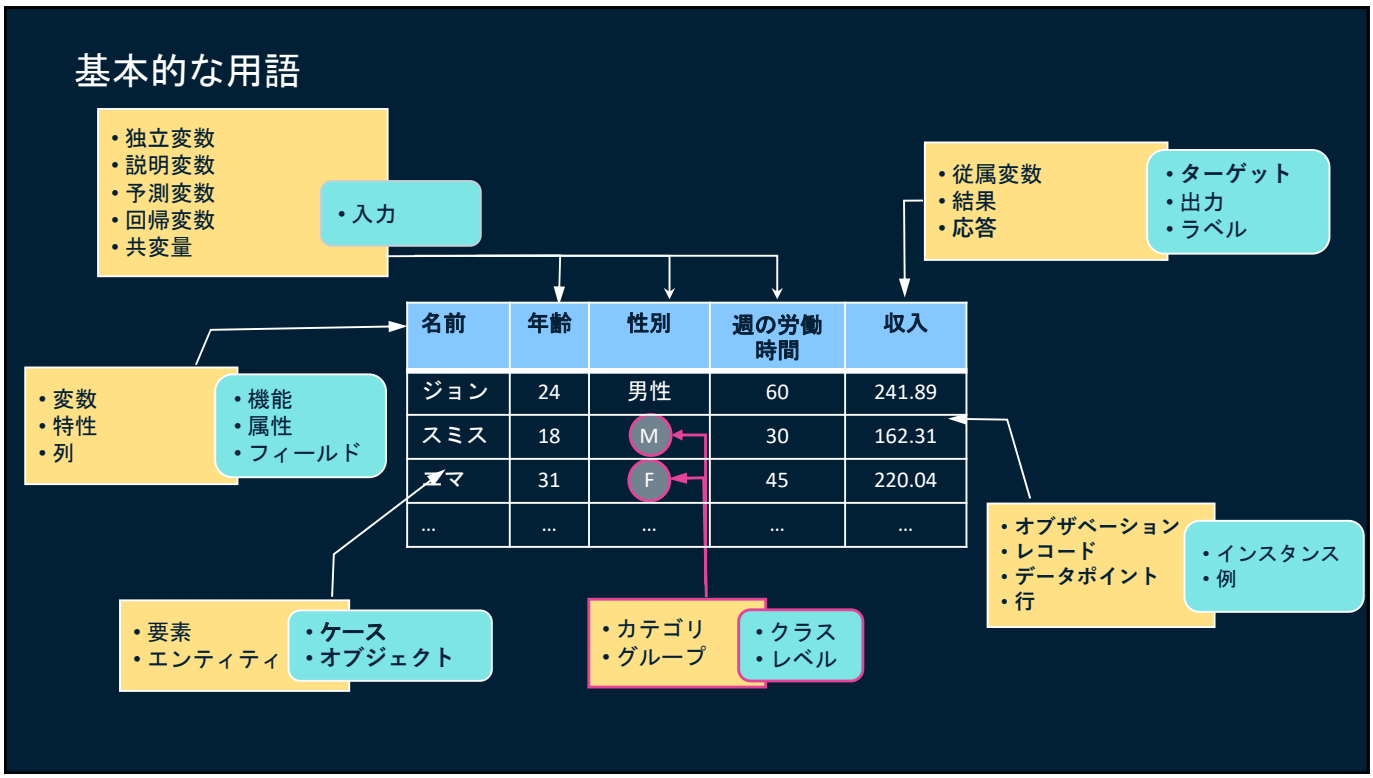
基本的な用語



基本的な用語




基本的な用語
• 独立変数
• 説明変数
• 予測変数
• 回帰変数
• 共変量
• 入力変数
• 従属変数
• 結果
• 応答
• ターゲット
• 出力
• ラベル
名前
年齢
性別
週の労働時間
収入
ジョン
24
男性
60
241.89
スミス
18
M
30
162.31
エマ
31
F
45
220.04
...
...
...
...
...
• 変数
• 特性
• 列
• 特徴
• 属性
• フィールド
• 要素
• エンティティ
• ケース
• オブジェクト
• オブザベーション
• レコード
• データポイント
• 行
• インスタンス
• 例


基本的な用語
• 独立変数
• 説明変数
• 予測変数
• 回帰変数
• 共変量
• 入力
• 従属変数
• 結果
• 応答
• ターゲット
• 出力
• ラベル
名前
年齢
性別
週の労働時間
収入
ジョン
24
男性
60
241.89
スミス
18
M
30
162.31
エマ
31
F
45
220.04
...
...
...
...
...
• 変数
• 特性
• 列
• 機能
• 属性
• フィールド
• 要素
• エンティティ
• ケース
• オブジェクト
• カテゴリ
• グループ
• クラス
• レベル
• オブザベーション
• レコード
• データポイント
• 行
• インスタンス
• 例

問題：基本用語（1）

次の用語の組み合わせのうち、データ表の強調表示された部分を定義しているのはどれか。（該当するものすべてを選択してください。）

a. ターゲット / 応答
b. インスタンス／例
c. 予測変数 / 入力
d. オブザベーション / レコード

会社	証券取引所	年間売上高（百万ドル）	1株当たり利益（$）
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
…	…	…	…

回答：基本用語 (1)

次の用語の組み合わせのうち、データ表の強調表示された部分を定義しているのはどれか。（該当するものをすべて選択してください。）

a. ターゲット / 応答
b. インスタンス／例
c. 予測変数 / 入力
d. オブザベーション / レコード

会社	証券取引所	年間売上高（百万ドル）	1株当たり利益（$）
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
…	…	…	…

問題：基本用語（2）

データ表の強調表示された部分を定義しているのは、以下の用語の組み合わせのうちどれですか？（該当するものをすべて選択してください。）

a. 変数／特性
b. カテゴリー／グループ
c. クラス／レベル
d. 特徴量／属性

会社	証券取引所	年間売上高（百万ドル）	1株当たり利益（$）
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
…	…	…	…

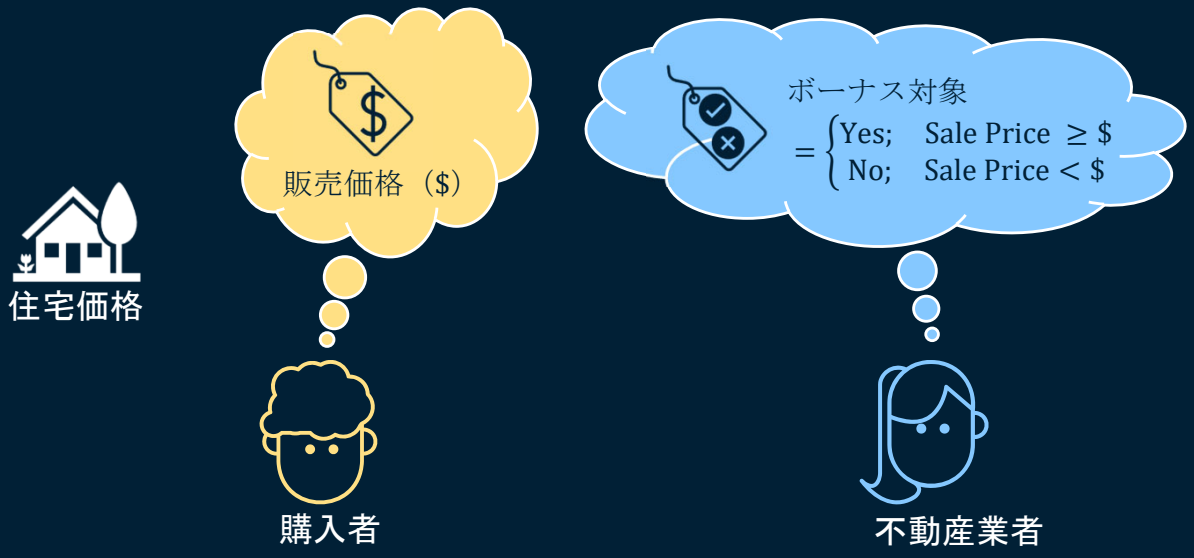
回答：基本用語 (2)

データ表の強調表示された部分を定義しているのは、次の用語の組み合わせのうちどれですか？（該当するものをすべて選択してください。）

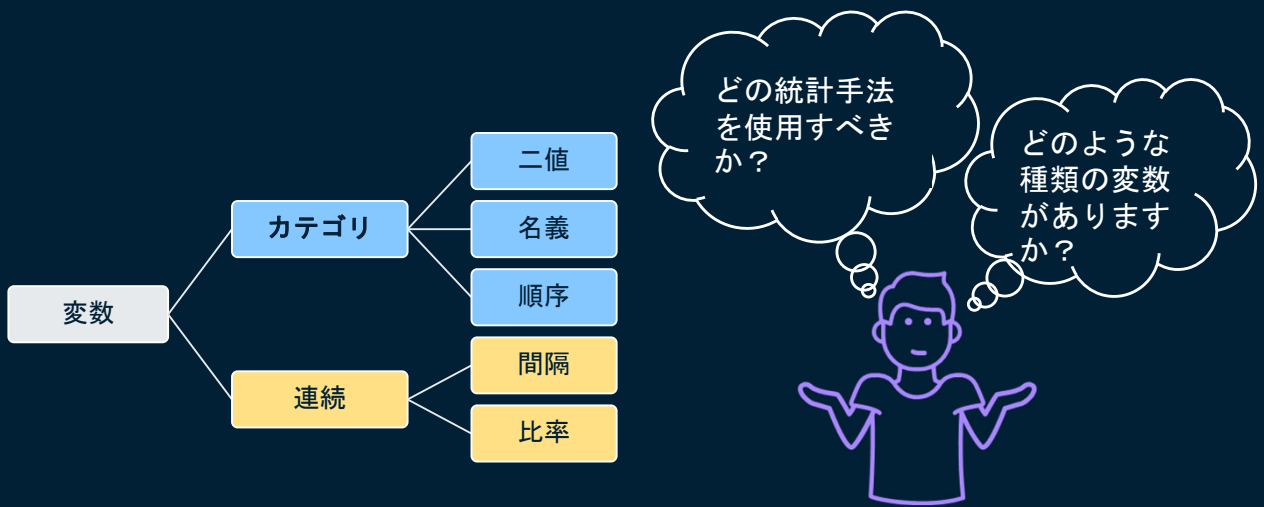
a. 変数／特性
b. カテゴリー／グループ
c. クラス／レベル
d. 特徴量／属性

会社	証券取引所	年間売上高（百万ドル）	1株当たり利益（$）
MedTech	AMEX	84.20	0.33
EnergyWest	NYSE	372.85	1.89
DataDig	BSE	74.65	2.10
…	…	…	…

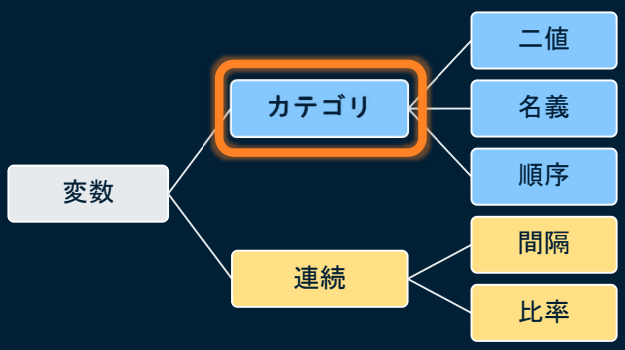
変数の種類と測定水準



変数の種類と測定水準



変数の種類と測定水準



カテゴリ

ステータス: オン / オフ

ステータス: リクエスト済 / 進行中 / 完了

変数の種類と測定水準

連続

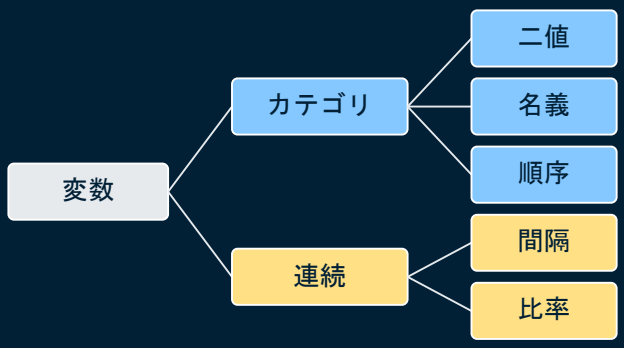
年齢:

22 22.45 22.6 22.98 23

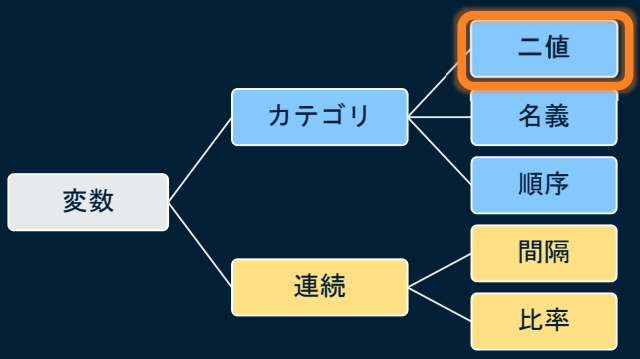
連続データ用の手法を適用できる

一部の数値変数、例えばカウントなど、は連続的**ではありません。**

変数の種類と測定水準

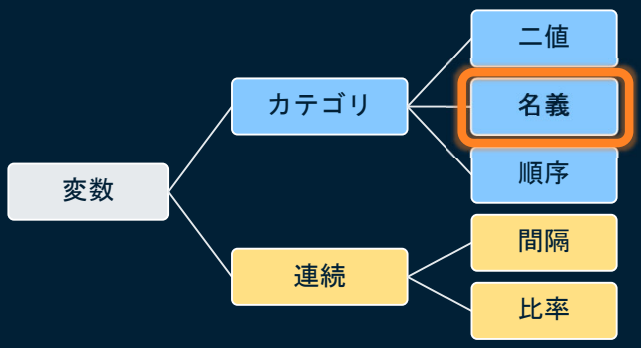


変数の種類と測定水準



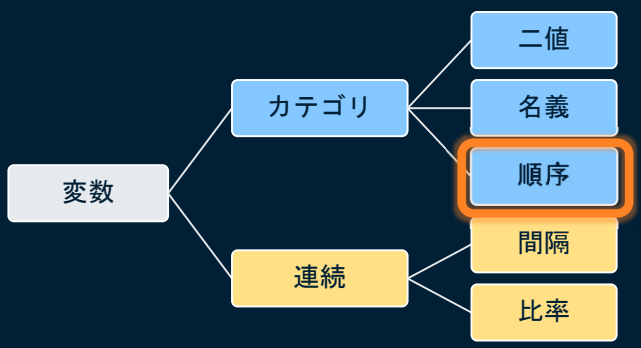
変数：飲料ありますか？
1 又は0、はい又はいいえ

変数の種類と測定水準



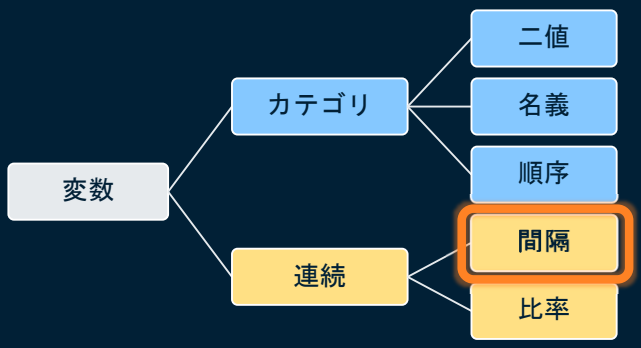
変数：飲料の種類

変数の種類と測定水準

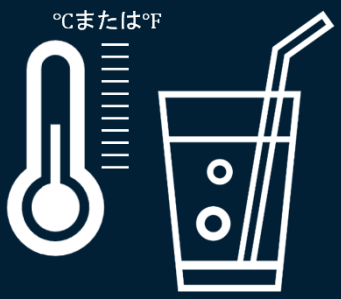


変数：飲料のサイズ

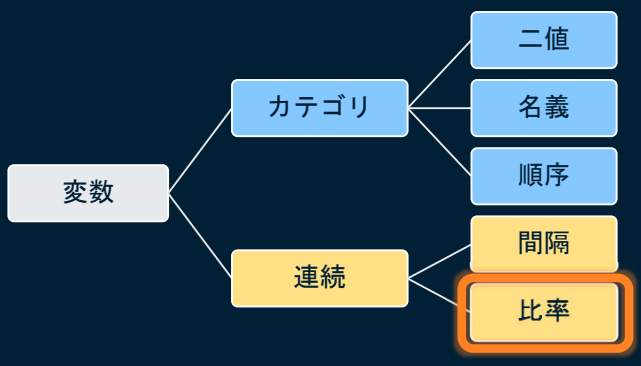
変数の種類と測定水準



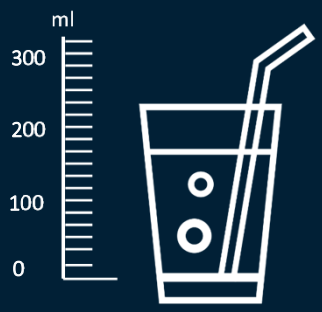
変数：飲料の温度



変数の種類と測定水準



変数：飲料の量



SASは比率尺度変数と間隔尺度変数を区別しません。

問題：測定尺度

次のうち、名義尺度を表すものはどれか。

a. ランナーに割り当てられた背番号			
b. 優勝者の順位			
c. 0〜10スケールのパフォーマンス評価	8.2	9.1	9.6
d. ゴールまでの時間（秒単位）	15.2	14.1	13.4

回答：測定尺度

次のうち、名義尺度を表すものはどれか。

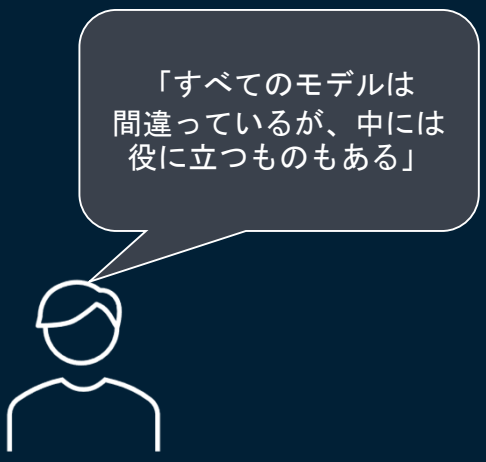
a. ランナーに割り当てられた背番号			
b. 優勝者の順位			
c. 0〜10スケールのパフォーマンス評価	8.2	9.1	9.6
d. ゴールまでの時間（秒）	15.2	14.1	13.4

モデリング用語集

分析データ／学習データ

ID　　*入力*　　*ターゲット*

名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…



すべてのモデルは間違っている（Box 1976）

モデリング用語集

モデル

モデルとは、入力とターゲットの関連性を簡潔に表現したもの

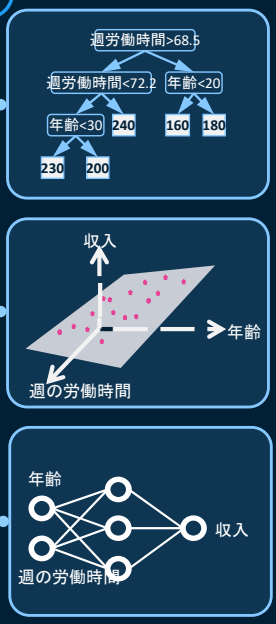
名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

モデル

モデリング用語集

モデルフレームワーク

名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…



決定木モデル

回帰モデル

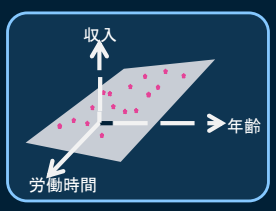
ニューラルネットワークモデル

モデリング用語集

回帰モデルの例

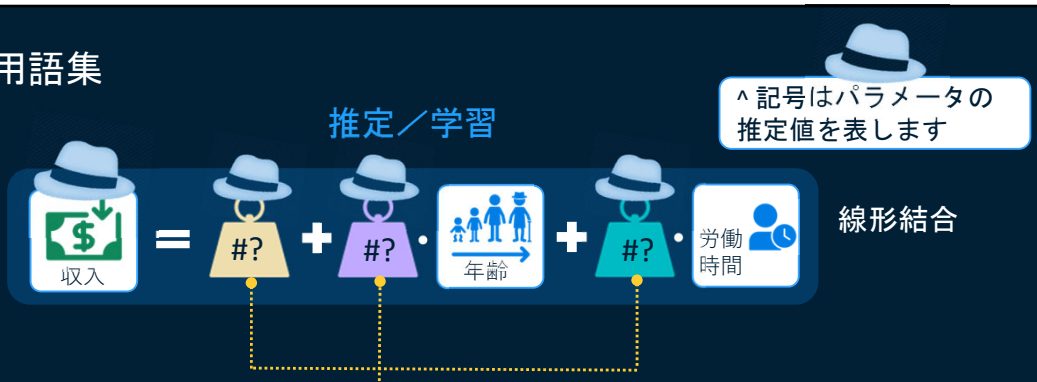
線形結合

名前	年齢	性別	労働時間	収入
ジョン	24	男性	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…



回帰モデル

モデリング用語集



推定基準を最適化するために、
パラメータの推定値を選択する。

- 従来の統計学の用語では、これを「*推定*」と呼ぶ。
- 現代の機械学習の世界では、これは「*学習*」と呼ぶ。

モデリング用語集

名前	年齢	性別	労働時間	収入
ジョン	24	男性	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

$\widehat{\text{収入}}$
274.19
143.63
215.51
…

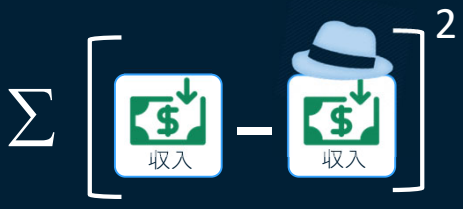
残差／誤差：

モデリング用語集

推定基準／学習指標

推定基準を最適化するために、
パラメータ推定値を選択します。

二乗誤差関数を最小化する：



モデリング用語

モデル適合 / モデル学習

モデリング用語集

NEW

名前	年齢	労働時間
ロバート	45	50
メアリー	21	40
ジェームズ	34	22
…	…	…

予測収入
245.85
187.81
120.43
…

*予測*とは、一連の入力測定値に対してモデルが出力する結果のこと。

質問：モデリング用語（モデル）

次のうち、分析用語としての「*モデル*」を定義しているのはどれか。

a. 推定基準を最適化することでモデルパラメータの推定値を選択すること
b. アルゴリズムがデータセットをどの程度適切にモデル化しているかを評価する尺度
c. 入力とターゲットの関連性を簡潔に表現したもの
d. 商業製品の宣伝、展示、または広告を行う役割を担う人

回答：モデリング用語（モデル）

次のうち、分析用語「*モデル*」を定義しているのはどれか。

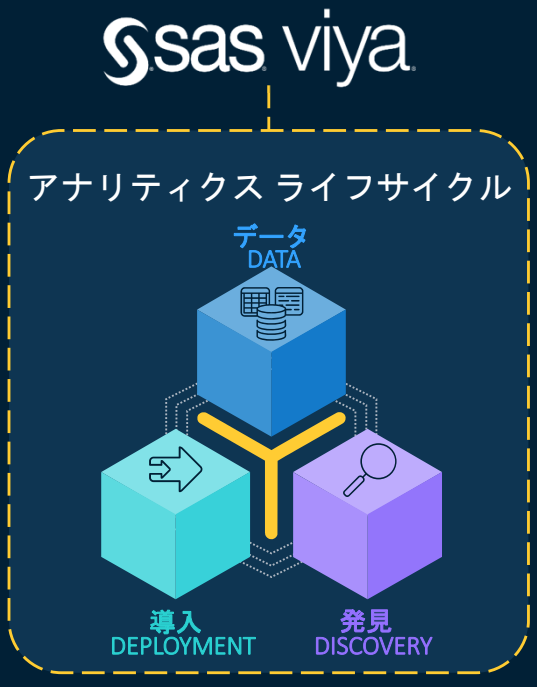
a. 推定基準を最適化することでモデルパラメータの推定値を選択すること
b. アルゴリズムがデータセットをどの程度適切にモデル化しているかを評価する尺度
c. 入力とターゲット目標の関連性を簡潔に表現したもの
d. 商業製品の宣伝、展示、または広告を行う役割を担う人

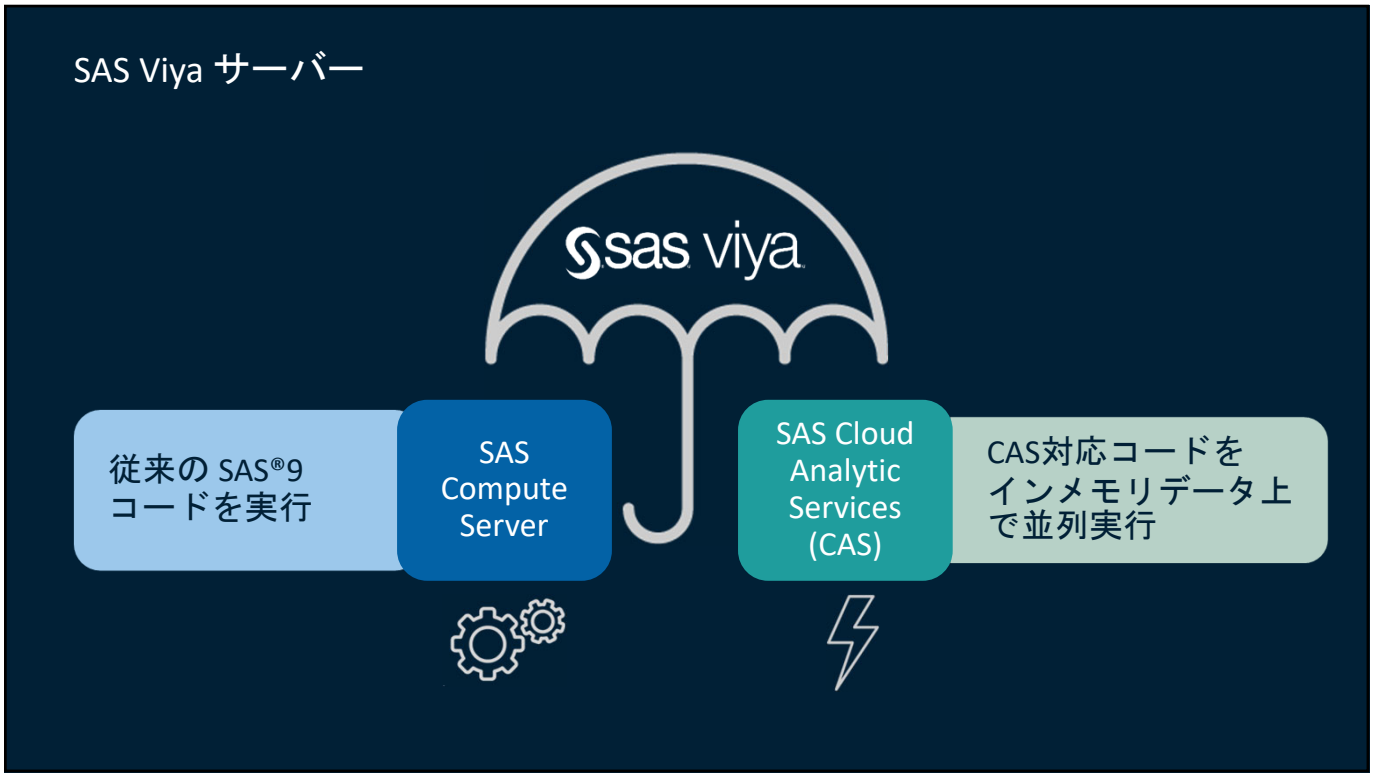
SAS Viya および SAS Studio の概要

Lesson 01, Section 03

SAS Viyaの概要

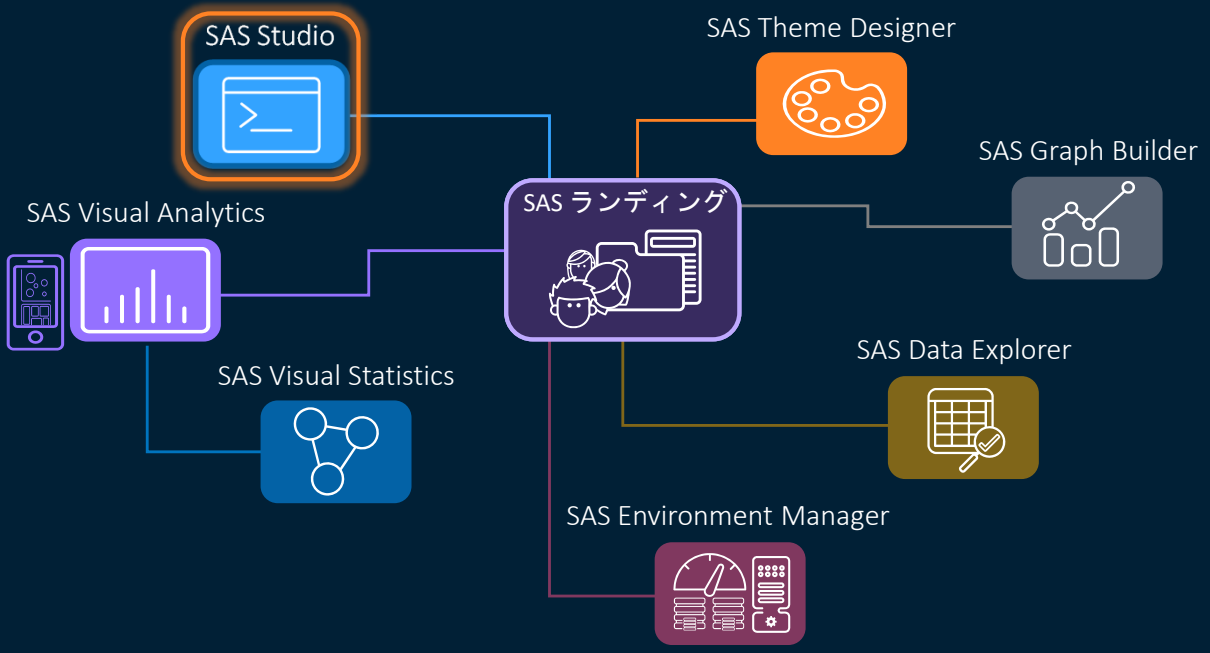
SAS Viyaの概要




SAS Viya サーバー
sas viya
従来の SAS®9 コードを実行
SAS Compute Server
SAS Cloud Analytic Services (CAS)
CAS対応コードをインメモリデータ上で並列実行

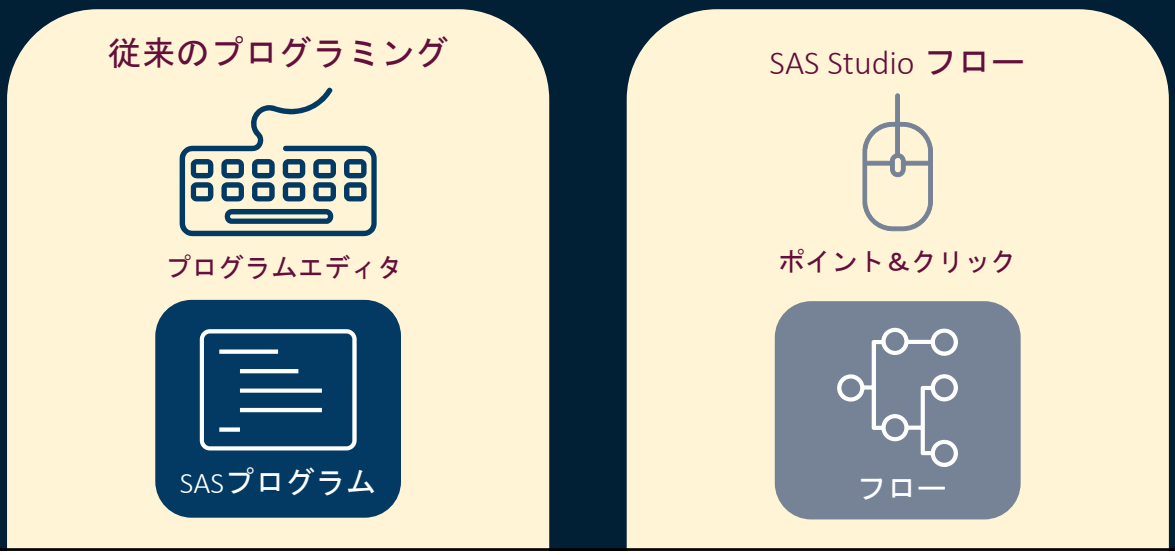
SAS Viyaの詳細

SAS Viya Applications

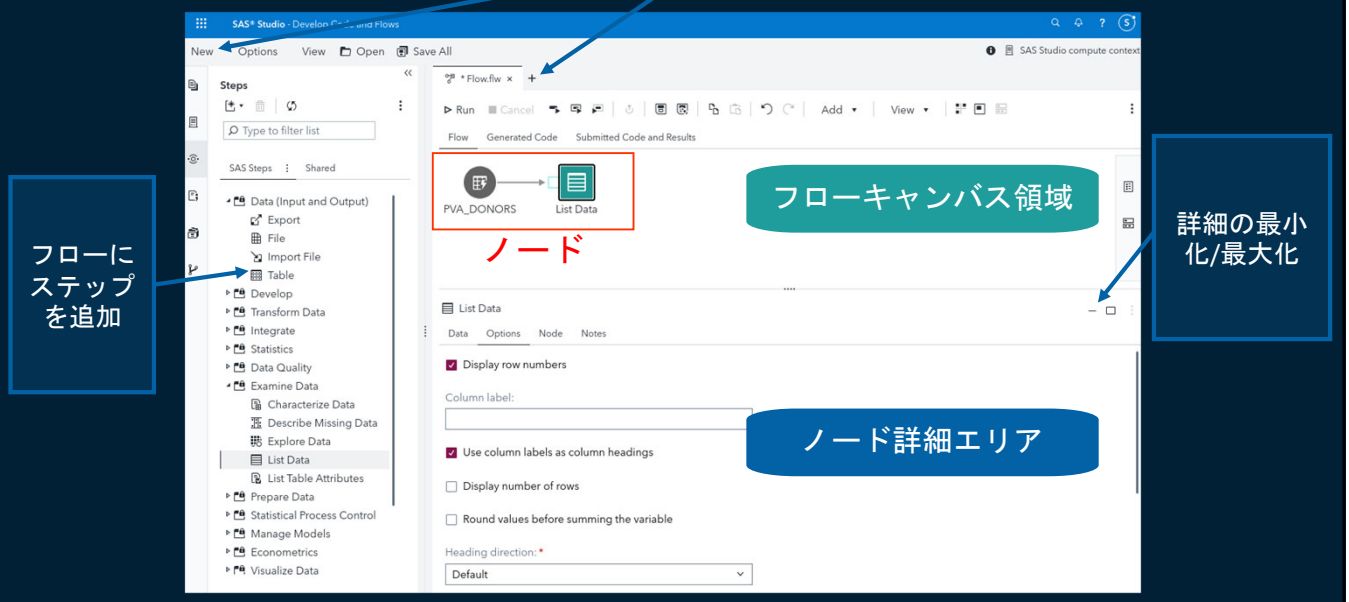


SAS Studio

- SAS Studioは、SASコードを扱うための複数の方法を提供するWebベースの開発インターフェースです。



SAS Studioのフロー例



SAS Studioのフローステップ

SASステップのカテゴリ		
データ（入力および出力）	エンリッチメント	データ準備
開発（コーディング）	計量経済学	データ検証
データ変換	データ品質	データ分析
統合	データ可視化	テキスト分析
データ統合	機械学習	カスタムステップ
モデル管理	最適化とネットワーク分析	

質問：SAS Viya

SAS Viyaは、並列タスクの実行を可能にするクラウド対応の分析およびデータ管理プラットフォームです。

a. 正しい
b. 誤り

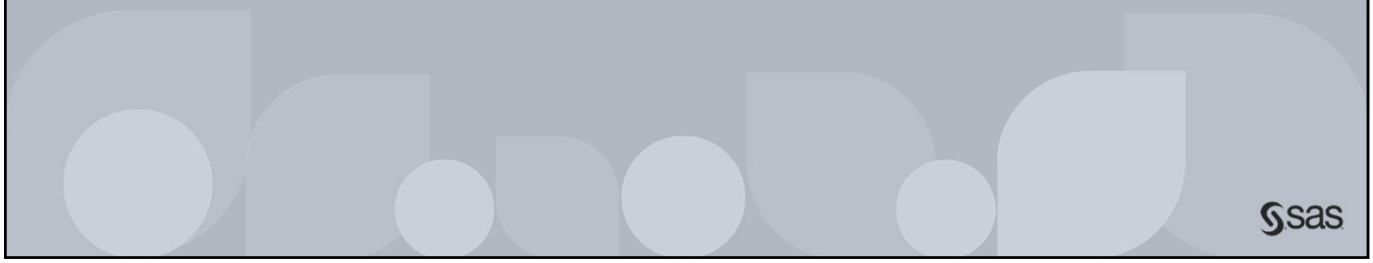
回答：SAS Viya

SAS Viyaは、並列タスクの実行を可能にするクラウド対応の分析およびデータ管理プラットフォームです。

a. 正しい
b. 誤り

ご質問はありますか？

レッスンのまとめ



Lesson 2: 統計学の基本概念

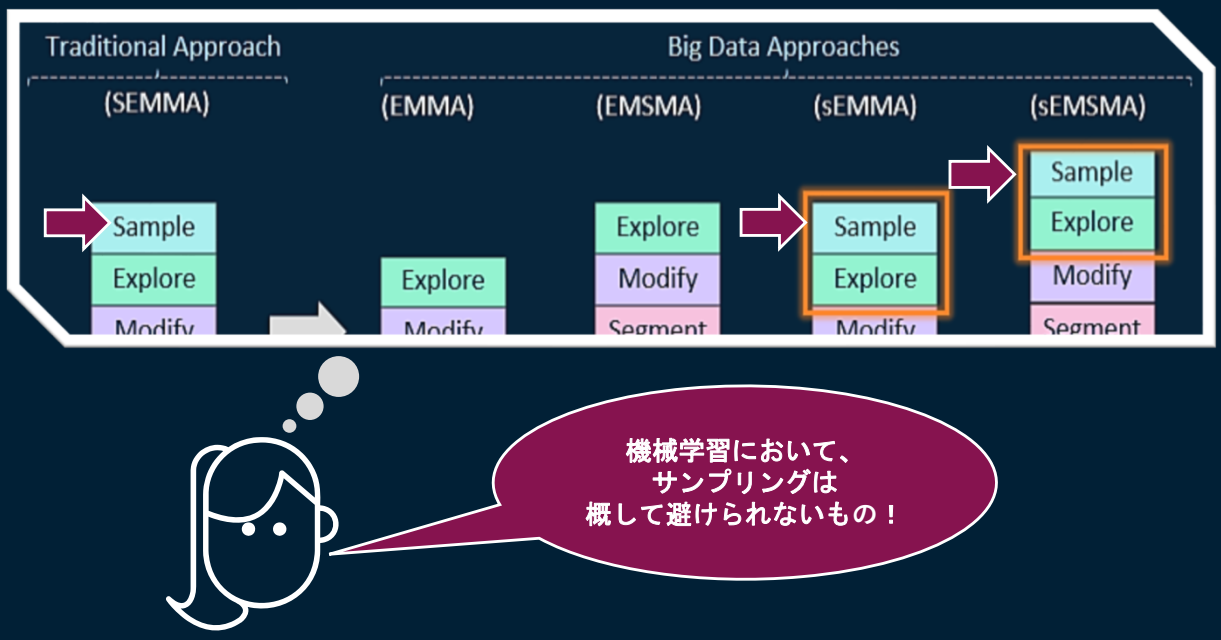
sas

統計分析入門

Lesson 02, Section 01

sas

母集団と標本

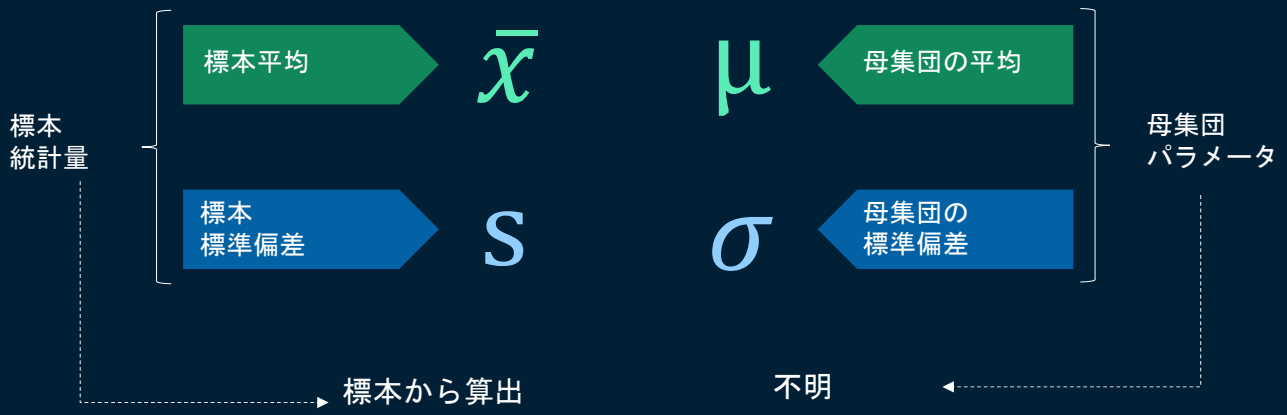


母集団と標本

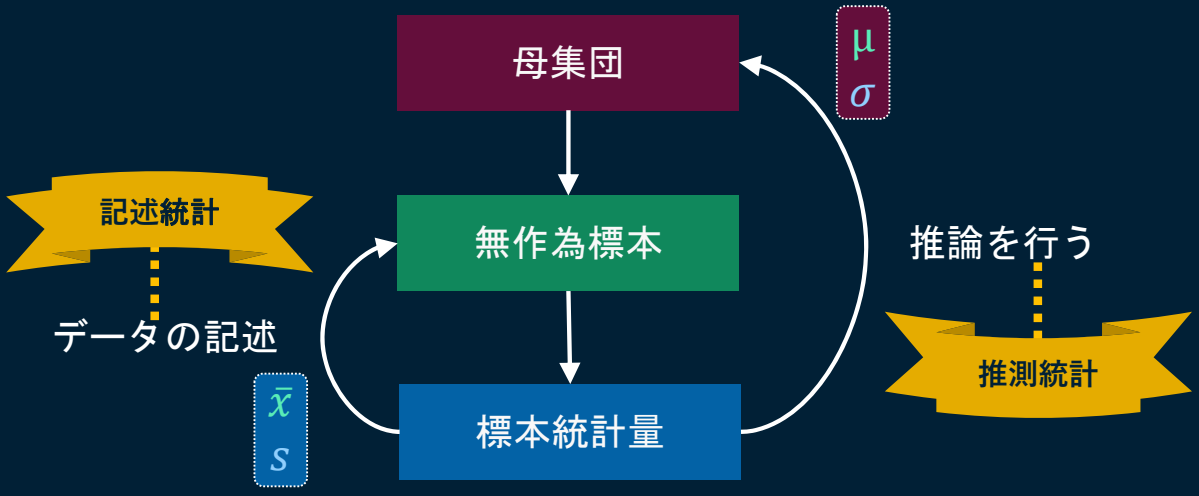
母集団 – 対象となる集団を構成する個体の全体

標本 –母集団への推論を可能にするために抽出された母集団の部分集合

パラメータと統計



統計分析の手順



問題：サンプル

標本とは、以下のうちどれから構成されるか。

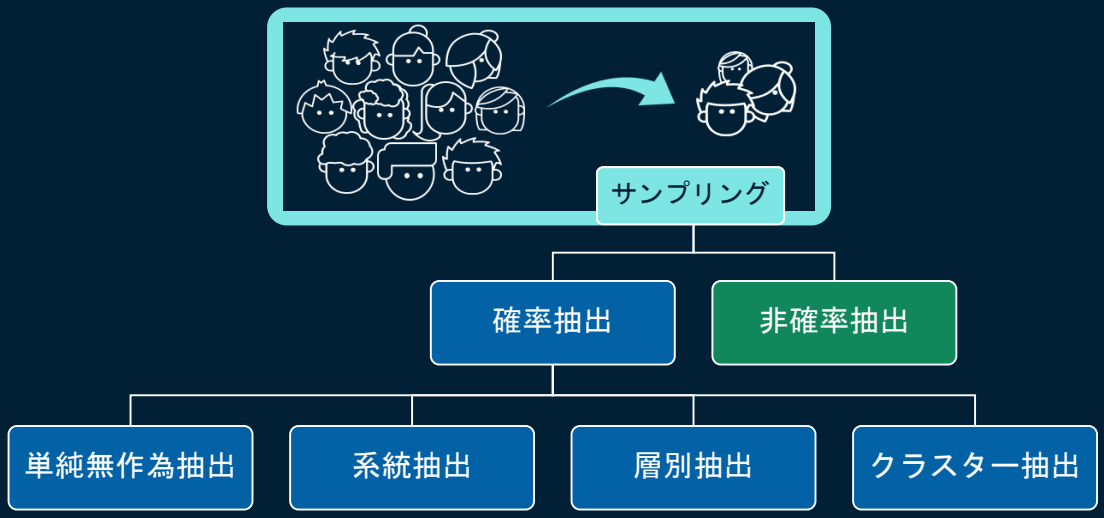
a. 母集団のすべて
b. 母集団の少なくとも20%
c. 母集団の最大50%
d. 母集団の任意の割合

回答：サンプル

標本とは、以下のうちどれから構成されるか。

a. 母集団のすべて
b. 母集団の少なくとも20%
c. 母集団の最大50%
d. 母集団の任意の割合

サンプリング



抽出法

18個から6個を抽出する

抽出法

18個から6個を抽出する

単純無作為抽出

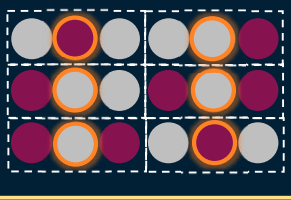
抽出法

18個から6個抽出する

単純無作為抽出

群のサイズ、k
= N/n
= 18/6 = 3

系統抽出

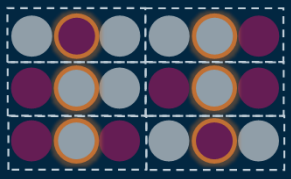


抽出法

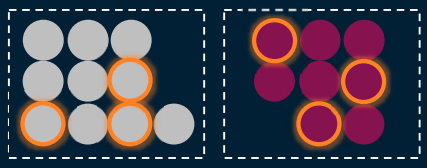
18個から6個を抽出する

単純無作為抽出

系統抽出



層別抽出

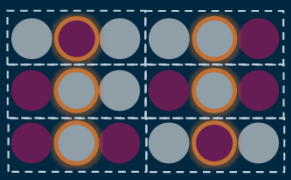


抽出法

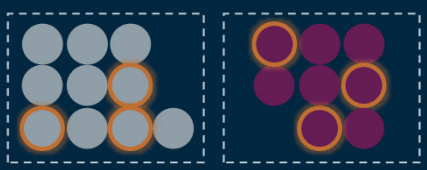
18個から6個を抽出する

単純無作為抽出

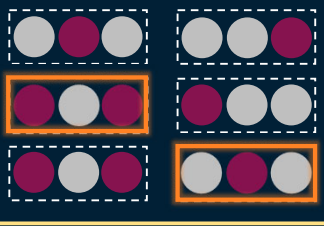
系統抽出



層別抽出



クラスター抽出



サンプリングについて詳しく

問題：サンプリング手法

FMラジオ市場において、リスナーの年齢は、放送局が採用する番組編成のタイプを決定する重要な要因です。リスナーの3つのサブグループ（25歳未満、25～40歳、40歳以上）それぞれから無作為抽出を行う場合、これはどのようなサンプリング手法でしょうか？

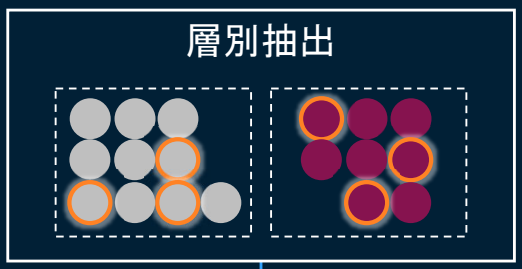
a. 単純無作為抽出
b. 層別抽出
c. 系統抽出
d. クラスター抽出

回答：サンプリング手法

FMラジオ市場において、リスナーの年齢は、放送局が採用する番組編成のタイプを決定する重要な要因です。リスナーの3つのサブグループ（25歳未満、25～40歳、40歳以上）それぞれから無作為抽出を行う場合、これはどのようなサンプリング手法でしょうか？

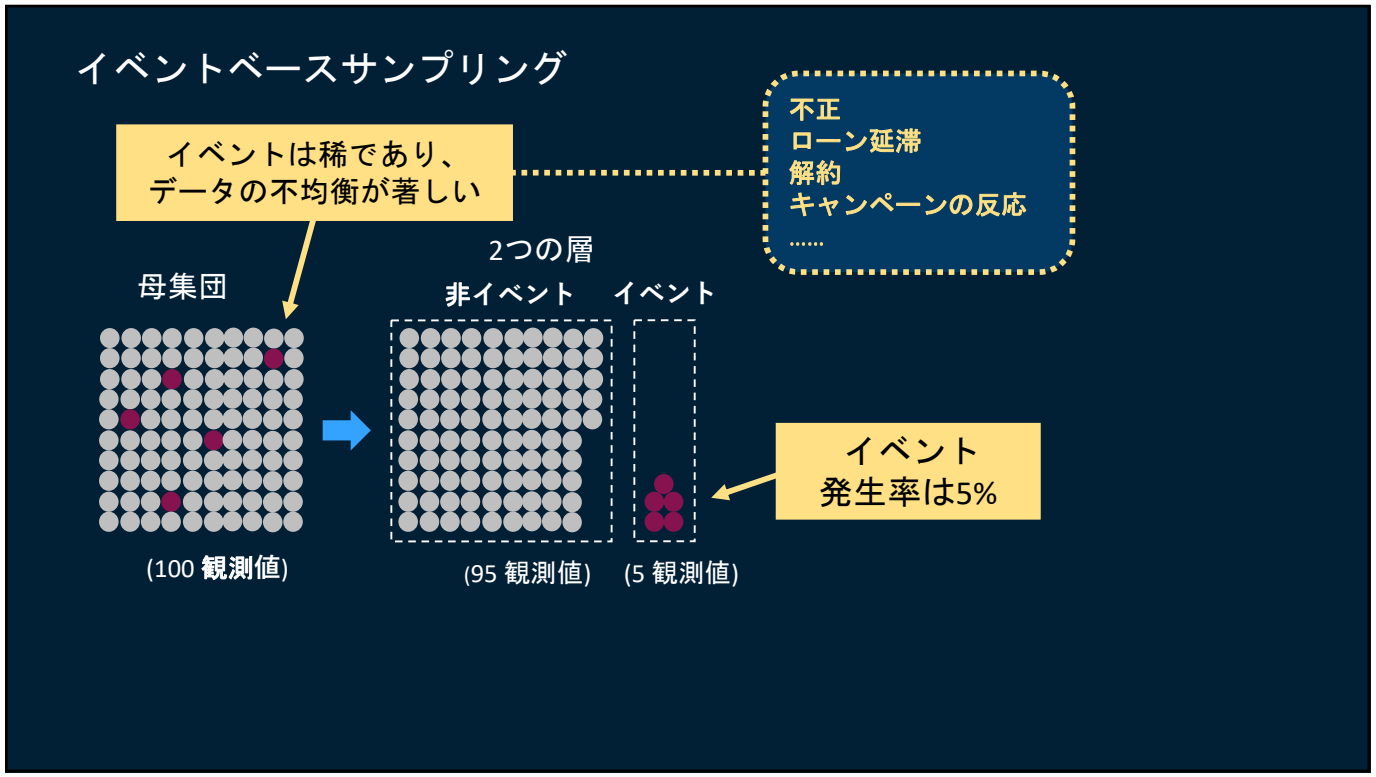
a. 単純無作為抽出
b. 層別抽出
c. 系統抽出
d. クラスター抽出

イベントベースサンプリング

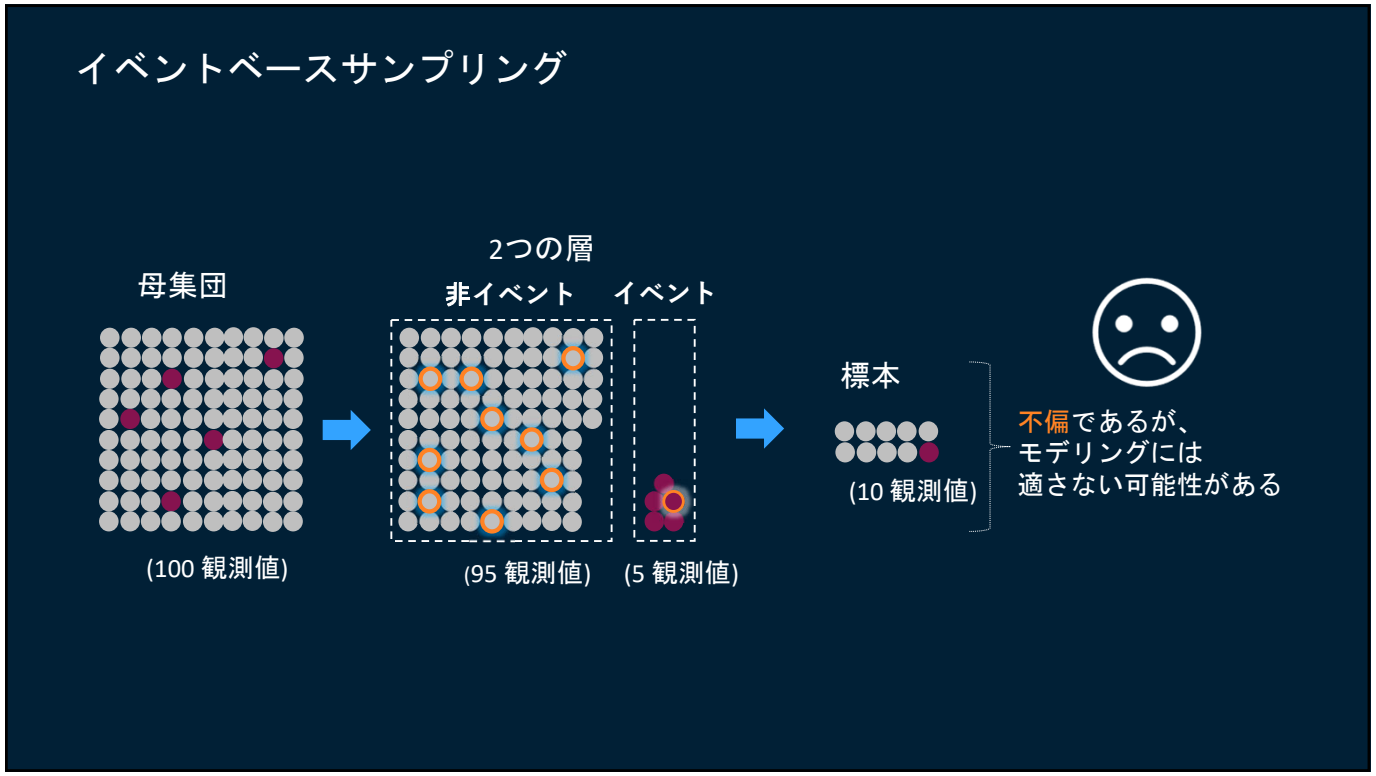


イベントベースサンプリング
（分離抽出）

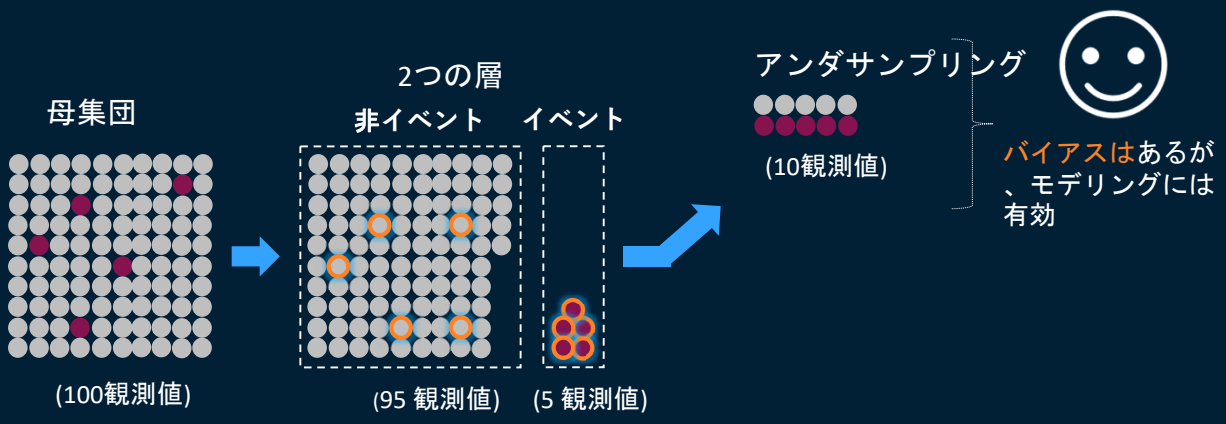
イベントベースサンプリング



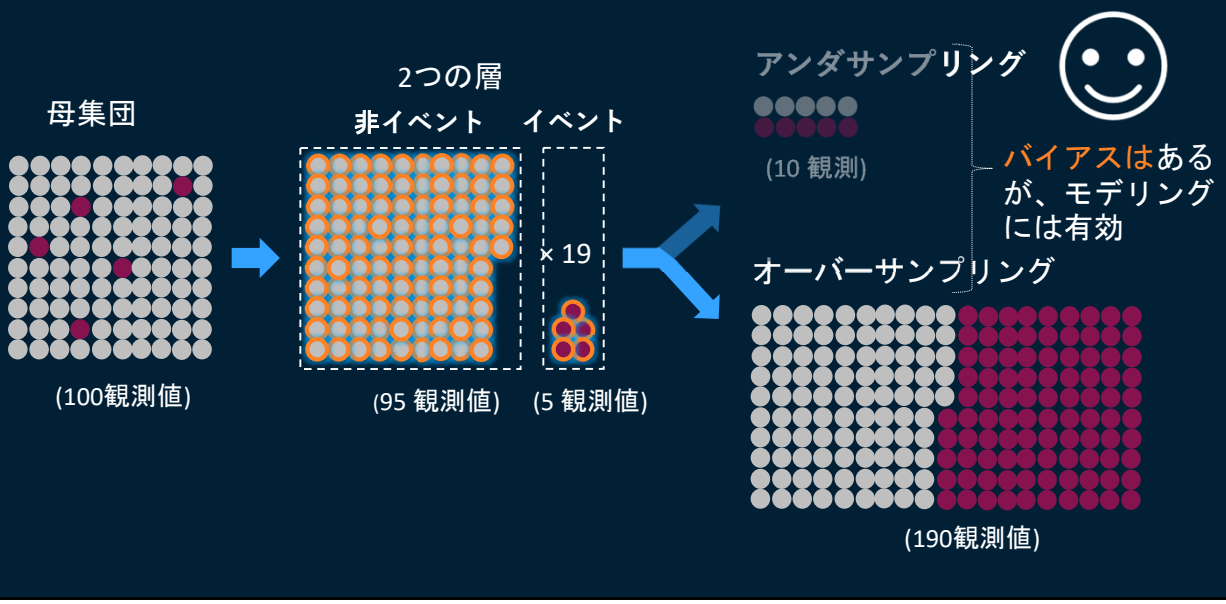
イベントベースサンプリング



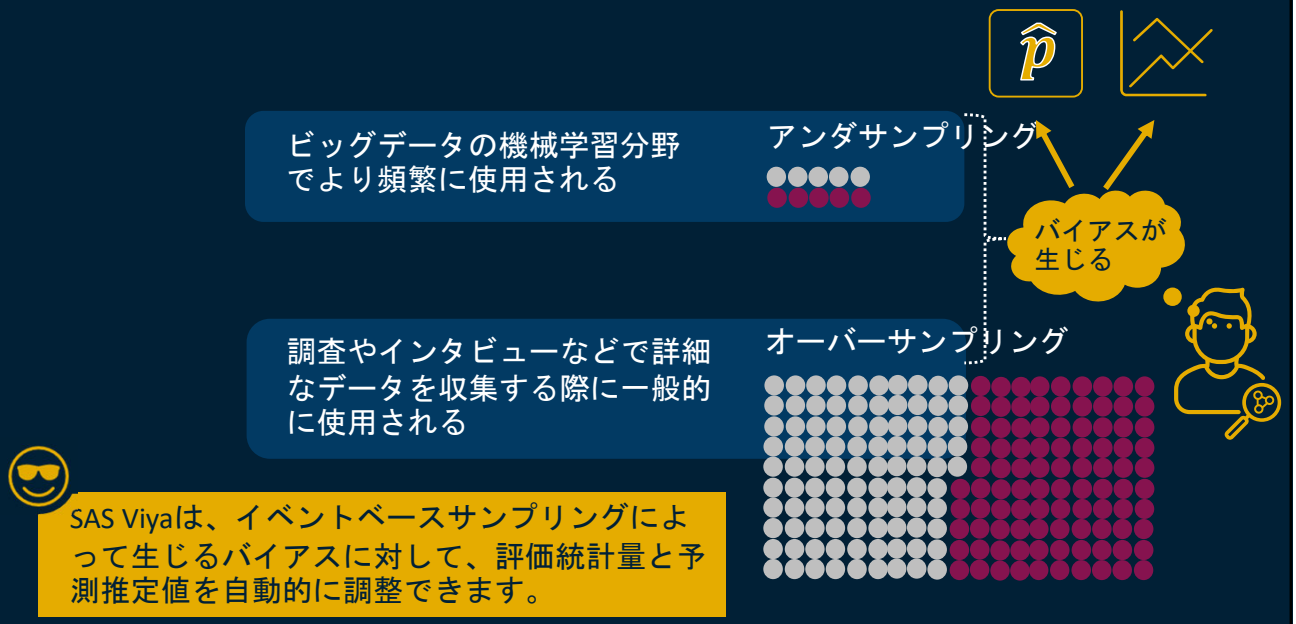
イベントベースサンプリング



イベントベースサンプリング



イベントベースサンプリング



SAS Viyaは、イベントベースサンプリングによって生じるバイアスに対して、評価統計量と予測推定値を自動的に調整できます。

オーバーサンプリングとアンダーサンプリング：長所と短所

問題：イベントベースサンプリング

オーバーサンプリングもアンダーサンプリングも、データ内の不均衡な分布を補正するために、あるクラスから他のクラスよりも多くの事例を取り入れることでバイアスを導入するものです。

a. 正しい
b. 誤り

回答：イベントベースサンプリング

オーバーサンプリングもアンダーサンプリングも、データ内の不均衡な分布を補正するために、あるクラスから他のクラスよりも多くの事例を取り入れることでバイアスを導入するものです。

a. 正しい
b. 誤り

分析データ

ある退役軍人団体が、寄付を中断した寄付者からの継続的な寄付を求めています。

対象層

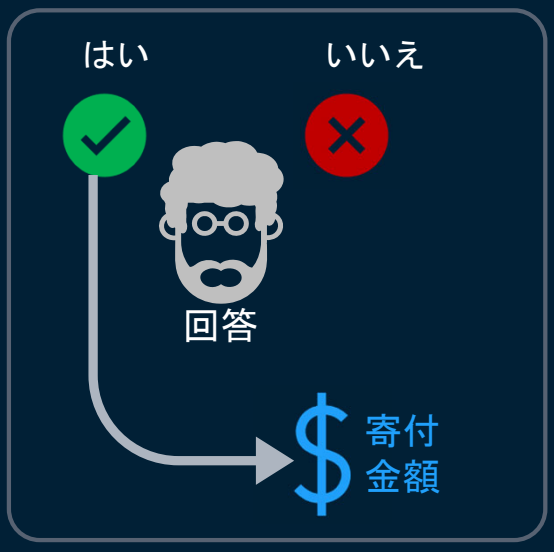
- 350万人
- 実際の回答率は約5%

対象

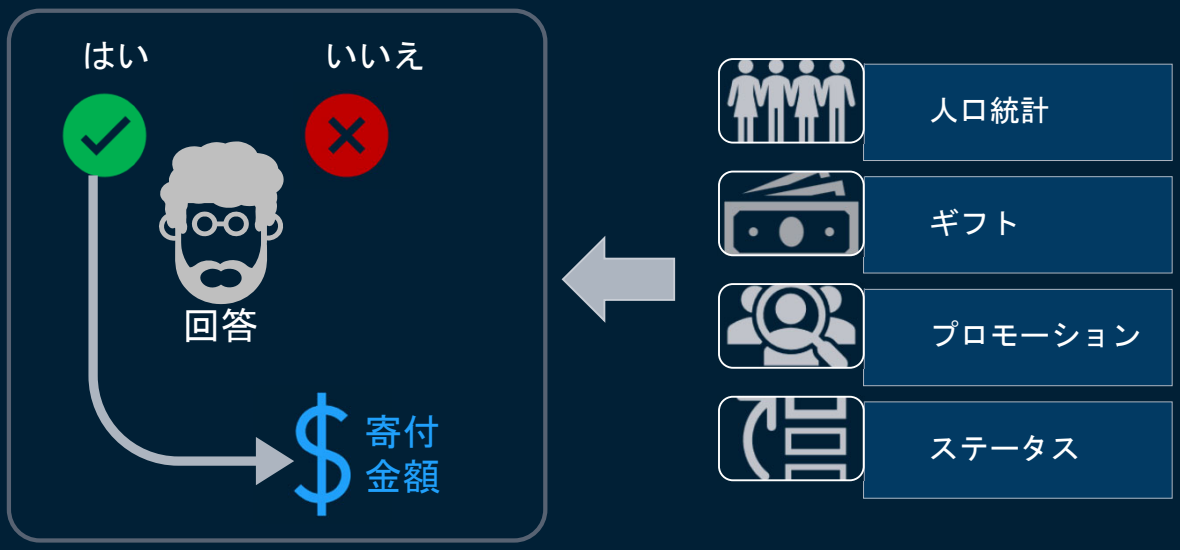
- 前年度のキャンペーンから選定された9,686名
- 標本では回答率と非回答率が均衡している

出典：ACM 1998 KDD-Cup コンテスト。アーカイブ：http://kdd.ics.uci.edu/

分析データ



分析データ



分析の目的

レッスン2、3、4、5の分析目標は以下の通りです：

2
- データを記述する。相関関係、傾向、および異常値を探る。
- 平均寄付額が100ドルであるかどうかを検証する。

→ 記述統計
→ 推測統計

3
- 人口統計、寄付、プロモーション、ステータスといった要因が、寄付額の説明に役立つかどうかを判断する。

→ 連続型ターゲットを用いた説明的モデリング

4
- 過去のキャンペーンにおける寄付中断者の反応データを用いて、将来の寄付中断者の反応を予測する。

→ 二値ターゲットを用いた予測モデリング

5
- データ前処理タスクの一部を実行し、機械学習アルゴリズムを実行する。

→ データ前処理と機械学習

観測結果の一覧

このデモでは、PVA_DONORSデータからいくつかの観測値をリスト表示する方法を示します。

記述統計

記述統計の理解

Lesson 02, Section 02

データの説明

1

データから無作為にサンプルを選択してください。

5 240 68 4021 149 509

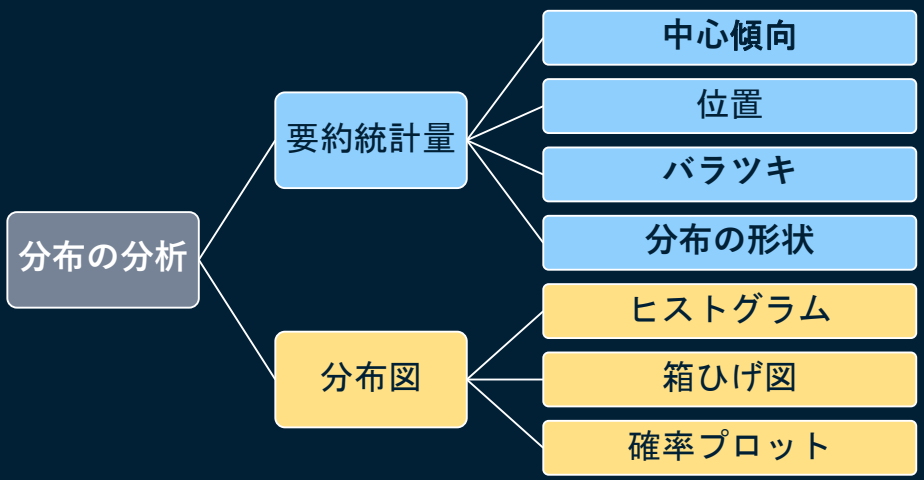
2

記述統計を用いてデータをより深く理解する。

データを記述する際の目標

- 異常なデータ値をスクリーニングする
- データの広がりと形状を確認する
- **中心的な傾向を特徴づける**
- 予備的な結論を導き出す

データの説明



中心傾向の尺度

- 平均
- 中央値
- 最頻値

平均=3　データセット内の全値の合計を値の数で割ったもの

中央値=1.5　中央の値（50th パーセンタイルとも呼ばれる）

最頻値=1　最も多く見られるデータ値

バラツキの尺度

- 四分位
- 十分位
- パーセンタイル

98
95
92　75th パーセンタイル=91
90
85
81　50th パーセンタイル=80
79
70
63　25th パーセンタイル=59
55
47
42

第3四分位

四分位数はデータを4等分

第1四分位数

問題：中心傾向

データは昇順に並べられています。最小値と最大値を除外した場合、以下のどの統計量に影響が出るでしょうか？

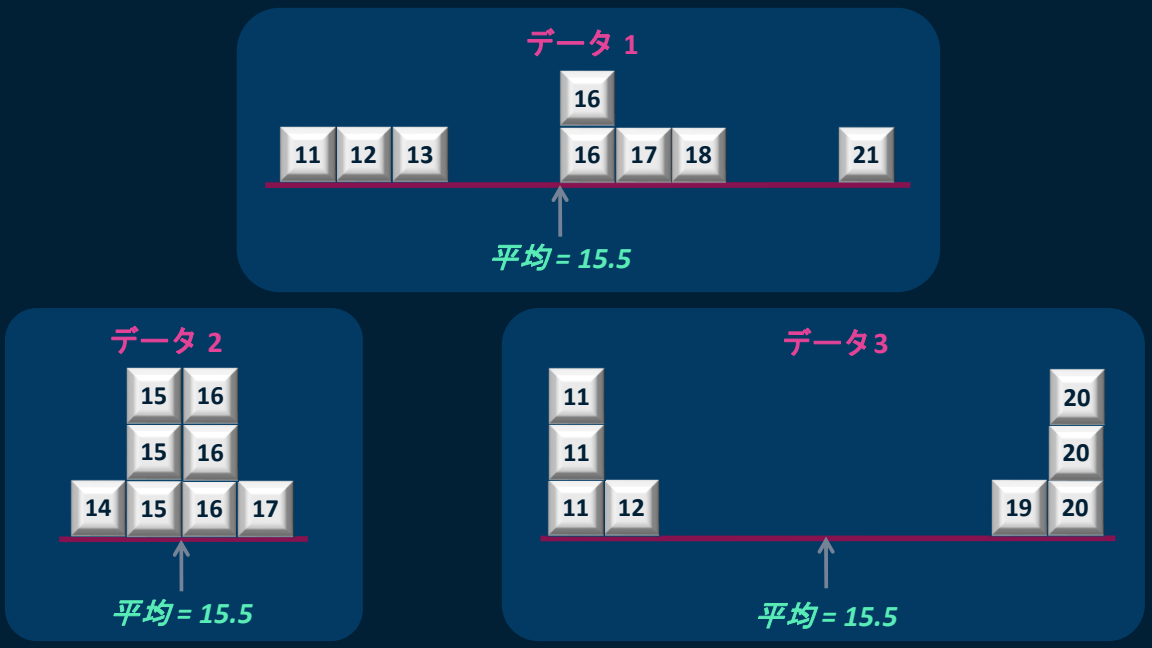
a. 平均
b. 中央値
c. 50パーセンタイル
d. いずれも該当しない

回答：中心傾向

データは昇順に並べられています。最小値と最大値を除外した場合、以下のどの統計量に影響が出るでしょうか？

a. 平均
b. 中央値
c. 50パーセンタイル
d. いずれも該当しない

バラツキの尺度



バラツキの尺度

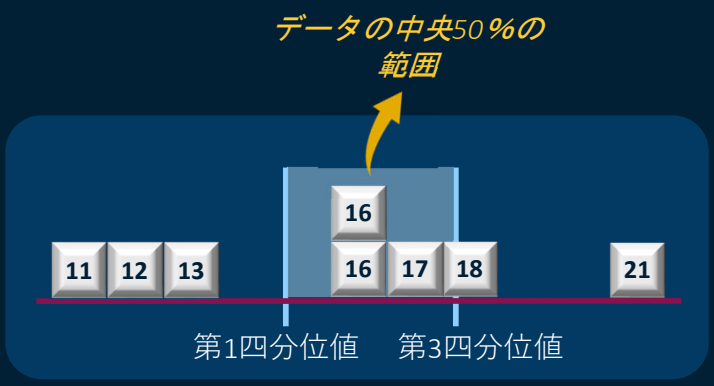
- 範囲

$$Range = x_{max} - x_{min}$$

バラツキの尺度

- **四分位範囲**

$$IQR = Q_3 - Q_1$$



バラツキの尺度

- **平均絶対偏差**

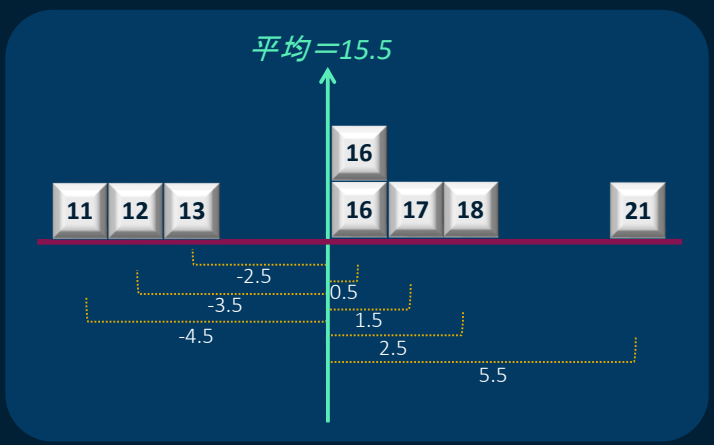
$$MAD = \frac{\sum|x_i - \bar{x}|}{n-1}$$

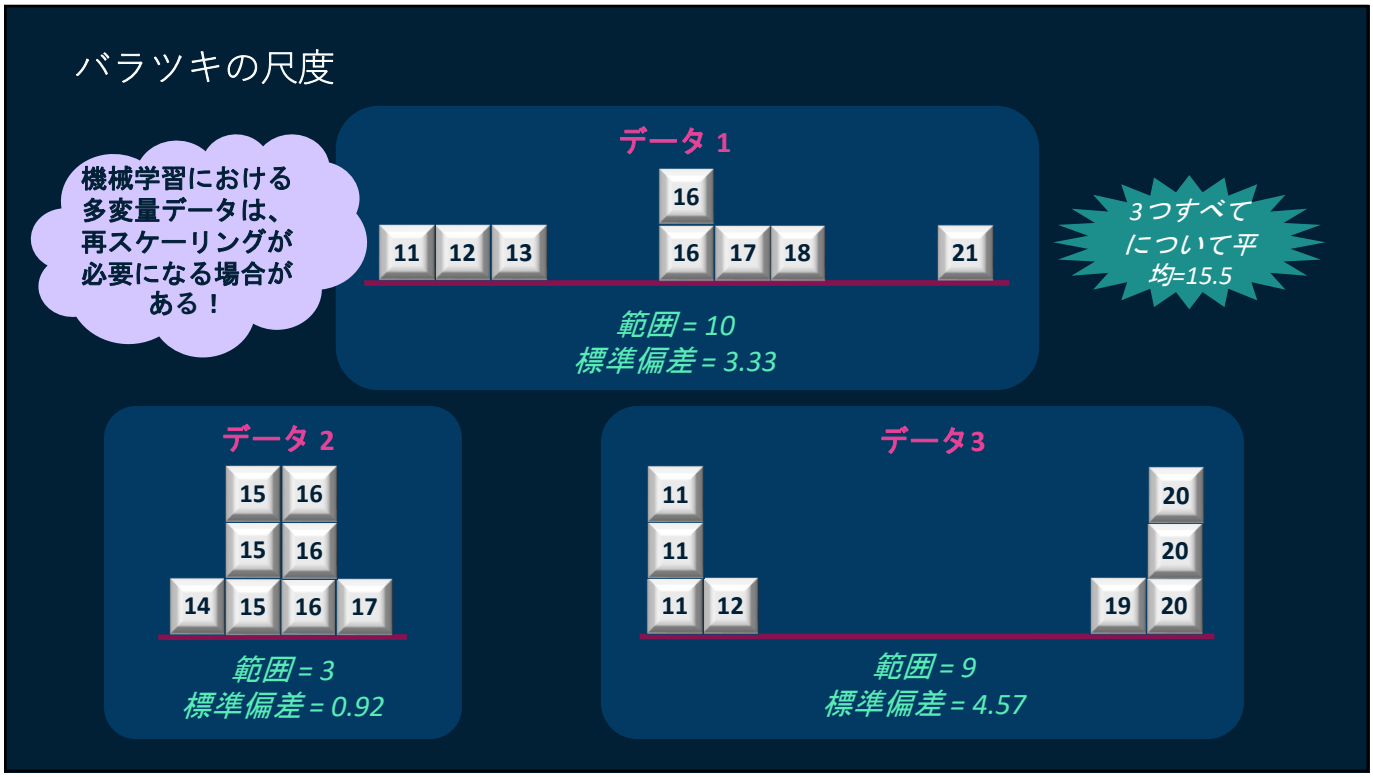
- **分散**

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- **標準偏差**

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$





問題：分散

標準偏差と分散は、本質的に以下のどれからの偏差を表すか？

a. 最大値
b. 中央値
c. 最頻値
d. 平均値

回答：分散

標準偏差と分散は、本質的に以下のどれからの偏差を示すか？

a. 最大値
b. 中央値
c. 最頻値
d. 平均値

問題：要約統計量

かわいそうなホセは、頭はオーブンの中に、足は冷凍庫の中だ！

あなたは統計学者として、ホセの体の各部位の温度のばらつきを調査しています。次のうち、妥当な記述はどれですか？

a. 平均的には、彼はまったく問題ないはずだ。
b. （温度）データの中心を特定すれば十分だ。
c. （温度）データの変動性やばらつきを特徴づける必要がある。
d. 彼は「熱血漢」でありながら「臆病者」なのだ！

回答：要約統計量

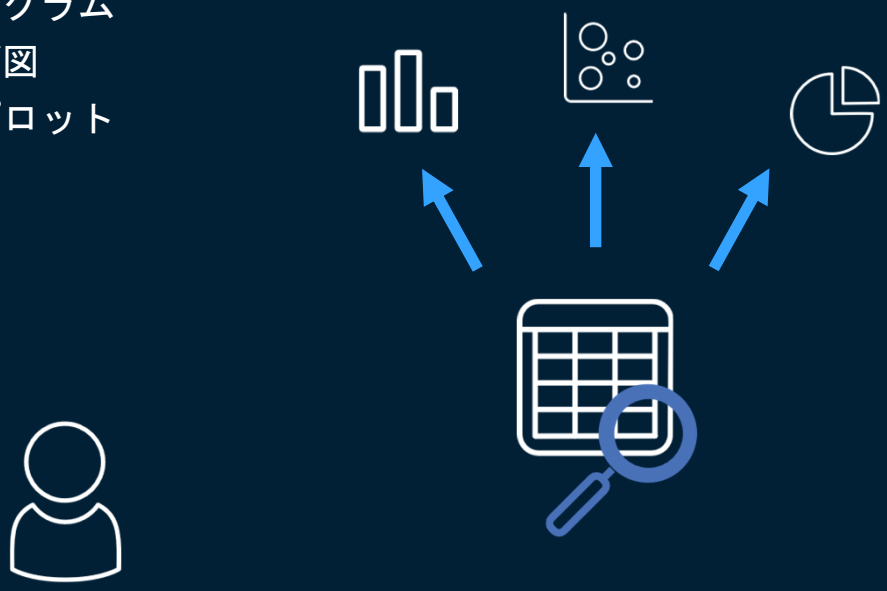
かわいそうなホセは、頭はオーブンの中に、足は冷凍庫の中だ！

あなたは統計学者として、ホセの体の各部位の温度のばらつきを調査しています。次のうち、妥当な記述はどれですか？

a. 平均的には、彼はまったく問題ないはずだ。
b. （温度）データの中心を特定すれば十分だ。
c. （温度）データの変動性やばらつきを特徴づける必要がある。
d. 彼は「熱血漢」でありながら「臆病者」なのだ！

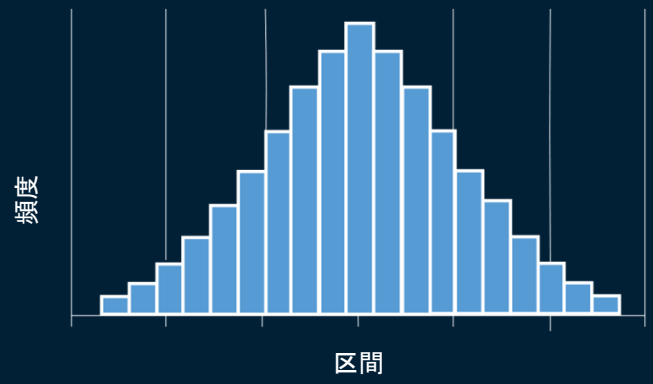
分布の可視化

- ヒストグラム
- 箱ひげ図
- 確率プロット

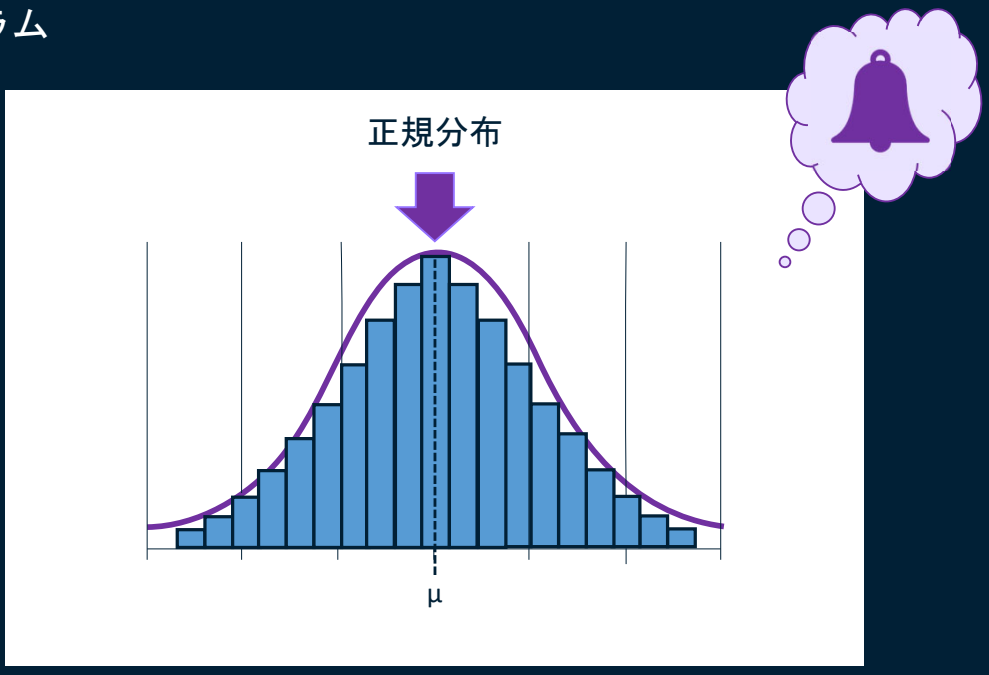


ヒストグラム

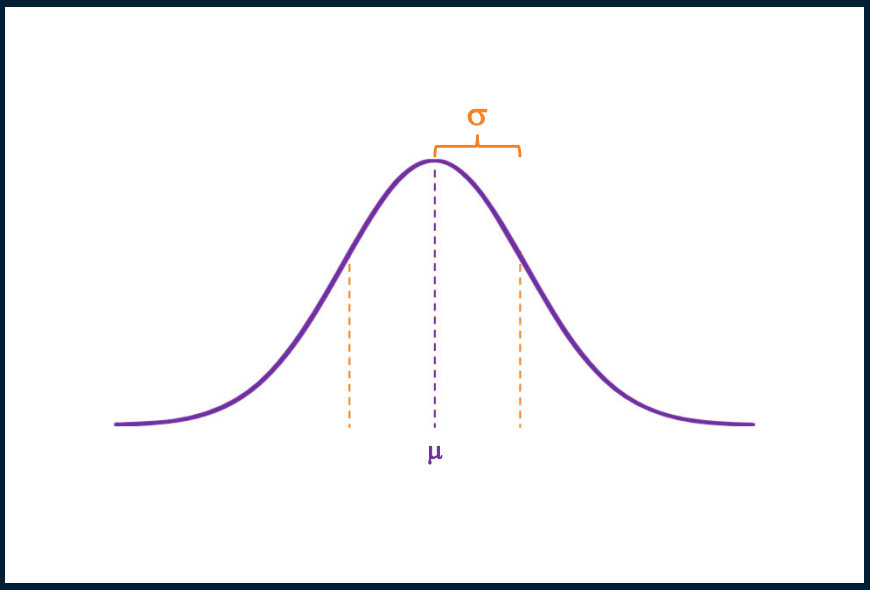
データの頻度分布を
視覚的に表したもの



ヒストグラム

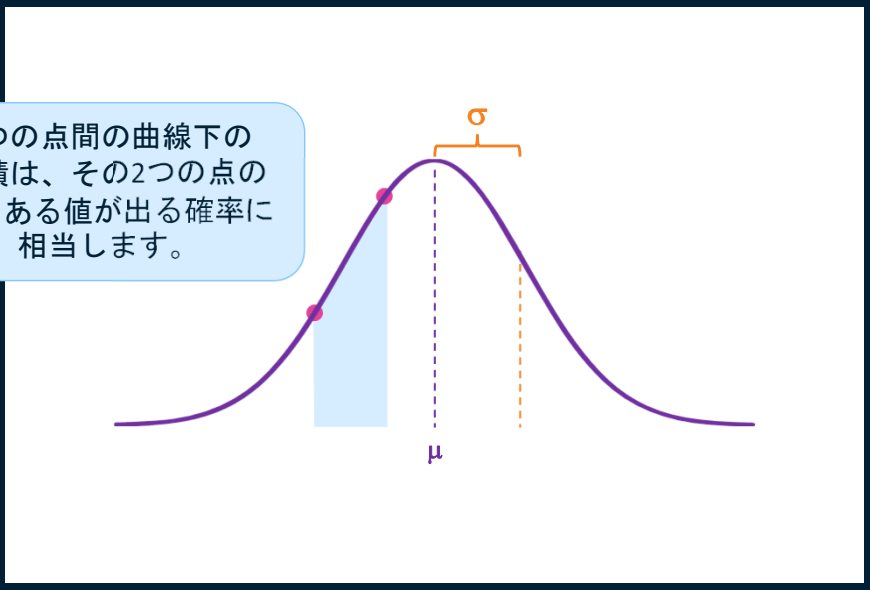


正規（ガウス）分布

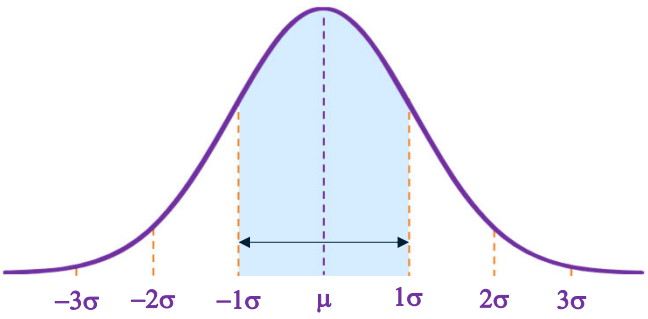


正規（ガウス）分布

2つの点間の曲線下の面積は、その2つの点の間にある値が出る確率に相当します。

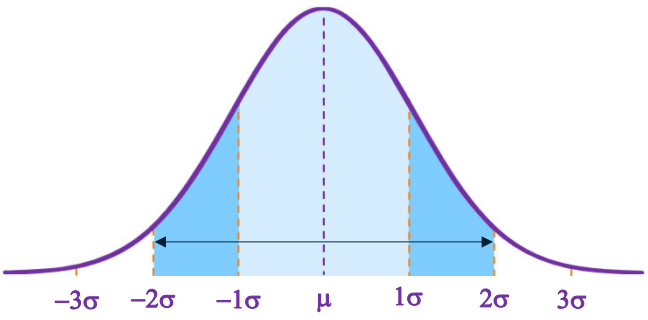


正規（ガウス）分布



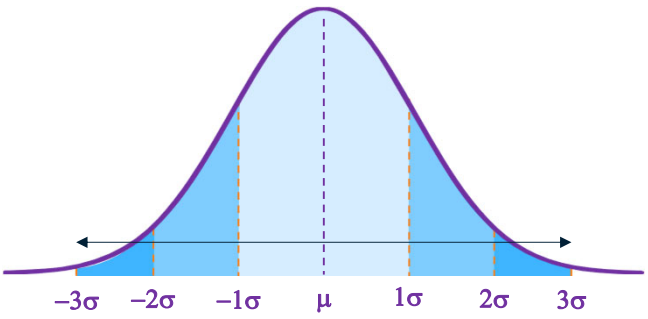
正規分布の約68%は、
平均値から1標準偏差の範囲内に収まる。

正規（ガウス）分布



正規分布の約95%は、
平均値から2標準偏差の範囲内に収まります。

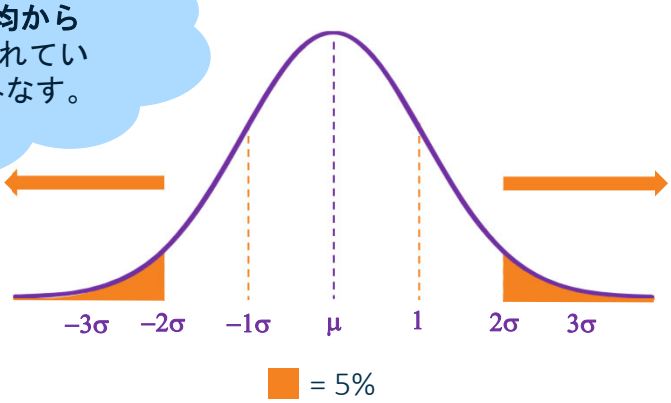
正規（ガウス）分布



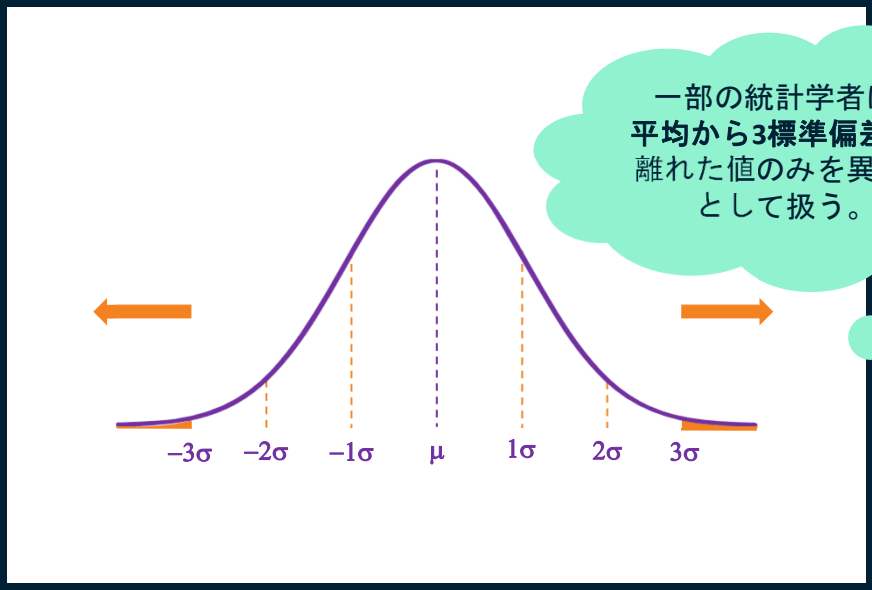
正規分布の約99.7%は、
平均値から3標準偏差の範囲内に収まる。

正規（ガウス）分布

統計学者は、**平均から2標準偏差**以上離れている値を異常値とみなす。



正規（ガウス）分布



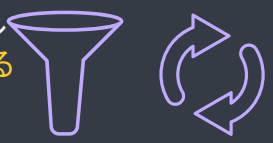
一部の統計学者は、
平均から3標準偏差以上
離れた値のみを異常値
として扱う。

機械学習における正規分布の有用性

機械学習モデルのためのデータ前処理

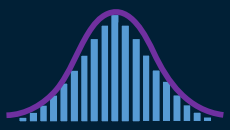
平均からの標準偏差の数を指定し、下限と上限を導出する……

…極端な観測値をフィルタリングおよび置換するために。



機械学習における正規分布の有用性

モデル構築



正規分布に従うデータ

計算が簡単になる！

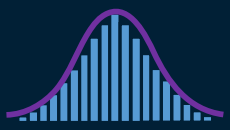
パラメトリック機械学習アルゴリズム

- ロジスティック回帰
- 線形判別分析
- ナイーブベイズ
- ニューラルネットワーク

関数を既知の形式に単純化する

機械学習における正規分布の有用性

モデル構築



正規分布に従うデータ

以下には不要

ノンパラメトリック機械学習アルゴリズム

- k-近傍法
- 決定木
- フォレスト
- グラディエントブースティング
- サポートベクターマシン
- 深層ニューラルネットワーク

写像関数の形式について強い仮定を置かない

問題：正規（ガウス）分布

次のうち、正規分布の典型的な特徴はどれか。
（該当するものをすべて選択）

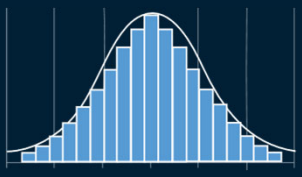
a. 多峰性
b. 対称で、ベル型
c. 平均、中央値、最頻値がすべて一致する
d. 値の約99.7%が平均から±3標準偏差の範囲内に収まる

回答：正規（ガウス）分布

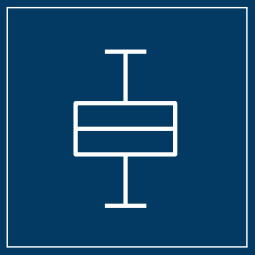
次のうち、正規分布の典型的な特徴はどれか。
（該当するものをすべて選択）

a. 多峰性
b. 対称で、ベル型
c. 平均、中央値、最頻値がすべて一致する
d. 値の約99.7%が平均値から±3標準偏差の範囲内に収まる

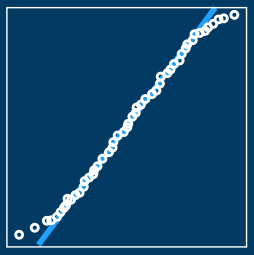
ヒストグラム以外のプロット



箱ひげ図

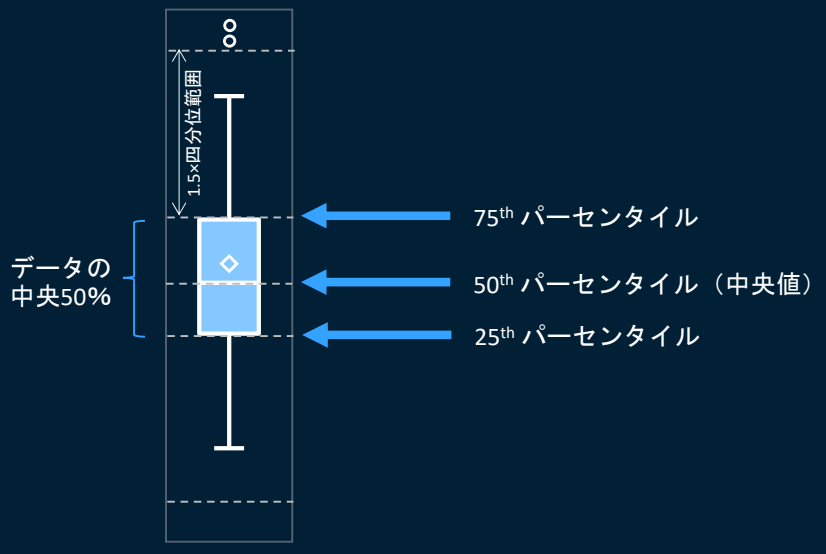


正規確率プロット



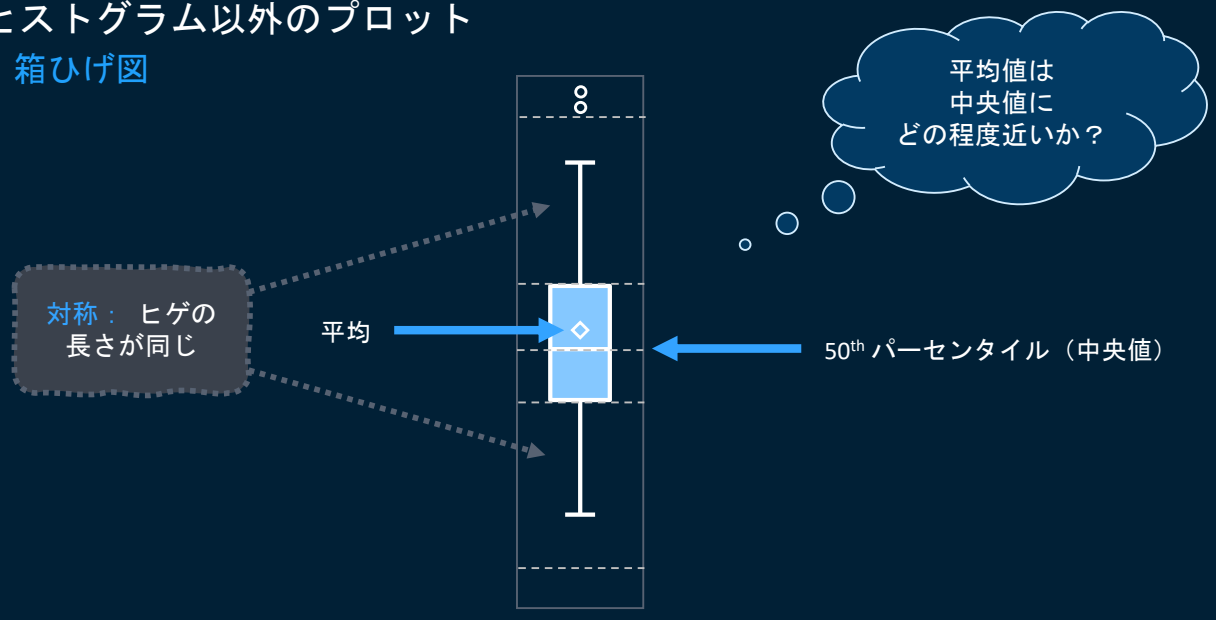
ヒストグラム以外のプロット

箱ひげ図



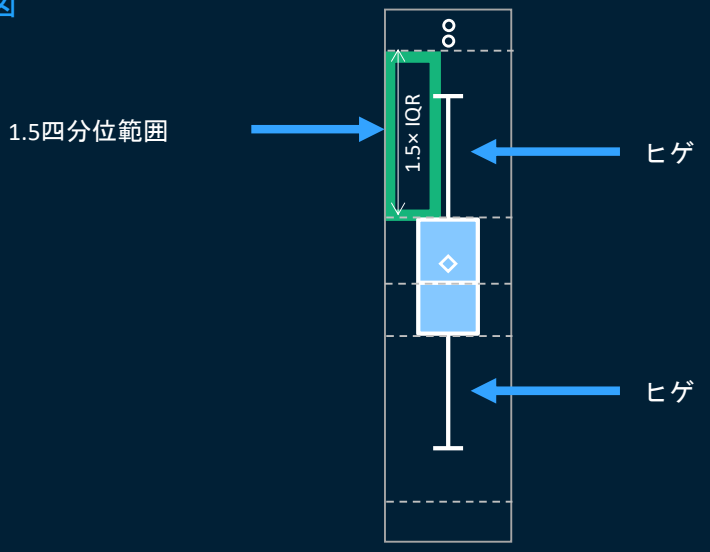
ヒストグラム以外のプロット

箱ひげ図



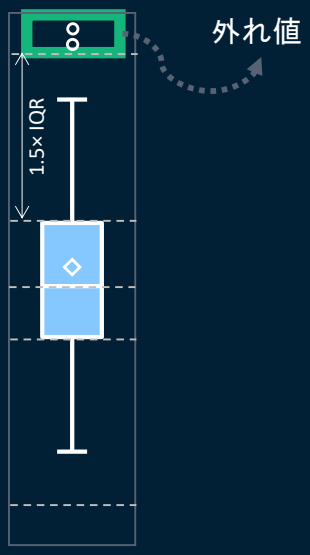
ヒストグラム以外のプロット

箱ひげ図



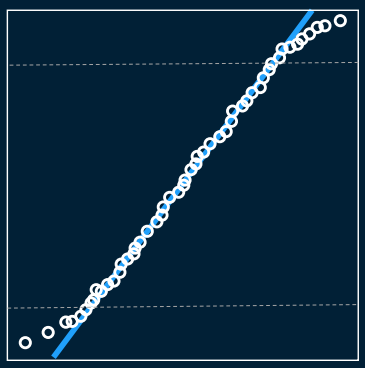
ヒストグラム以外のプロット

箱ひげ図



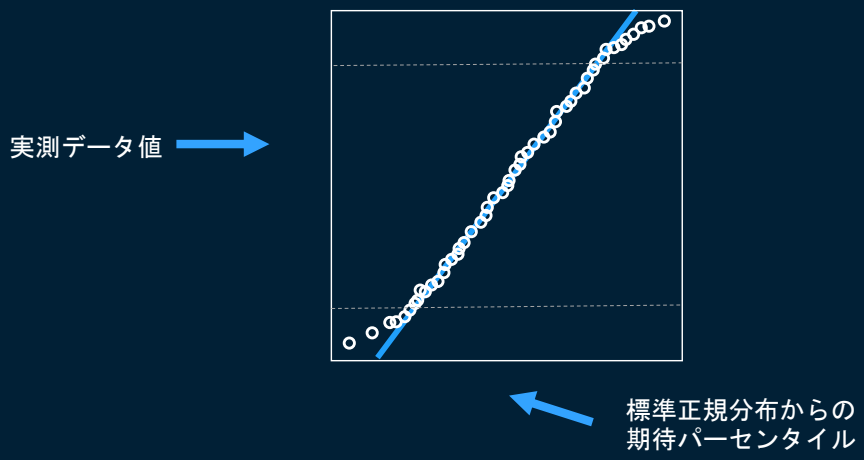
ヒストグラム以外のプロット

正規確率プロット



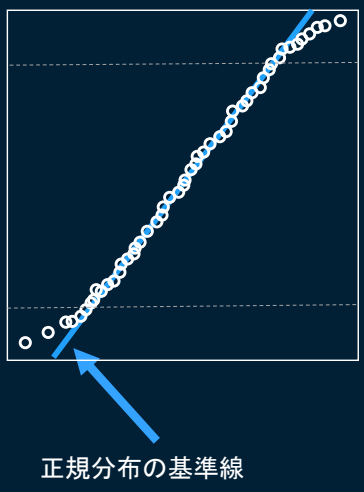
ヒストグラム以外のプロット

正規確率プロット



ヒストグラム以外のプロット

正規確率プロット

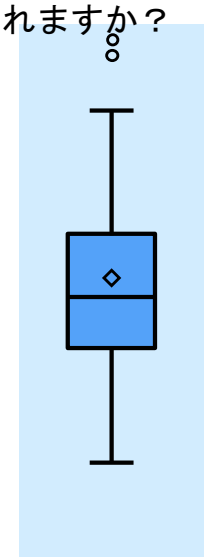


外れ値検出のIQR法でなぜ「1.5」なのか？

問題：分布の可視化

SASでは、箱ひげ図において中央値はどのように表示されますか？

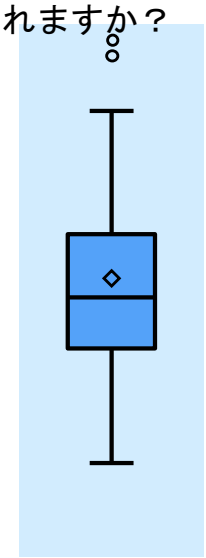
a. 箱
b. 箱ひげ
c. 箱内の線
d. 四分位範囲（IQR）の1.5倍の外側の点



回答：分布の可視化

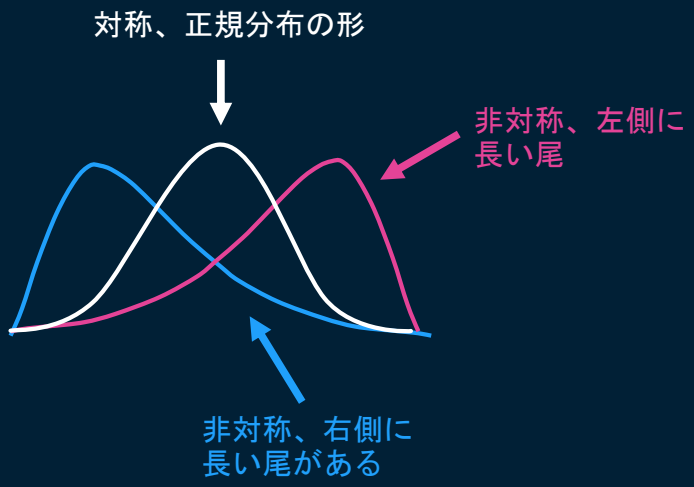
SASでは、箱ひげ図において中央値はどのように表示されますか？

a. 箱
b. 箱ひげ
c. 箱の中の線
d. 四分位範囲（IQR）の1.5倍の外側の点

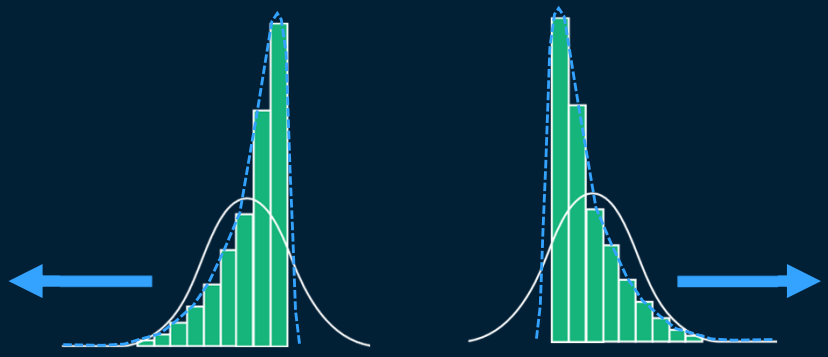


分布形状の尺度：歪度

歪度：分布の非対称性の度合い



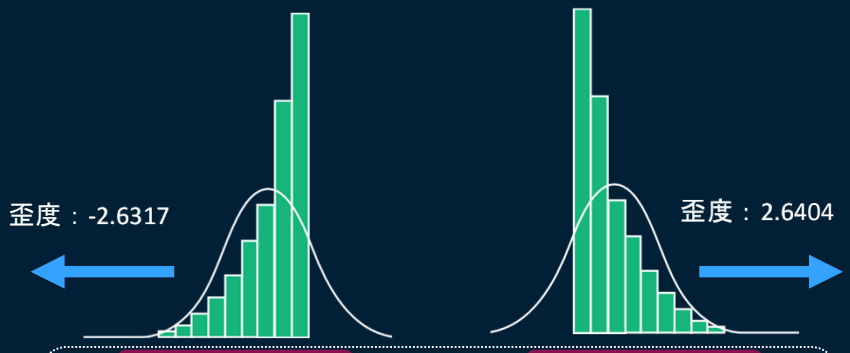
分布形状の尺度：歪度



左偏りの分布の例　　右偏りの分布の例

分布形状の尺度：歪度

歪度の統計量が0に近いほど、データは正規分布に近い、あるいは対称的である。

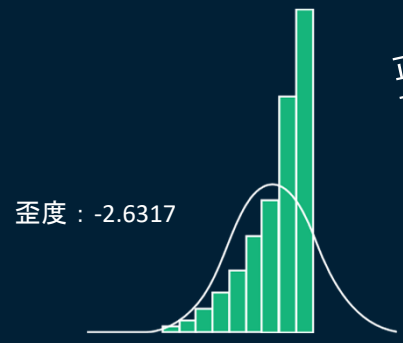


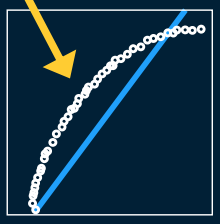
平均 < 中央値　　平均 > 中央値

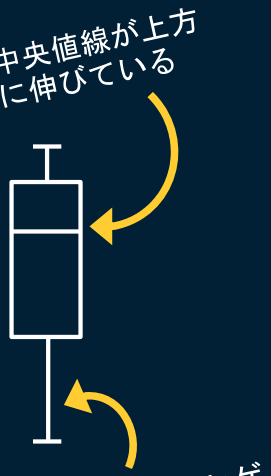
- 多峰性分布
- 片方の尾が長く、もう片方が重い分布

教科書の定説の修正（Hippel 2005）

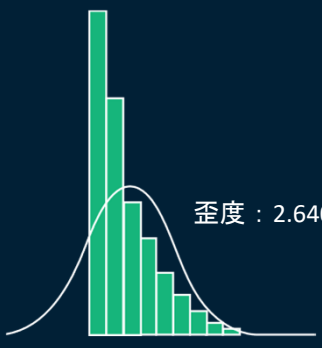
分布形状の尺度：歪度

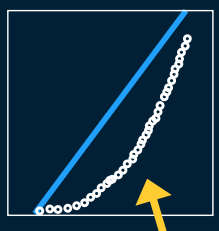


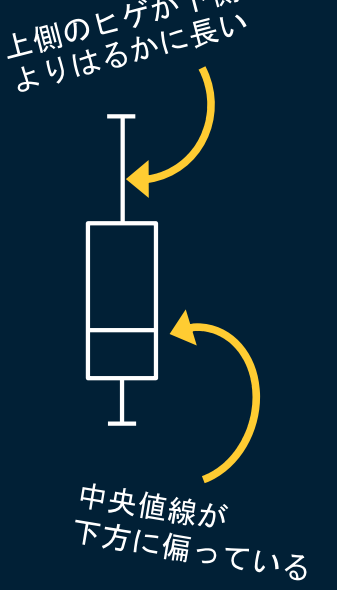




分布形状の尺度：歪度







問題：歪度

正の歪みを持つ分布の場合、以下のうち正しいものはどれか？（該当するものをすべて選択）

a. 非対称な分布である
b. 平均 = 中央値 = 最頻値
c. 分布の左側尾部に値が集中しており、右側尾部はより長い
d. 左偏り分布とも呼ばれる

回答：歪度

正の歪みを持つ分布の場合、以下のうち正しいものはどれか？（該当するものすべてを選択）

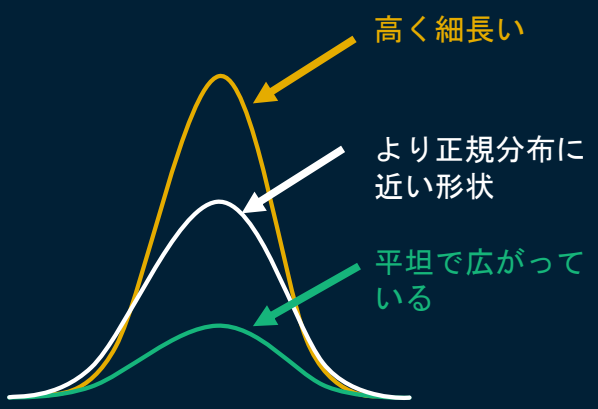
a. 非対称な分布である
b. 平均 = 中央値 = 最頻値
c. 分布の左側尾部に値が集中している一方で、右側尾部はより長い
d. 左偏り分布とも呼ばれる

分布形状の尺度：尖度

尖度：データの「尖り具合」を示す指標

尖度：分布が外れ値を生み出す度合いを測る指標

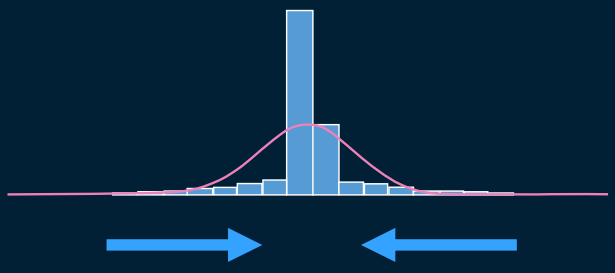
尾部の端（Westfall 2014）



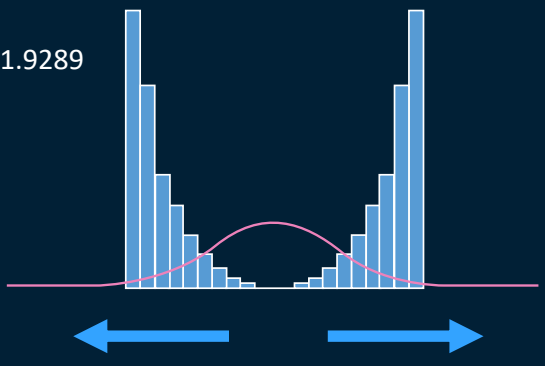
分布形状の尺度：尖度

過剰尖度分布の例

尖度：6.5557



尖度：-1.9289

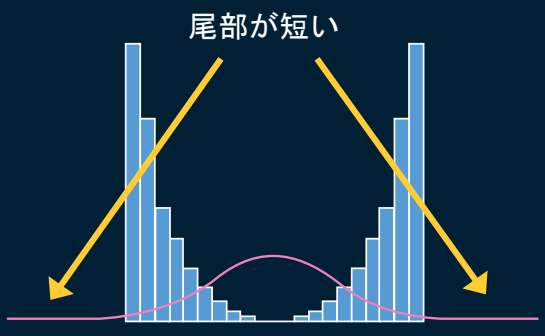


SASでは、正規分布の尖度値は0になります。

分布形状の尺度：尖度

過剰尖度分布の例

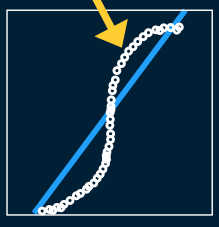
分布形状の尺度：尖度



尖度：-1.9289

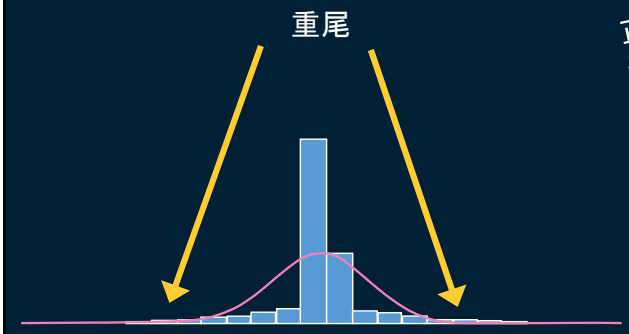
右尾が薄い分布

正規基準線に沿った
S字状の曲線



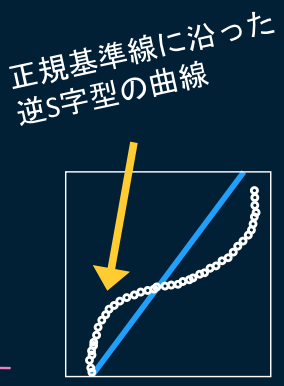
ヒゲの長さは単一の観測値によって決まる

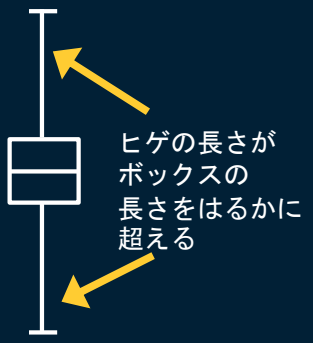
分布形状の尺度：**尖度**

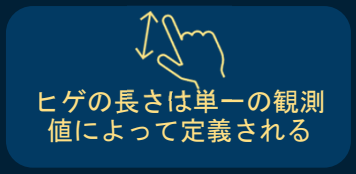


尖度：6.5557

重尾分布







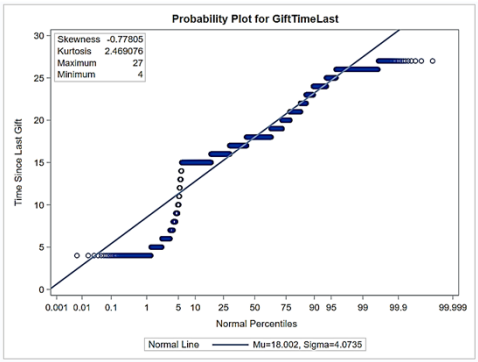
プラティクルティック分布と
レプトクルティック分布

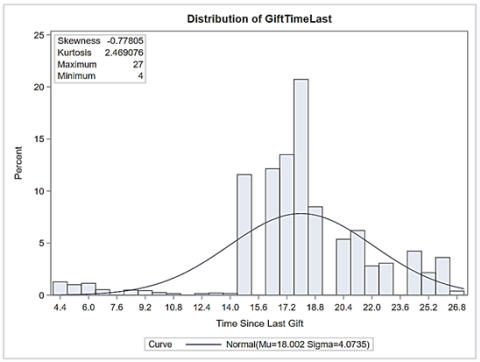
問題：正規分布

「前回のギフトからの経過時間」の正規確率プロットとヒストグラムを調べてください。データがほぼ正規分布に従っていることを示していますか？

a. 正しい

b. 誤り



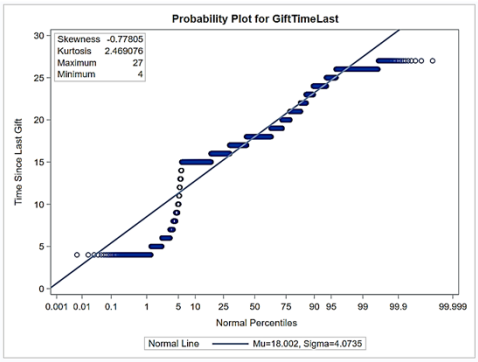


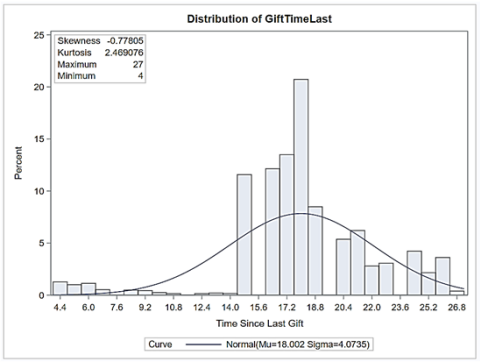
回答：正規分布

「前回のギフトからの経過時間」の正規確率プロットとヒストグラムを調べてください。データがほぼ正規分布に従っていることを示していますか？

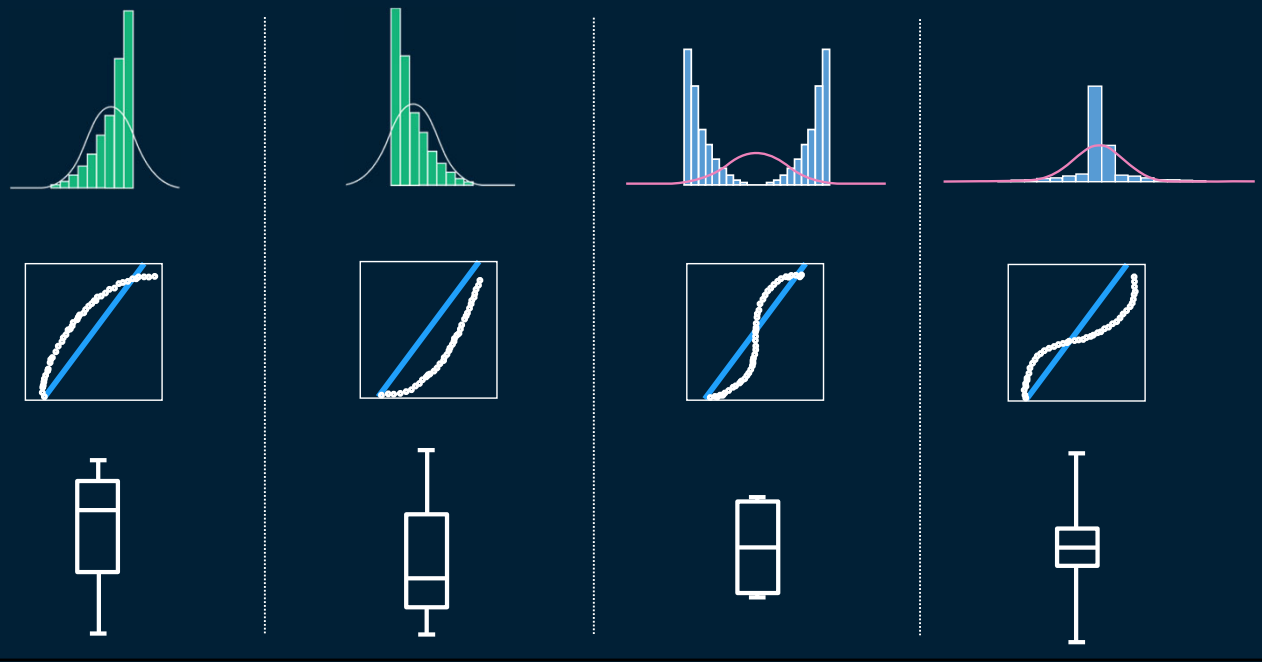
a. 正しい

b. 偽





機械学習における分布分析の有用性



デモ：データの探索

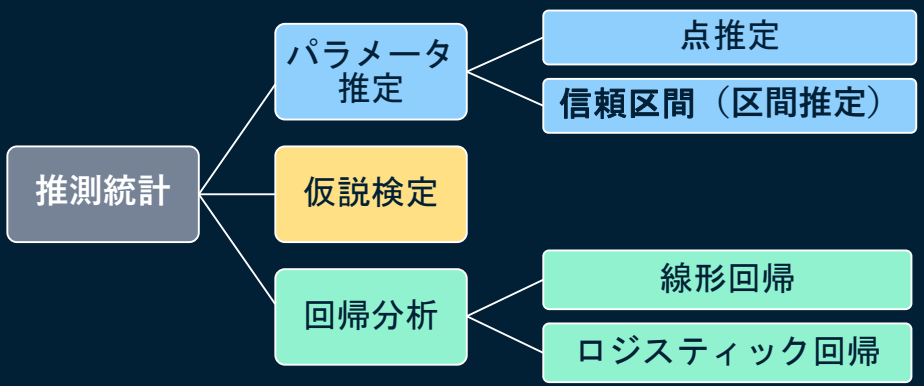
このデモでは、データの初期探索の手順を説明します。連続変数の要約統計量を算出し、その分布を確認します。最後に、カテゴリ変数の頻度表やプロットを作成する方法を学びます。

推測統計

推測統計の理解

Lesson 02, Section 03

データからの推論



機械学習における推測統計学 …

「すでに学習済みのモデルを評価して予測を行う」

点推定

点推定量　標本統計量…

…母集団のパラメータを
推定するために用いられる

標本統計量		母集団のパラメータ
$\bar{x}$		μ
r		ρ
s		σ
s^2		σ^2

$\bar{x}$ は μ を推定する

sはσを推定する

標準誤差

分散

推定量 が標本ごとに
どれほど変動するか

この ***変動性を*** どのように推定すればよいでしょうか？

標準誤差

$$\frac{s}{\sqrt{n}} = s_{\bar{x}}$$

$\bar{x}$ = 1.0

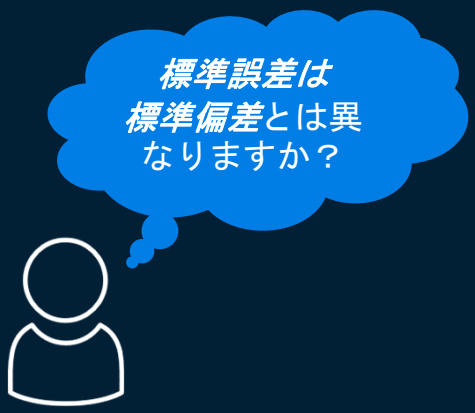
$\bar{x}$ = 0.9

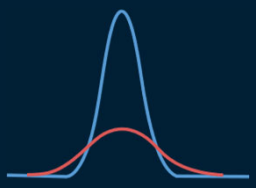
$\bar{x}$ = 1.3

$\bar{x}$ = 0.8

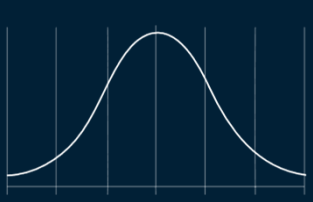
標本統計量は、収集された標本の
直接的な関数である。

標準誤差





標本統計量の変動性



データの変動性

問題：標本統計量の変動

母集団から同じサイズの異なる標本を抽出したとしても、標本統計量の値は変わらない。

a. 正しい
b. 誤り

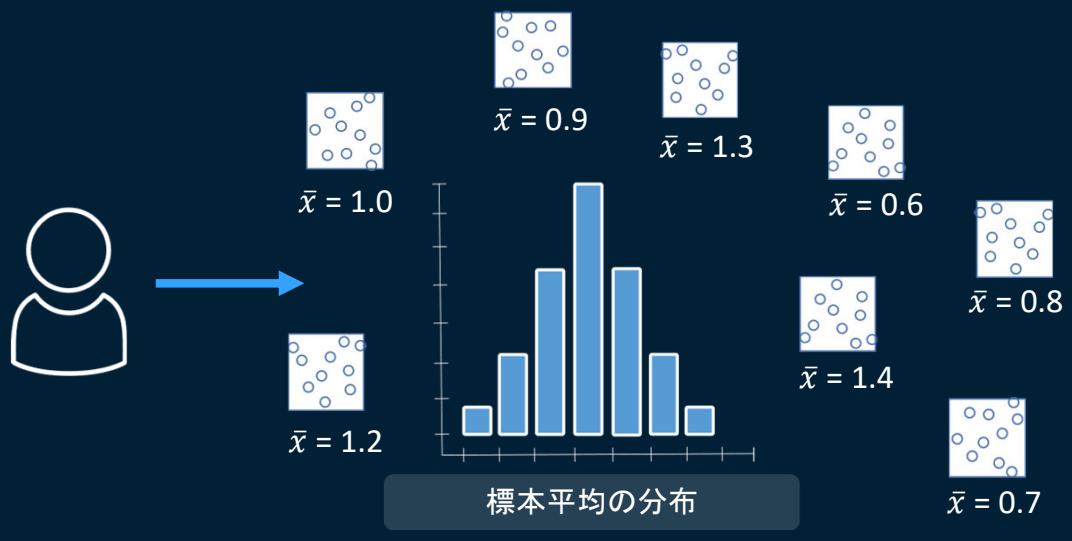
回答：標本統計量の変動

母集団から同じサイズの異なる標本を抽出したとしても、標本統計量の値は変わらない。

a. 正しい

b. 誤り

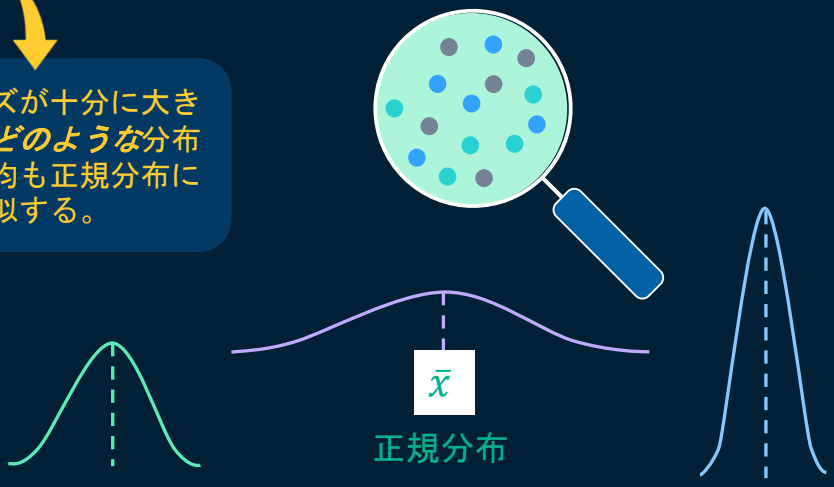
標本分布



標本分布

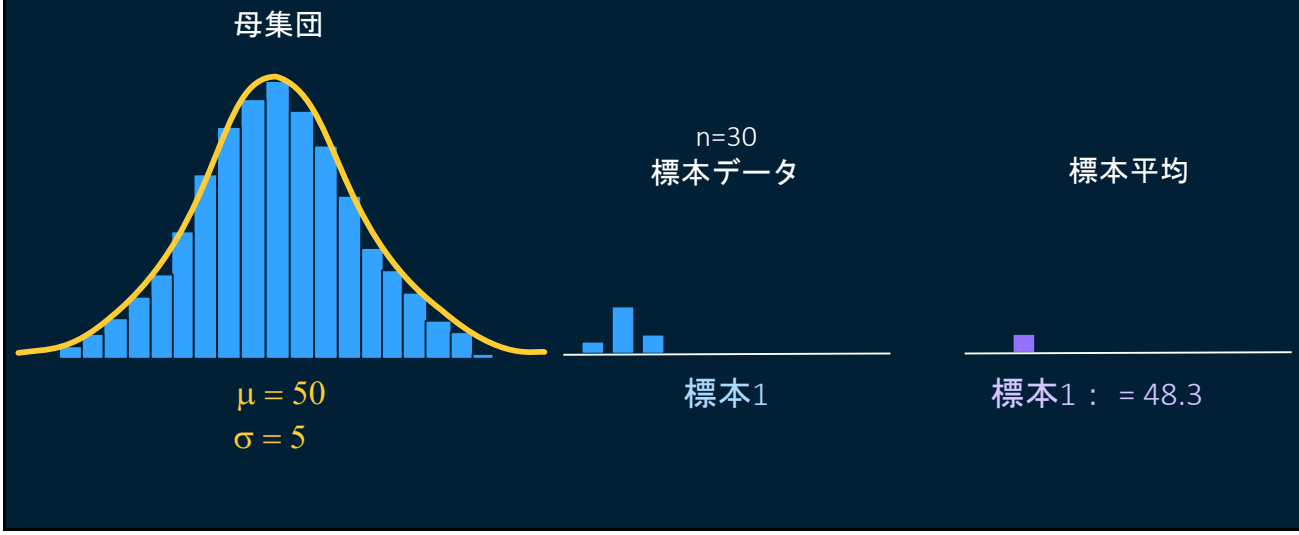
中心極限定理

標本サイズが十分に大きい場合、***どのような***分布の標本平均も正規分布に近似する。



標本分布

中心極限定理

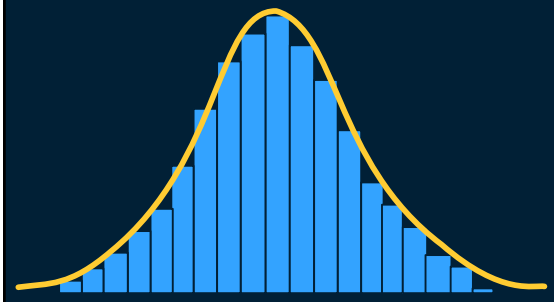


標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3

母集団



$\mu = 50$

$\sigma = 5$

n=30

標本データ

標本2

標本平均

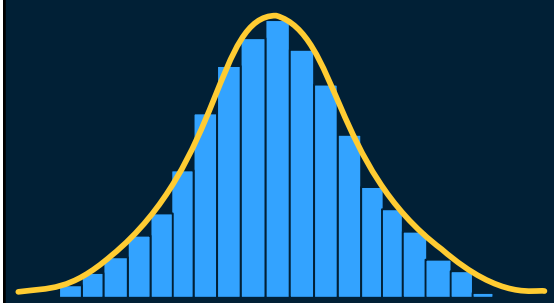
サンプル2： = 52.1

標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1

母集団



$\mu = 50$

$\sigma = 5$

n=30

標本データ

標本3

標本平均

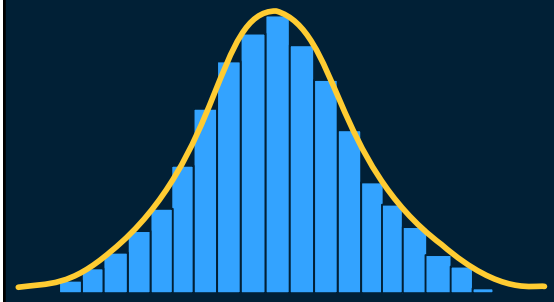
サンプル3： = 47.4

標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4

母集団



$\mu = 50$
$\sigma = 5$

n=30
標本データ

標本4

標本平均

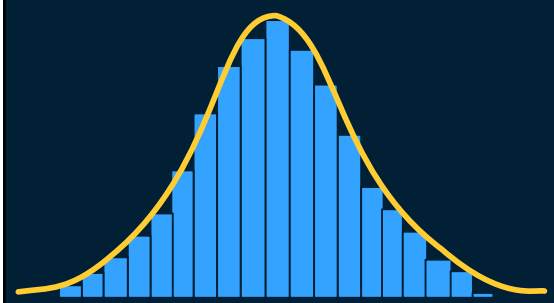
サンプル4： = 51.5

標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4
標本4: $\bar{x}$ = 51.5

母集団



$\mu = 50$
$\sigma = 5$

n=30
標本データ

標本5

標本平均

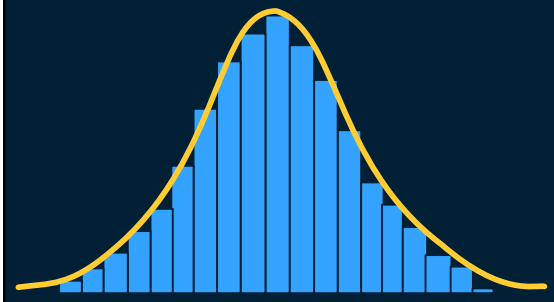
サンプル5： = 50.2

標本分布

中心極限定理

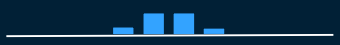
標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4
標本4: $\bar{x}$ = 51.5
標本5: $\bar{x}$ = 50.2

母集団



$\mu = 50$
$\sigma = 5$

n=30
標本データ



標本6

標本平均

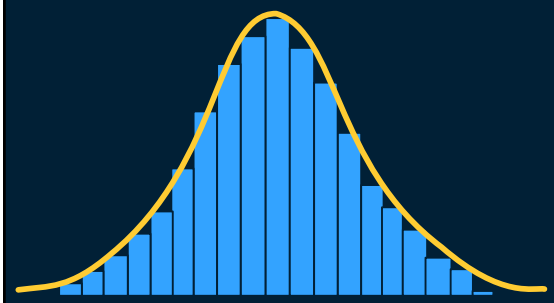
サンプル6： = 49.5

標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4
標本4: $\bar{x}$ = 51.5
標本5: $\bar{x}$ = 50.2
標本6: $\bar{x}$ = 49.5

母集団

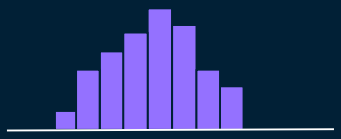


$\mu = 50$
$\sigma = 5$

n=30
サンプルデータ

…

標本平均



…

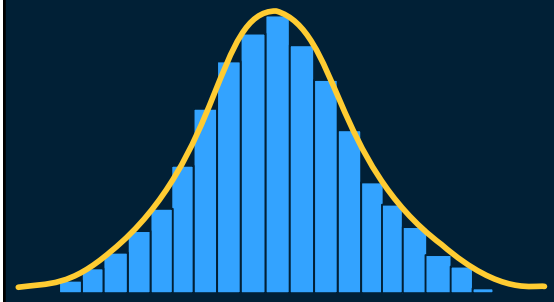
標本分布

中心極限定理

標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4
標本4: $\bar{x}$ = 51.5
標本5: $\bar{x}$ = 50.2
標本6: $\bar{x}$ = 49.5
…

n=30
サンプルデータ

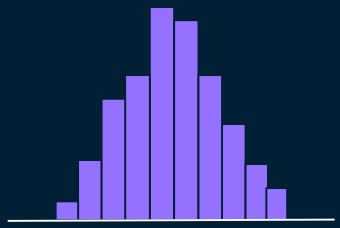
母集団



$\mu = 50$
$\sigma = 5$

…

標本平均



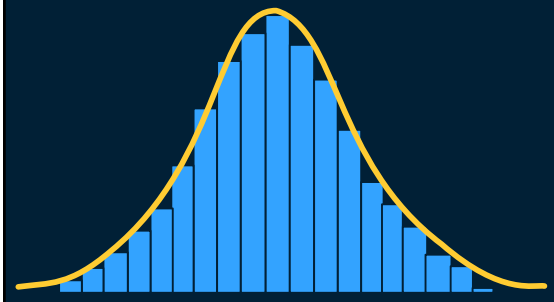
…

標本分布

中心極限定理

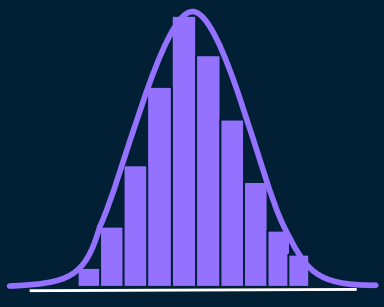
標本1: $\bar{x}$ = 48.3
標本2: $\bar{x}$ = 52.1
標本3: $\bar{x}$ = 47.4
標本4: $\bar{x}$ = 51.5
標本5: $\bar{x}$ = 50.2
標本6: $\bar{x}$ = 49.5
…
標本N: $\bar{x}$ = 49.9

母集団



$\mu = 50$
$\sigma = 5$

標本平均

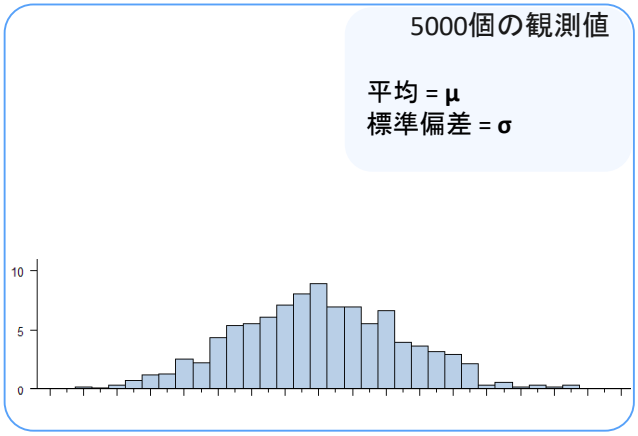


$\mu = 50$
$\sigma_{\bar{x}} = \frac{5}{\sqrt{30}}$

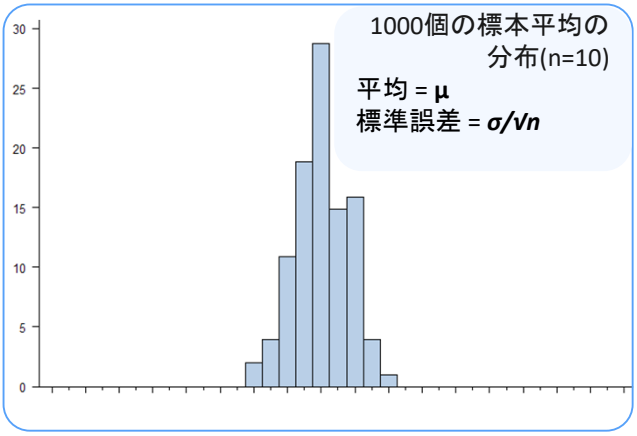
問題：標本分布

SATスコアは、5000件の観測値を持つ対象変数である。10件ずつ1000回の標本抽出を行った。次のうち、標本分布はどれか。

a. SATスコアの分布



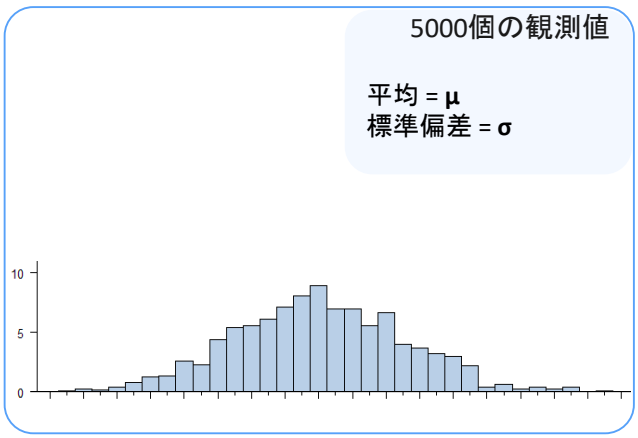
b. SATスコアの平均値の分布



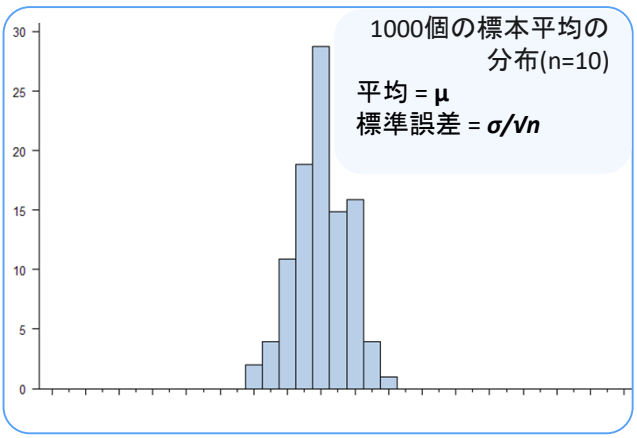
回答：標本分布

SATスコアは、5000件の観測値を持つ対象変数です。10件ずつ1000回の標本抽出を行いました。次のうち、標本分布はどれですか？

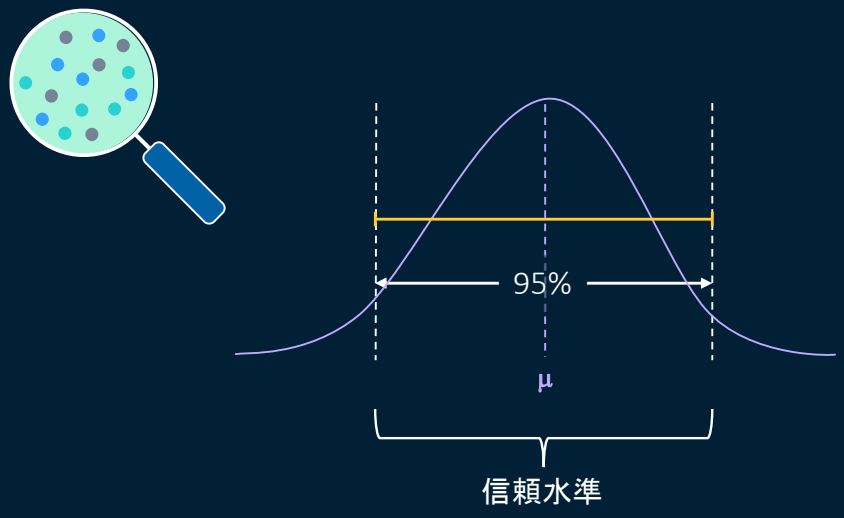
a. SATスコアの分布



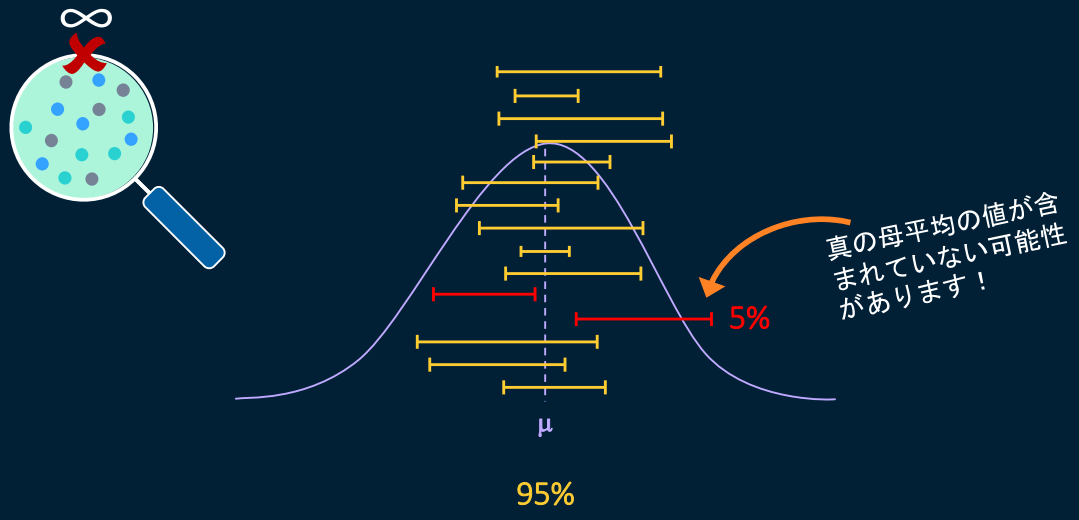
b. SATスコアの平均値の分布



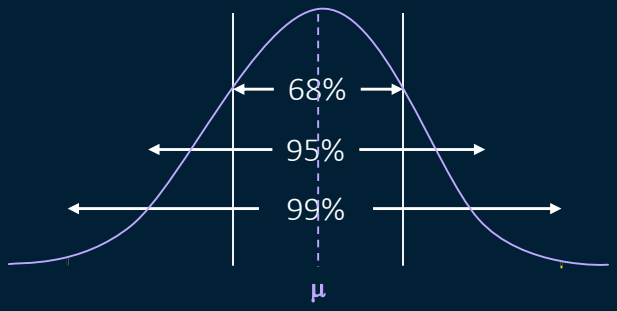
信頼区間



信頼区間

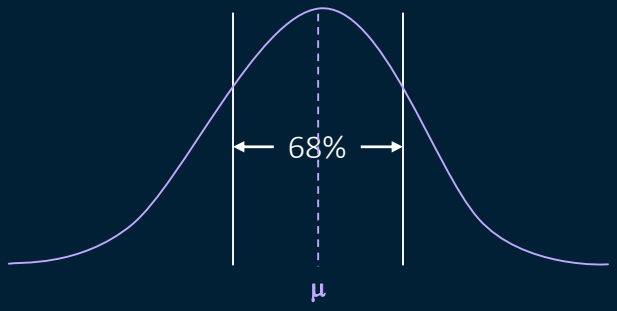


信頼区間



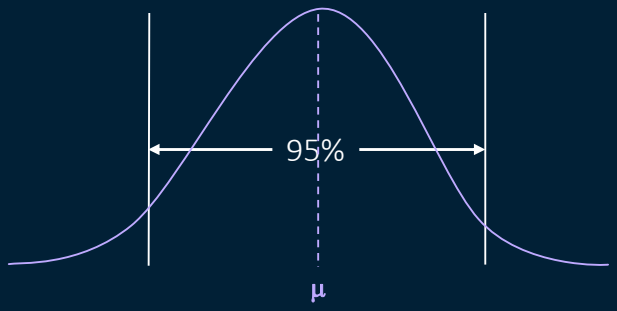
信頼水準を上げると、
区間の幅が広がり、
情報量が減少します。

信頼区間



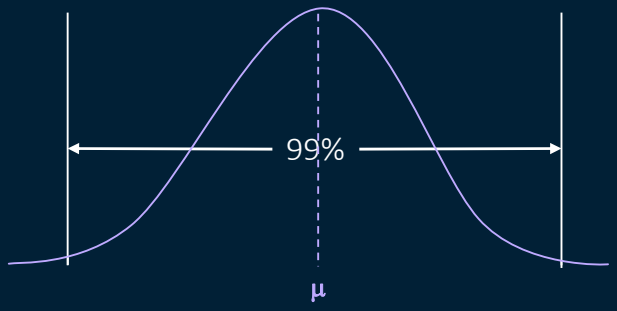
信頼水準を上げると、
区間の幅が広がり、
情報量が減少します。

信頼区間



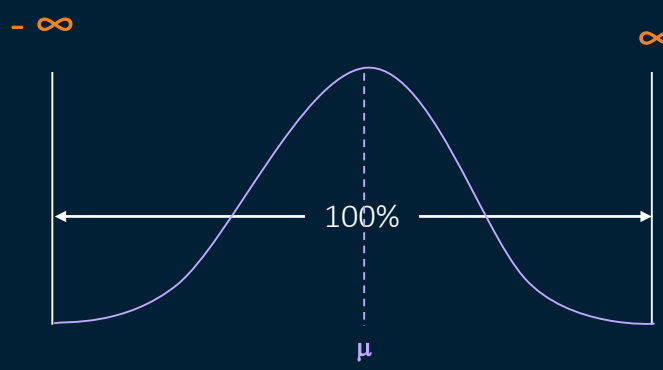
信頼水準を上げると、
区間の幅が広がり、
情報量が減少します。

信頼区間



信頼水準を上げると、
区間の幅が広がり、
情報量が減少します。

信頼区間



信頼水準を上げると、区間の幅が広がり、情報量が減少します。

平均の信頼区間

問題：区間推定

ある研究者が母平均の95%信頼区間を($100, $130)と算出した。標本平均の最も妥当な値はどれか。

a. $100
b. $115
c. $125
d. $230

回答：区間推定

ある研究者が母平均の95%信頼区間を($100, $130)と算出した。標本平均の最も妥当な値はどれか。

a. $100
b. $115
c. $125
d. $230

統計的仮説検定

帰無仮説と対立仮説

仮説検定

H_0 – 等しい
(帰無仮説)

H_a – 等しくない
(対立仮説)

統計的仮説検定

*p*値

帰無仮説が真であると仮定した場合、データで観察されたものと同程度、あるいはそれ以上に極端な検定統計量が得られる確率

p値が低い場合、
帰無仮説の真偽に疑いを抱くことになる。

仮説検定

H_0 – 等しい
(帰無仮説)

H_a – 等しくない
(対立仮説)

帰無仮説を棄却するのに十分な証拠を集められるかどうかを確認。

統計的仮説検定

有意水準

帰無仮説を誤って棄却する一定の確率

仮説検定

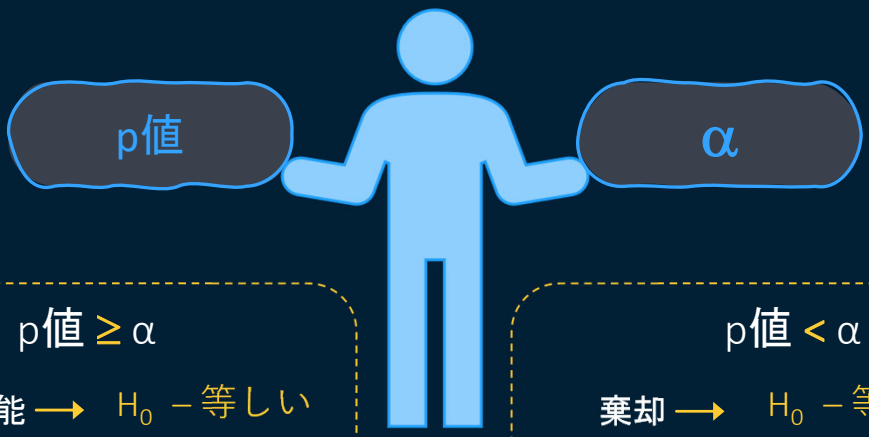
H_0 – 等しい
（帰無仮説）

H_a – 等しくない
（対立仮説）

一般的な有意水準は**0.05**（20分の1の確率）である。

仮説検定の実施

決定規則



p値 ≥ α

棄却不能 ⟶ H_0 – 等しい

✗ H_a – 等しくない

p値 < α

棄却 ⟶ H_0 – 等しい

✓ H_a – 等しくない

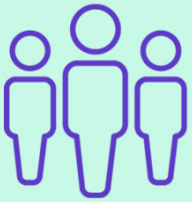
統計的仮説検定

コインの例

そのコインが公正なものであると、どうやって分かるのですか？

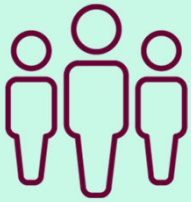
コイン投げの勝者が、どちらのゴールを攻めるかを決める。

表



コイン投げの負けた方がキックオフを行い、試合を開始する。

裏



統計的仮説検定

コインの例

まず、コインが公平であると仮定する。

H_0 – コインは公平である

H_a – コインは公平ではない

有意水準 = ?

次に、有意水準を選択する：

コインを5回投げる 

観察結果：

あなたは、そのコインが不公平であると結論づけます。

または

あなたは、そのコインが不公平であると結論づけます。

そうでなければ、コインが偏っていることを示す証拠が不十分であると判断する。

統計的仮説検定

コインの例

公正なコインの場合…

✓ H_0 – コインは公平である

H_a – コインは公平ではない

有意水準 = 0.0625

…コインを5回投げたとき、5回連続で裏または表が出る確率はどれくらいか？

コインを5回投げる

50% 50% 50% 50% 50%

$(1/2)^5 = 1/32$

50% 50% 50% 50% 50%

$(1/2)^5 = 1/32$

$1/32 + 1/32 = 2/32 = 1/16$

統計的仮説検定

コインの例

もし間違っていたら？

✓ H_0 – コインは公平である

H_a – コインは公平ではない

有意水準 = 0.0625

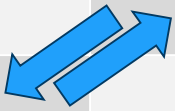
証拠を集める。

コインを5回投げる

統計的仮説検定

誤りの種類

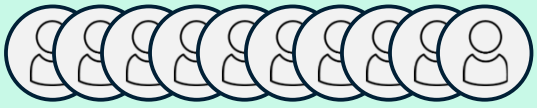
決定＼真実	H_0 が真である	H_0 は偽
H_0を棄却できない	正しい判定 確率 = (1-α) 「信頼水準」 真陰性	**第II種誤差** 確率 = β 偽陰性
H_0を棄却する	**第I種誤り** 確率 = α 「有意水準」 偽陽性	正しい判定 確率 = (1-β) 「検出力」 真陽性



統計的仮説検定

効果量の影響

100回



50 表 - 妥当と思われる

40または60が表 - ?

37または63が表 - ??

15または85が表 - ???

効果量：
*観測された統計量*と
*仮説値*との差

H_0 ：表50%、裏50%

55% → 効果量 = 5%

45%

統計的仮説検定
効果量の影響

100回

p値：帰無仮説が真であると仮定した場合、観測された値と同程度（またはそれ以上）の極端な値が観測される確率。

p値 = 0.3682

55 表

45 裏

差 = 10

p値 = 0.0569

40 表

60 裏

差 = 20

p値 = 0.0120

37 表

63 裏

差 = 26

p値 < 0.0001

15 表

85 裏

差 = 70

統計的仮説検定
標本サイズの影響

x回

p値

40%

60%

効果量

これは公平なコインか？

統計的仮説検定
標本サイズの影響

40%
が表

10回

p値 = 0.7539

4 表

6 裏

統計的仮説検定
標本サイズの影響

40%
表

40回

p値 = 0.7539

4 表

6 裏

10回

p値 = 0.2682

16 表

24 裏

統計的仮説検定
標本サイズの影響

40%
表

100回

p値 = 0.7539

4 表

6 裏

10回

p値 = 0.2682

16 表

24 裏

40回

p値 = 0.0569

40 表

60 裏

統計的仮説検定
標本サイズの影響

40%
が表

400回

公正なコインで
p値が0.0001を下回る
ことはめったにない。

p値 = 0.7539

4 表

6 裏

10回

p値 = 0.2682

16 表

24 裏

40回

p値 = 0.0569

40 表

60 裏

100回

p値 < 0.0001

160 表

240 裏

p値と統計的有意性

p値では測定できないもの
- 効果の大きさを測定することはできない
- 効果の大きさ
- 証拠の強さを

p値が示すのは、特定の仮説的な説明に対して、その結果が得られる確率についてのみである。

問題：仮説検定の用語

以下の用語のうち、観測された統計量と仮説上の値との差として定義できるものはどれか。

a. 効果量
b. 第I種誤り
c. アルファ
d. 検出力

回答：仮説検定の用語

以下の用語のうち、観測された統計量と仮説上の値との差として定義できるものはどれか。

a. 効果量
b. 第I種誤り
c. アルファ
d. 検出力

問題：第I種誤りと第II種誤り

標本サイズを一定とした場合、誤りの種類に関して、以下の記述のうち正しいものはどれか。（該当するものすべてを選択）

a. 第I種誤差率と第II種誤差率は互いに影響し合わない。
b. 第I種誤りと第II種誤りは反比例の関係にある。
c. 信頼水準が高くなると、検定の検出力は高くなる。
d. 第I種の誤り率を設定することで、間接的に第II種の誤り率の大きさにも影響を与える。

回答：第I種誤りと第II種誤り

標本サイズを一定とした場合、誤りの種類に関して、以下の記述のうち正しいものはどれか。（該当するものすべてを選択）

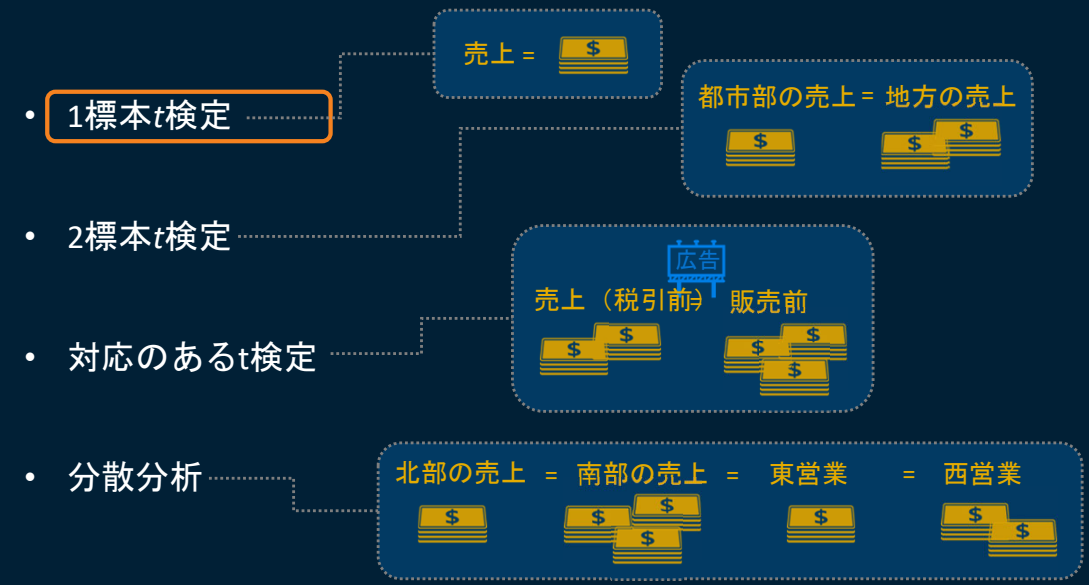
a. 第I種誤差率と第II種誤差率は互いに影響し合わない。

b. 第I種誤りと第II種誤りは反比例の関係にある。

c. 信頼水準が高くなると、検定の検出力は高くなる。

d. 第I種誤差率を設定することで、間接的に第II種誤差率の大きさにも影響を与える。

平均の仮説検定



1標本t検定のシナリオ

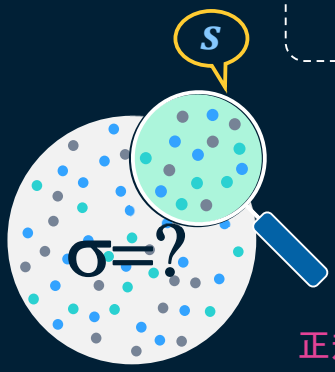
PVA_DONORS データ

寄付額の平均は
100ドルですか？

t検定の実施

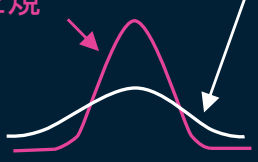
帰無仮説
$H_0 : \mu = \mu_0$

対立仮説
$H_a : \mu \neq \mu_0$

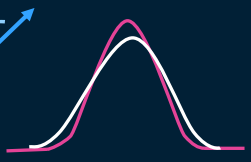


スチューデントのt分布 ： $t = \dfrac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$

正規



標本サイズ



t検定の実施

PVA_DONORS データ

$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

$$t = \frac{(101.64 - 100.00)}{0.3690}$$

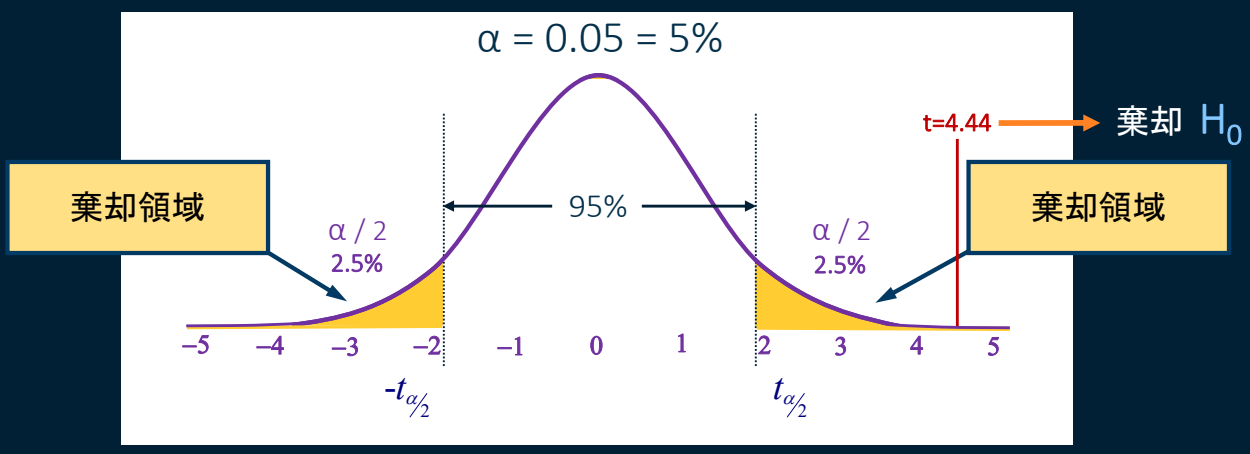
4.44

標本平均が得られる確率

p値 ← 確率

t検定の実施

$H_a : \mu \neq \mu_0$　　両側検定



問題：t分布

標本サイズが大きくなるにつれて、t分布は正規分布に近づいていきます。

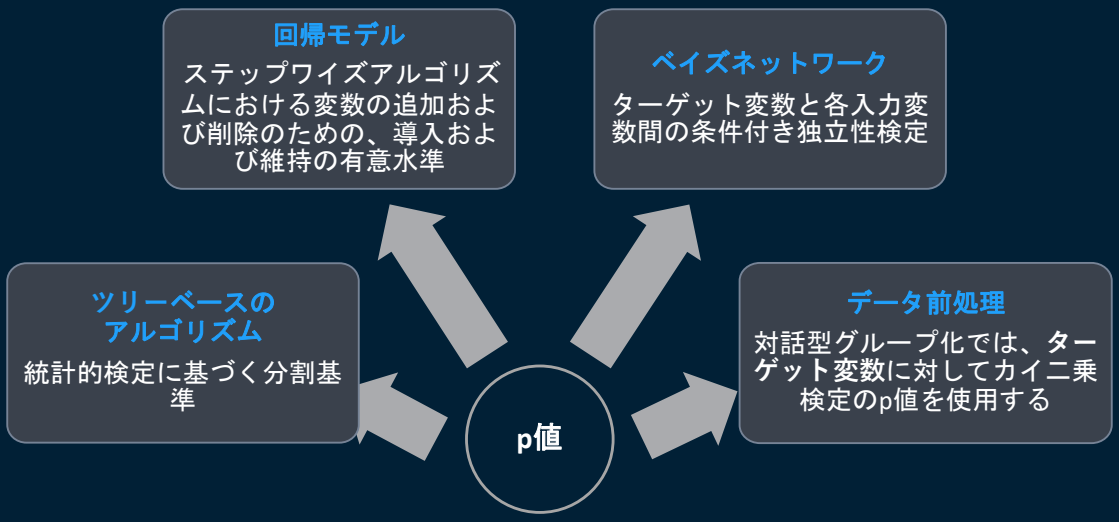
a. 正しい
b. 誤り

回答：t分布

標本サイズが大きくなるにつれて、t分布は正規分布に近づいていきます。

a. 正しい
b. 誤り

機械学習におけるp値



機械学習におけるp値

問題：5%の有意水準（α = 0.05）という恣意的な閾値への過度な依存

統計的有意性

実用的な有意性

統計的な関連性のみを求め、因果関係は求めない

ビッグデータを用いた探索的研究

- より大きな標本サイズ
- より小さい*p値*
- 帰無仮説（H_0）を棄却しやすい

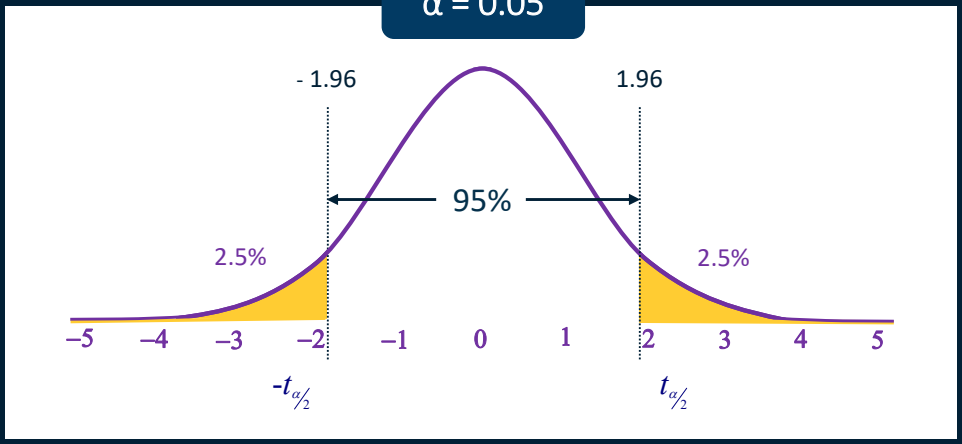
有意水準（α）を小さく設定する

ビッグデータにおいては、有意水準0.01以下がしばしば推奨される

機械学習におけるp値

有意水準の影響

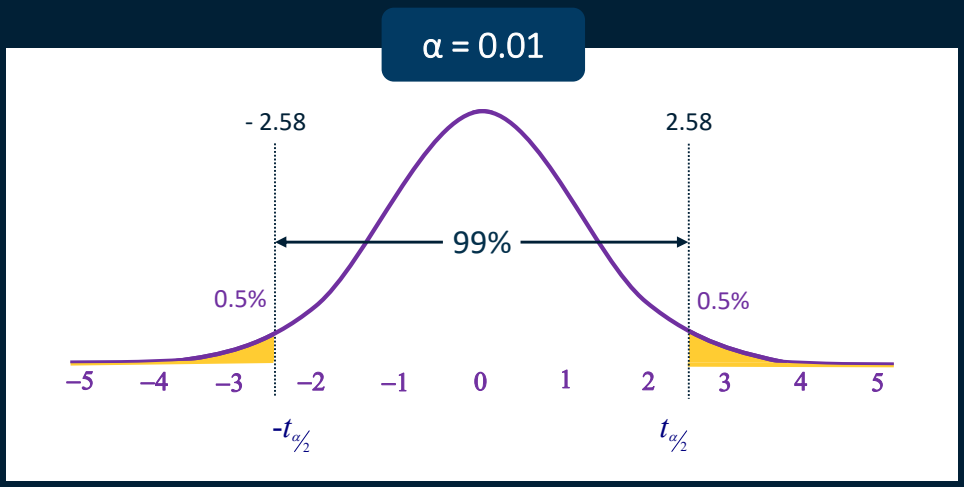
α = 0.05



機械学習におけるp値

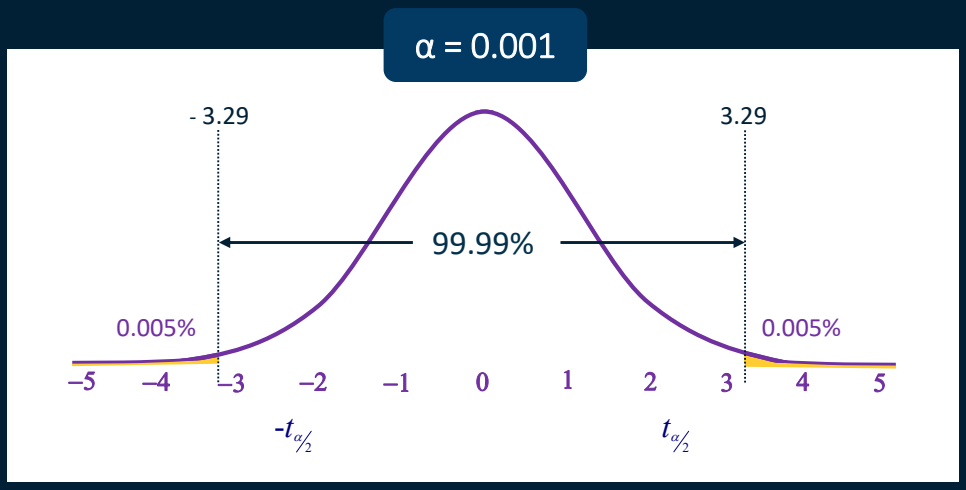
有意水準の影響

α = 0.01



機械学習におけるp値

有意水準の影響



デモ：仮説検定
*t*検定

このデモでは、*t*検定を使用して、母集団の平均が仮説値と等しいかどうかを判定します。

問題：信頼区間

「**過去36ヶ月間の平均贈与額**」の平均値に対する95%信頼区間は（$14.67, $15.08）です。これをもとに、α=0.05の場合、どのような結論が導き出せますか？

a. 「**36ヶ月平均贈与額**」の真の平均値は、$15と有意に異なる。
b. 「36ヶ月平均贈与額」の真の平均値は、$15と有意な差がない。
c. 「36ヶ月平均贈与額」の真の平均は$15未満である。
d. 上記のいずれでもない。信頼区間から統計的有意性を判断することはできない。

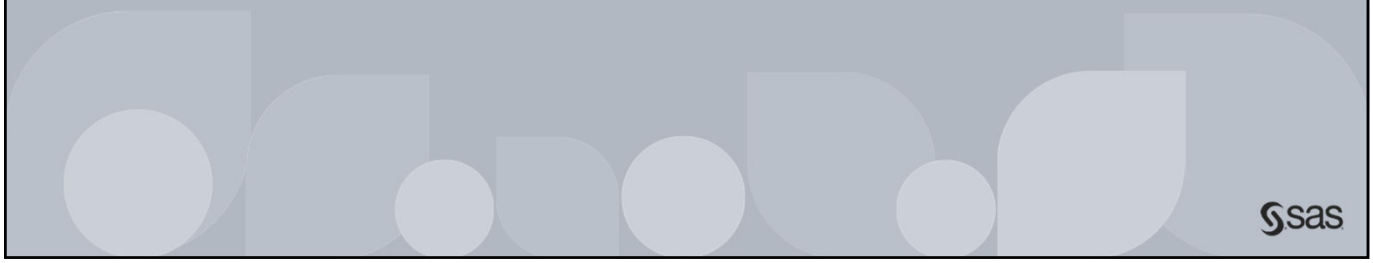
回答：信頼区間

「**過去36ヶ月間の平均贈与額**」の平均値に対する95%信頼区間は（$14.67, $15.08）です。これをもとに、α=0.05の場合、どのような結論が導き出せますか？

a. 「**36ヶ月間の平均贈与額**」の真の平均値は、$15と有意に異なる。
b. 「36ヶ月平均贈与額」の真の平均値は、$15と有意な差がない。
c. 「36ヶ月平均贈与額」の真の平均は$15未満である。
d. 上記のいずれでもない。信頼区間から統計的有意性を判断することはできない。

ご質問はありますか？

レッスンのまとめ



Lesson 3: 線形回帰による説明的モデリング

sas

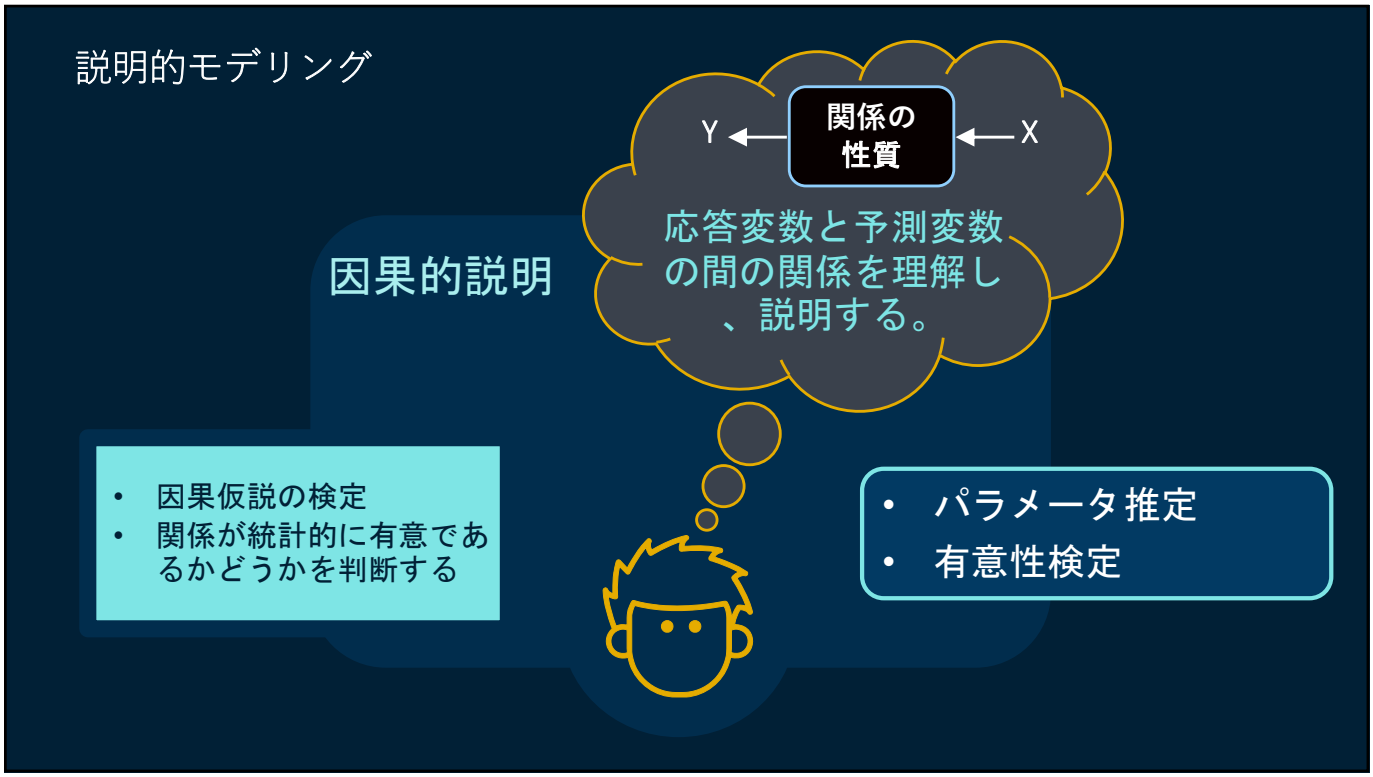
相関と単純線形回帰

相関と単純線形回帰の使い方

Lesson 03, Section 01

sas

説明的モデリング



説明的モデリング

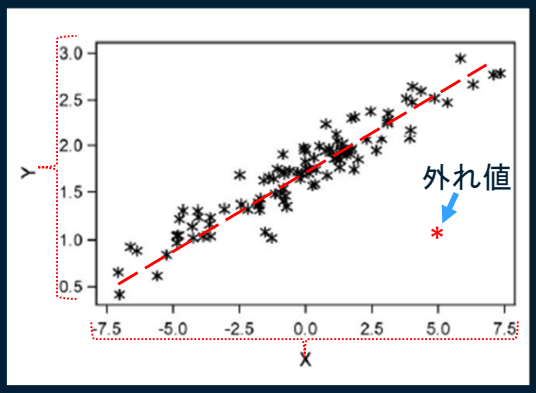
データサイエンスや機械学習に取り組んでいるのに、なぜ説明的モデリングについて学ぶ必要があるのでしょうか？

- 規制当局から、結論（例えば、クレジットスコアリングなど）の説明を求められる場合があります。
- 説明可能な方法で重要な入力を選択する必要があるかもしれません。
- 機械学習モデルによって生成された予測結果を解釈する必要があるかもしれません。
- 必ずしも複雑な機械学習モデルが必要とは限りません。
- 業務において推論を行う必要があるかもしれません。（例：メールマーケティングと郵便マーケティングのどちらが効果的か？）

回帰モデル構築の前にデータを分析する

散布図

連続変数



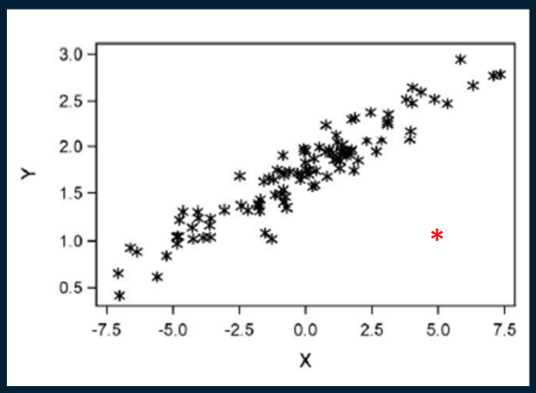
連続変数

散布図：
- **関係性を**特定する
- 値の**範囲**を表示する
- **外れ**値を目立たせる

回帰モデル構築の前にデータを分析する

散布図

連続変数



連続変数

線形相関：
散布図の全体的な形状は**直線**です。

この関係の強さをどのように定量化できるでしょうか？

ピアソン相関係数

相関係数

相関係数

仮説検定

	母集団 パラメータ	標本 統計量
相関	ρ	r

「ロー」

ρを推定する
標本統計量

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

p 値は関連の強さを測定するものでは**ない**。

標本相関係数（r）***は***、
関連の強さを測定する。

問題：散布図

散布図は、以下のうちどの目的を達成するのに役立ちますか？
（該当するものをすべて選択してください）

a. 2つの変数間の関係を調べる
b. 外れ値や異常値を見つける
c. 2つの変数の値の範囲を特定する
d. 相関関係を定量的に評価する

回答：散布図

散布図は、以下のうちどの目的を達成するのに役立つか？
（該当するものをすべて選択してください）

a. 2つの変数間の関係を調べる
b. 外れ値や異常値を見つける
c. 2つの変数の値の範囲を特定する
d. 相関関係を定量的に評価する

相関と共分散の違い

$$correlation\ (r) = \frac{covariance}{std\ dev\ x\ * std\ dev\ y}$$

値の範囲

共分散	$[-\infty, +\infty]$
相関係数	$[-1, +1]$

- PCAにおける固有値のソース行列
- 変数クラスタリングにおける精度行列
- 教師無し変数選択におけるソース行列

相関と共分散の違い

共分散：

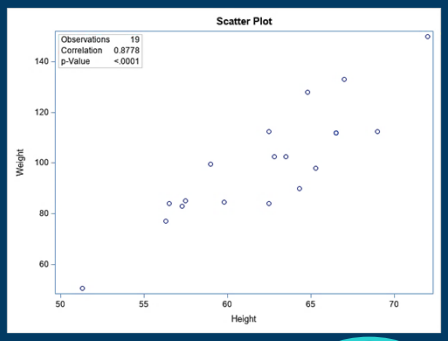
- 単位を持ち、変数のスケーリングの影響を受ける
- 変数の単位を変更すると、変数間の共分散も変化する

相関係数：

- 単位を持たない関連性の尺度
- 変数のスケーリングの影響を受けない

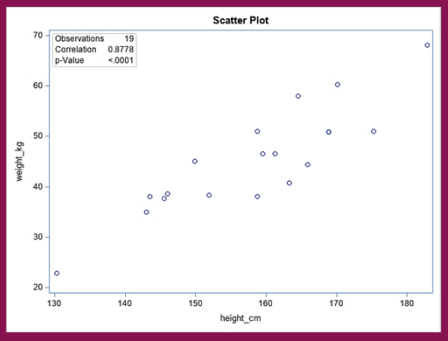
相関と共分散の違い

英単位



体重（ポンド）　相関係数 = 0.8778
身長（インチ）　共分散 = 102.49 ポンド・インチ

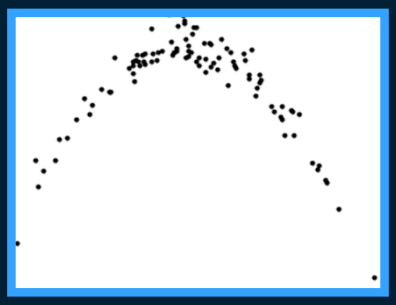
メートル法



体重 (kg)　相関係数 = 0.8778
身長 (cm)　共分散 = 118.08 kg-cm

相関係数に単位なし

一部のデータではピアソン相関係数は不適切である



強い曲線的な関係は相関では捉えられない。

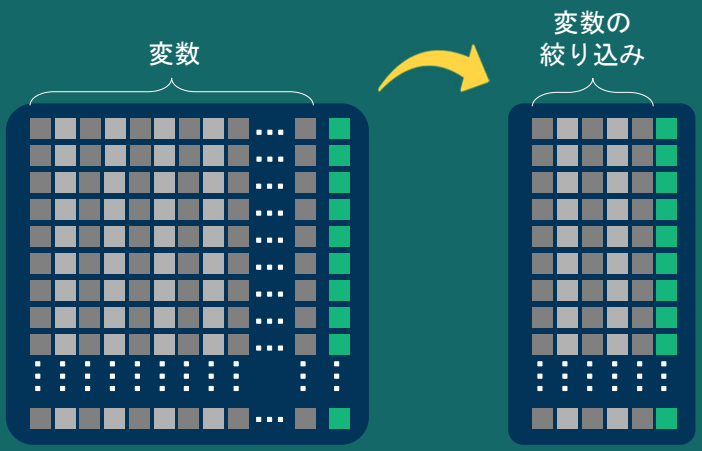
強い周期的関係があるにもかかわらず、相関係数はほぼゼロになる。

1つの異常な観測値が存在するだけで、強い相関が生じることがある。

☛ 相関は線形関係のみを測定する。

相関を用いた変数の選定

機械学習モデル向けのデータ準備

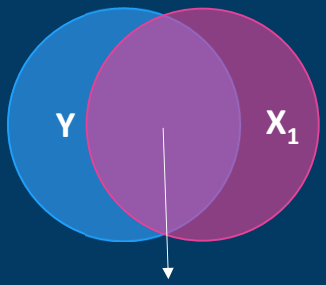


変数スクリーニングは、
無関係な予測変数を特定します。

変数スクリーニングは、
冗長な予測変数を特定します。

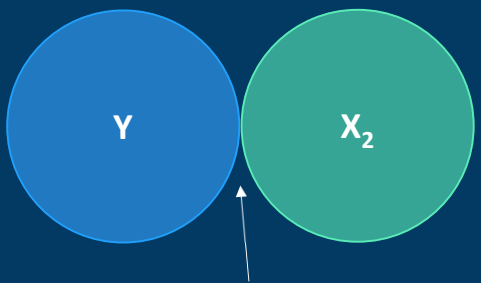
関連性のある予測因子と関連性のない予測因子

関連性
（応答変数と予測変数との間に強い相関がある）



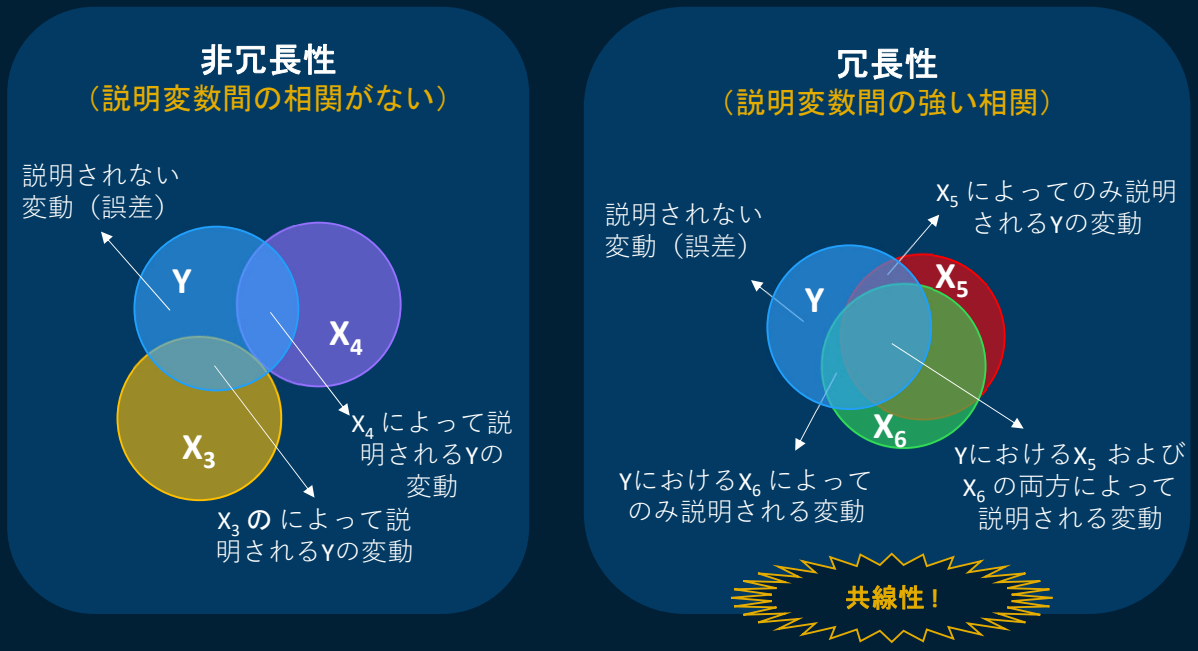
YにおいてX₁によって説明される変動がある

無関係性
（応答変数と予測変数との間に相関がない）

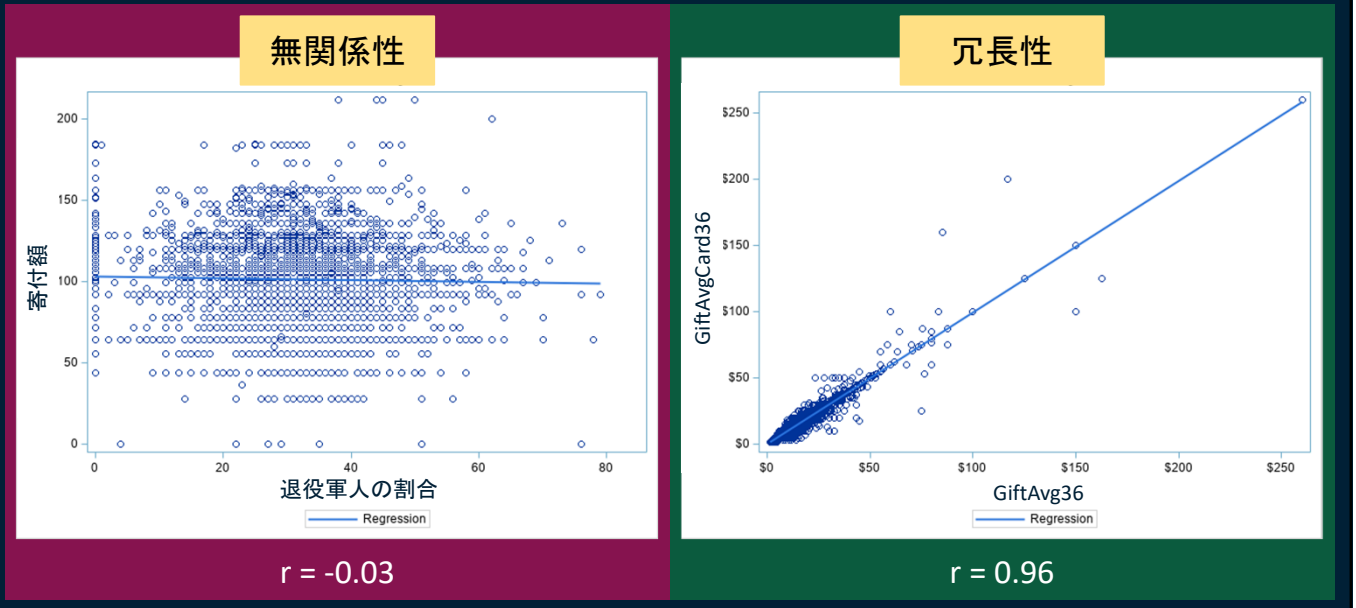


YにおいてX₂によって説明される変動がない

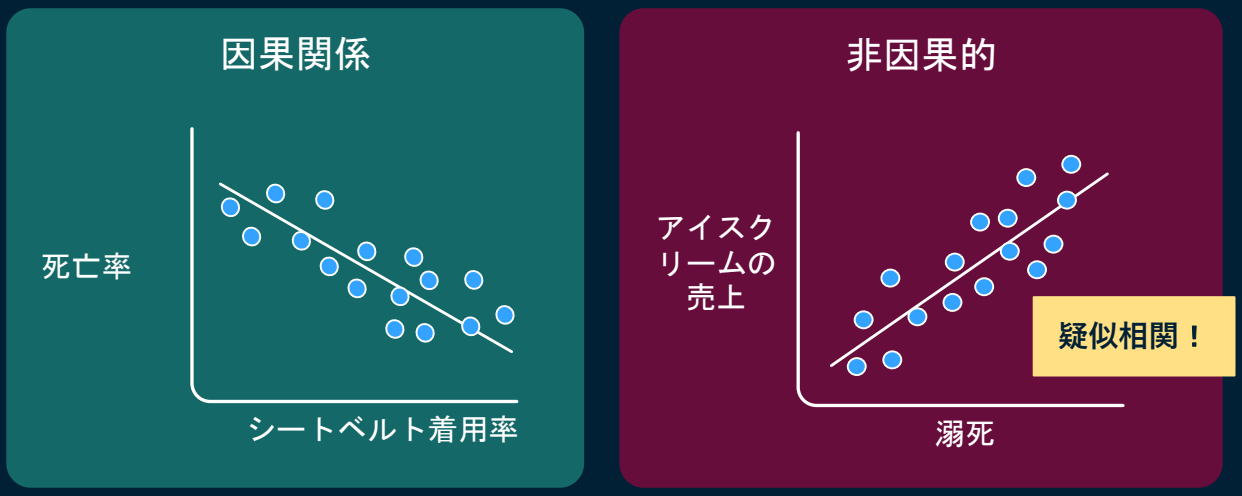
冗長な予測変数と非冗長な予測変数



無関係な情報や冗長な情報の例



相関関係は因果関係を意味しない



問題：ピアソン相関係数

次のうち、正しい記述はどれか。

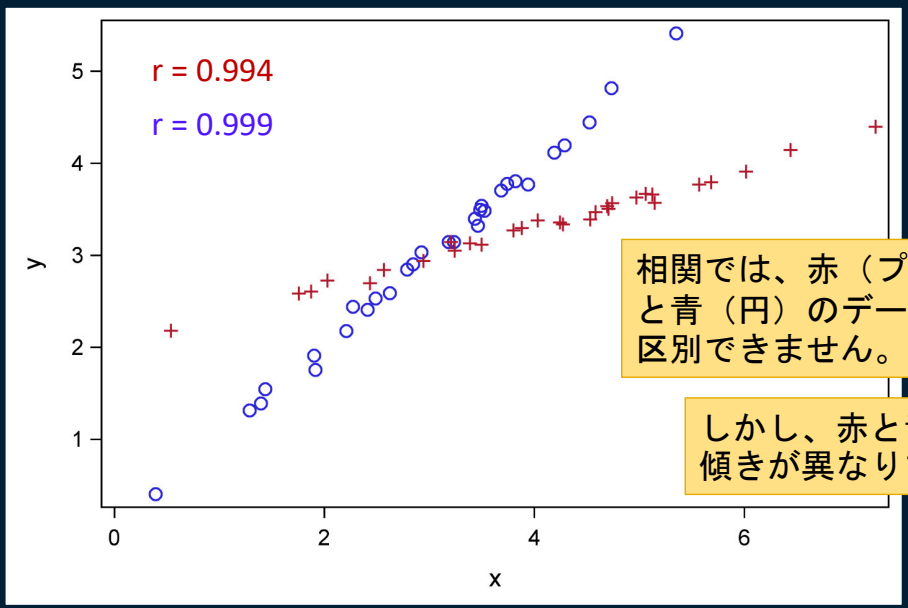
a. 負のピアソン相関は、線形関係の強さが弱いことを示す。
b. ピアソン相関のp値は、関係の強さを測定することができる。
c. データの散布図がランダムな分布の場合、負の相関が示唆される。
d. ピアソン相関係数は、線形関係の度合いを測る指標である。

回答：ピアソン相関係数

次のうち、正しい記述はどれか。

a. 負のピアソン相関は、線形関係の強さが弱いことを示す。
b. ピアソン相関のp値は、関係の強さを測定することができる。
c. データの散布図がランダムな分布の場合、負の相関が示唆される。
d. ピアソン相関係数は、線形関係の度合いを測る指標である。

相関と回帰

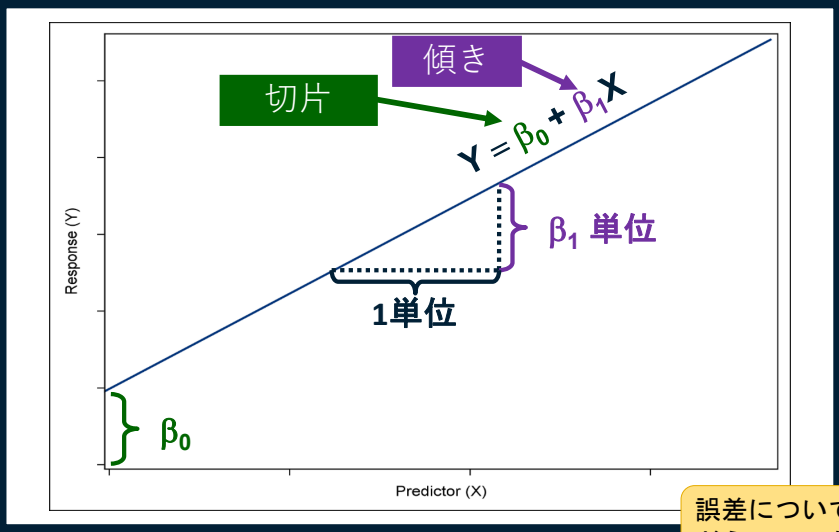


相関では、赤（プラス記号）と青（円）のデータセットを区別できません。

しかし、赤と青の直線は傾きが異なります。

「回帰」という用語の歴史

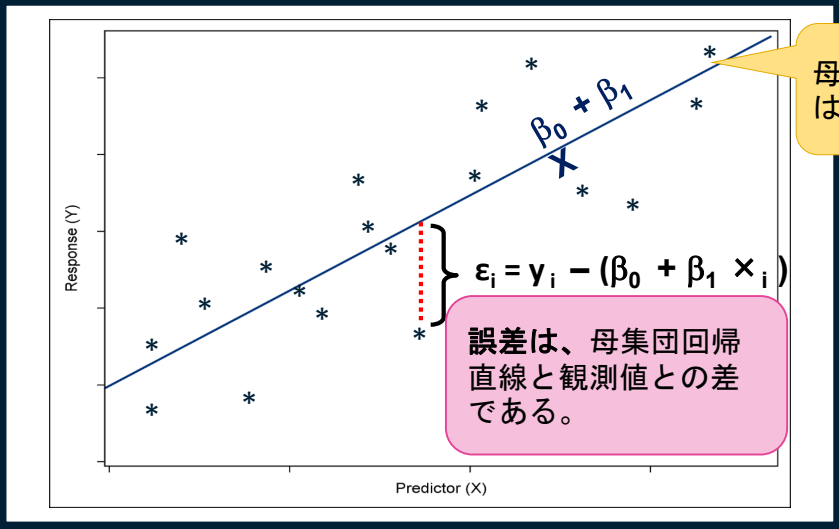
単純線形回帰



誤差についてはどうでしょうか？

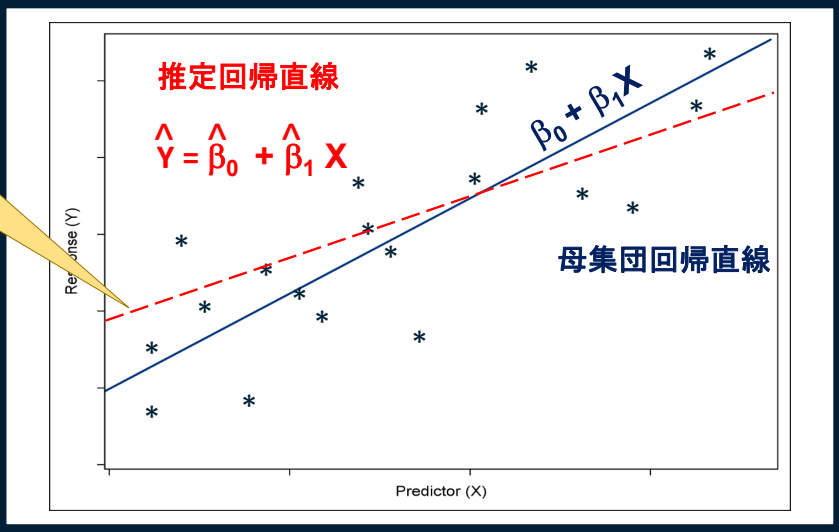
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

単純線形回帰



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

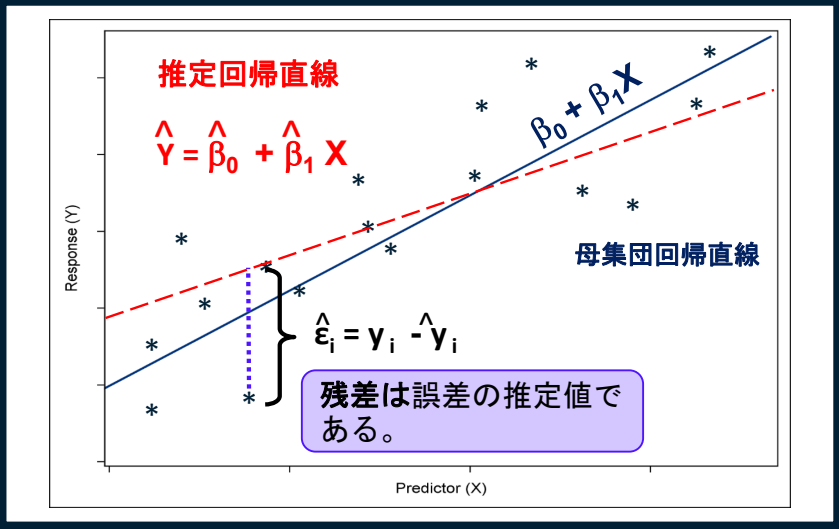
単純線形回帰



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

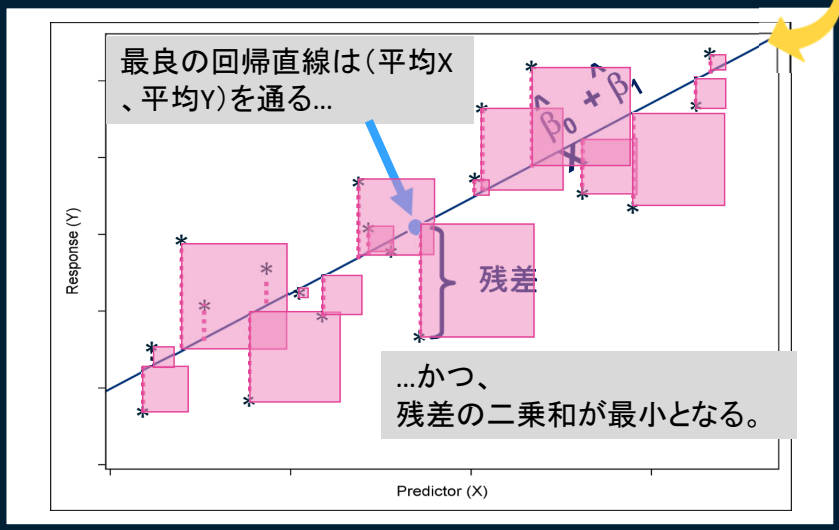
単純線形回帰

回帰の推定と評価には、残差が関わります。



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

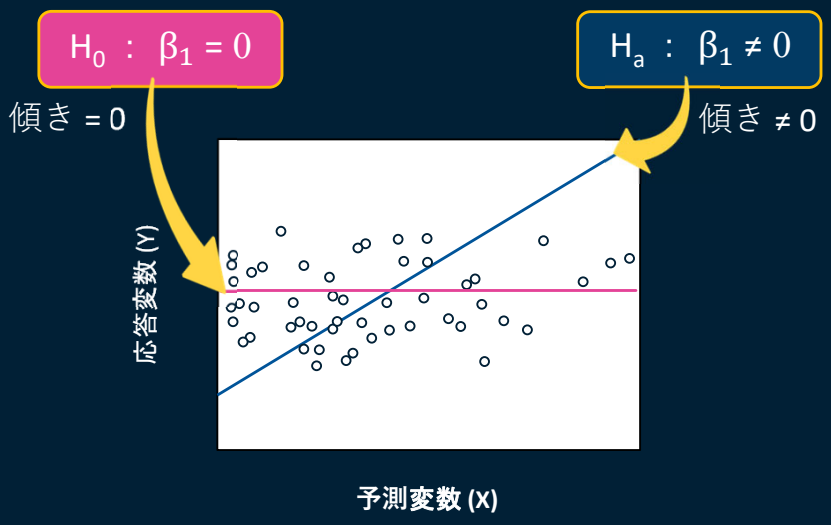
最小二乗法



X +

OLS回帰のパラメータ推定値

線形回帰の仮説検定



問題：単純線形回帰

単純線形回帰において、X1のパラメータ推定値（傾き）が0の場合、X1 = 15のときのYの最良の推定値（予測値）は、次のうちどれか。

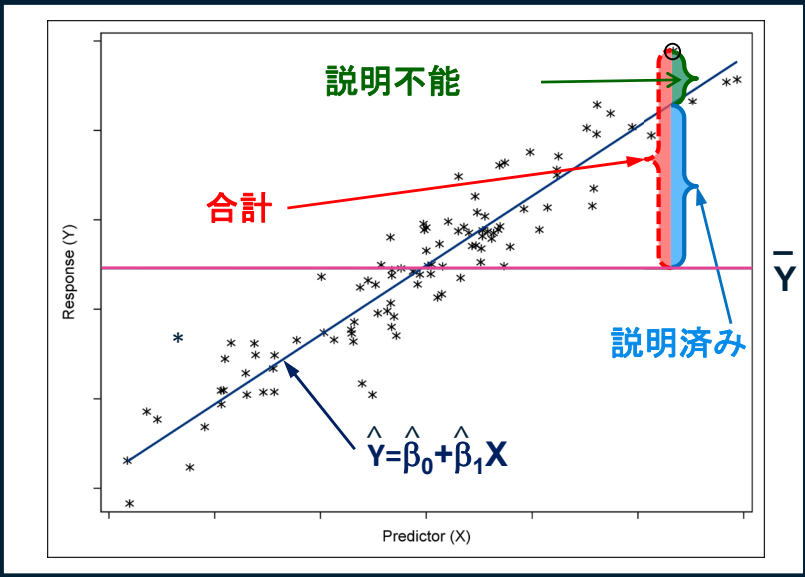
a. 15
b. Yの平均
c. ランダムな数値
d. X1の平均
e. 0

回答：単純線形回帰

単純線形回帰において、X1のパラメータ推定値（傾き）が0の場合、X1 = 15のときのYの最良の推定値（予測値）は、次のうちどれか。

a. 15
b. Yの平均
c. 乱数
d. X1の平均
e. 0

説明可能な変動と説明不可能な変動



モデルによって説明される変動の割合は全体の何%か？

決定係数

私の回帰モデルはデータにどの程度適合しているか？

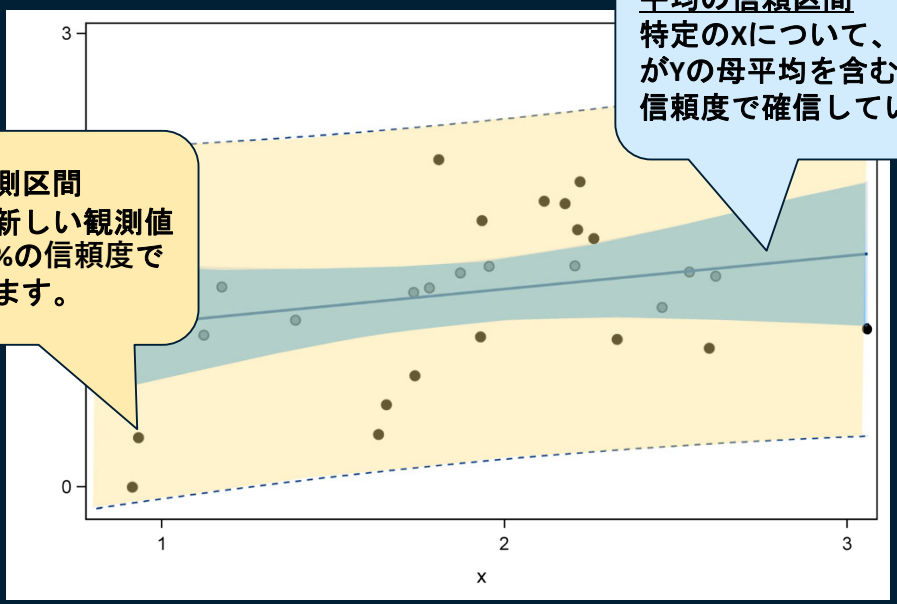
$$R^2 = \frac{\text{説明される変動}}{\text{全ての変動}}$$

「モデルによって説明できるYの変動の割合」

$$0 \leq R^2 \leq 1$$

最小二乗回帰について

信頼区間と予測区間



観測値の予測区間
その区間が新しい観測値を含むと95%の信頼度で確信しています。

平均の信頼区間
特定のXについて、その区間がYの母平均を含むと95%の信頼度で確信しています。

問題：変数間の関係

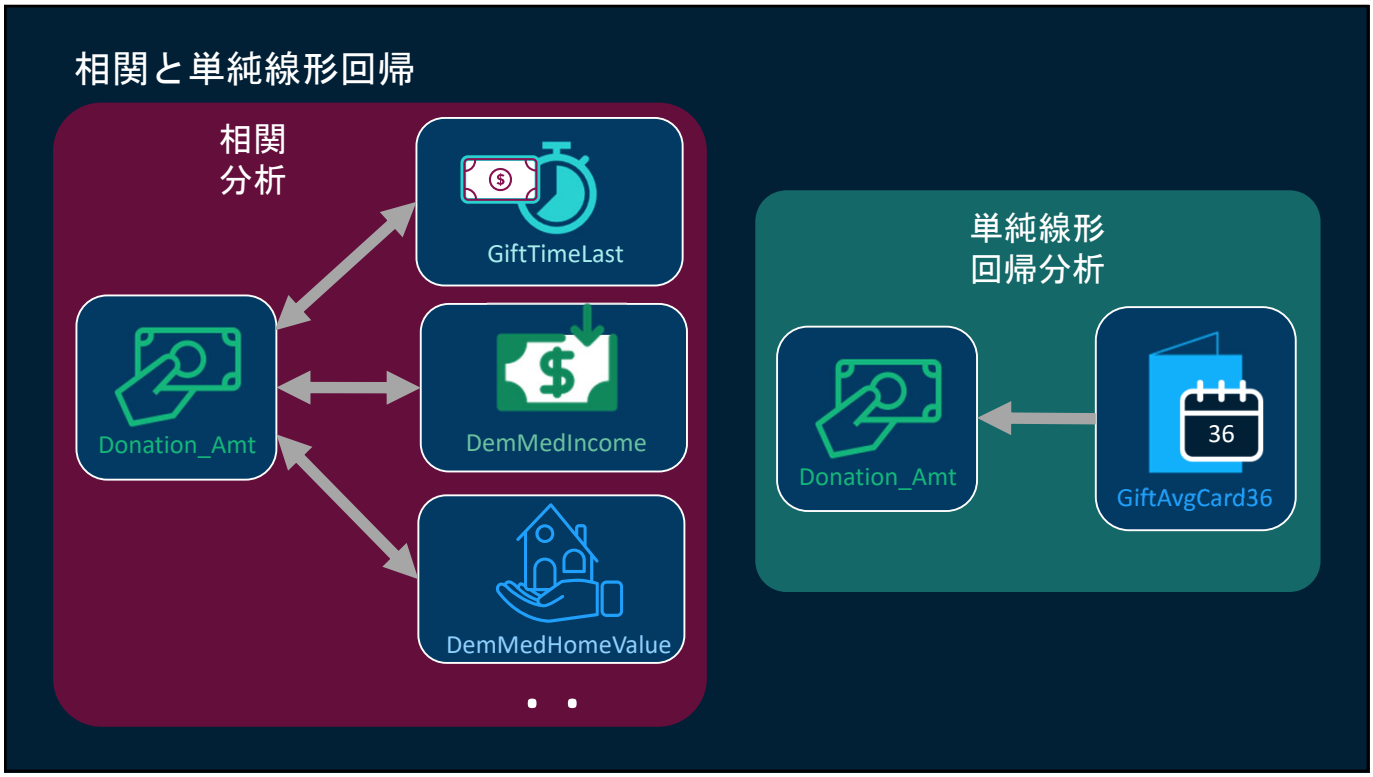
次のうち、正しい記述はどれか。

a. 相関係数が-1に近い場合は、関連性がほとんどないことを示している。
b. 散布図は相関係数の解釈に役立たない。
c. 決定係数（R^2）は、-1 から +1 の間の値をとる。
d. 誤差は母集団のパラメータであり、残差は誤差の推定値です。

回答：変数間の関係

次のうち、正しい記述はどれか。

a. 相関係数が-1に近い場合は、関連性がほとんどないことを示している。
b. 散布図は相関係数の解釈に役立たない。
c. 決定係数（R^2）は、 -1 から +1 の間の値をとる。
d. 誤差は母集団のパラメータであり、残差は誤差の推定値です。



デモ：相関と線形回帰

このデモでは、相関分析と単純線形回帰を用いて関係性を調べる方法について解説します。

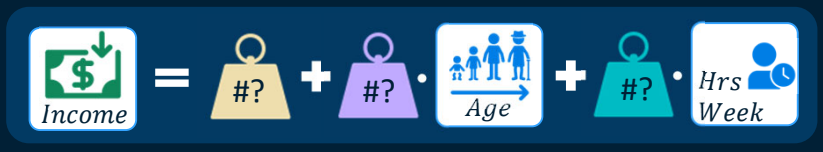
多重回帰とモデル選択

多重回帰とモデル選択手法の使い方

Lesson 03, Section 02

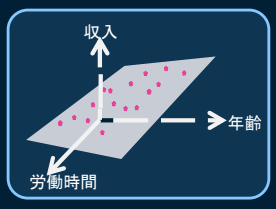
多重回帰

次の例を思い出してください：



線形結合

名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

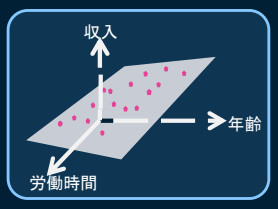


多重回帰

次の例を思い出してください：

$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Age + \hat{\beta}_2 \cdot HrsWeek$$

名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

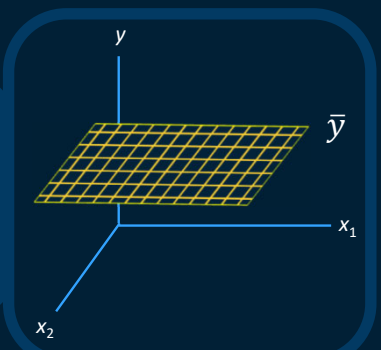


$\hat{\beta}_1$ = 労働時間が一定の場合の、年齢が収入に与える影響
$\hat{\beta}_2$ = 年齢が一定の場合の、労働時間が収入に与える影響

多重回帰の仮説検定

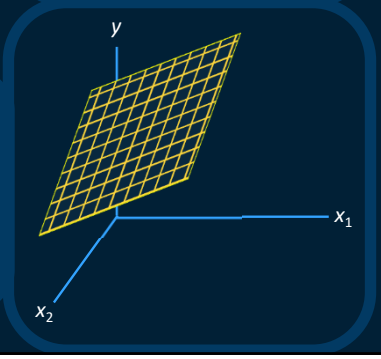
帰無仮説： 説明変数の係数はすべてゼロである。

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$



対立仮説： 少なくとも1つの係数が0ではない。

$$H_a: \text{not}(\beta_1 = \beta_2 = \cdots = \beta_k = 0)$$

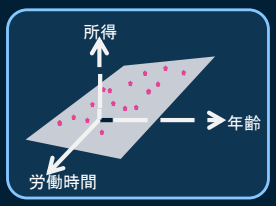


回帰分析におけるカテゴリカルな説明変数

$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Age + \hat{\beta}_2 \cdot HrsWeek$$

名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

これらはどのように数値変数に変換されるのでしょうか？



回帰モデル

カテゴリカルな入力変数のダミー変数化

カテゴリ変数	値	ダミー変数		
		性別F	性別M	性別O
性別	*女性*	1	0	0
	男性	0	1	0
	その他	0	0	1

「女性」ダミー変数

カテゴリカルな入力に対するダミー変数化

カテゴリ変数	値	ダミー変数		
		性別F	性別M	性別O
性別	*女性*	1	0	0
	男性	0	1	0
	その他	0	0	1

「男性」
ダミー変数

カテゴリカルな入力変数のダミー変数化

カテゴリ変数	値	ダミー変数		
		性別F	性別M	性別O
性別	*女性*	1	0	0
	男性	0	1	0
	その他	0	0	1

「その他」
ダミー変数

カテゴリ変数を用いた多重回帰

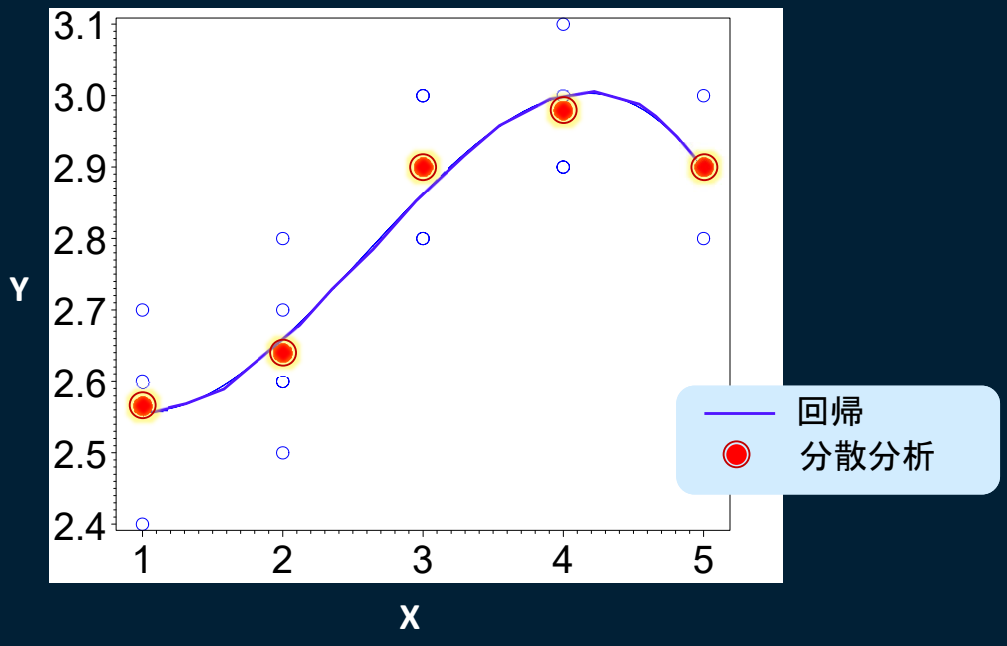
$$\widehat{Income} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Age + \hat{\beta}_2 \cdot HrsWeek + \hat{\beta}_3 \cdot GenderF + \hat{\beta}_4 \cdot GenderM + \hat{\beta}_5 \cdot GenderO$$

性別に関する3つのパラメータ

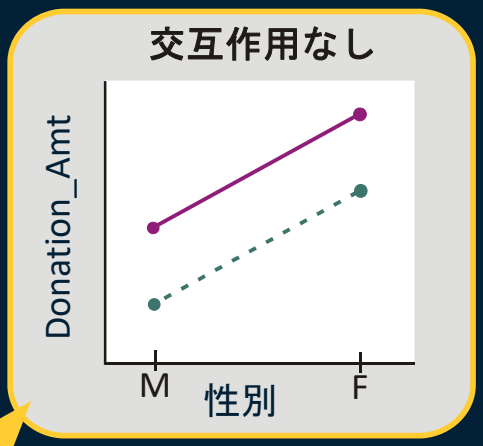
名前	年齢	性別	労働時間	収入
ジョン	24	M	60	241.89
スミス	18	M	30	162.31
エマ	31	F	45	220.04
…	…	…	…	…

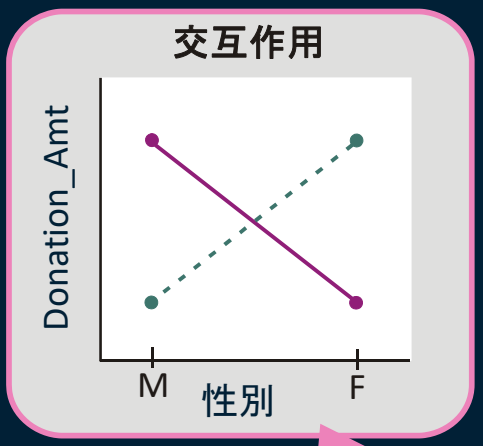
カテゴリカルな説明変数のみを用いた回帰分析は、分散分析（ANOVA）と呼ばれる。

回帰分析と分散分析



相互作用効果





ある予測変数が応答に及ぼす影響は、別の予測変数の値に依存しない。

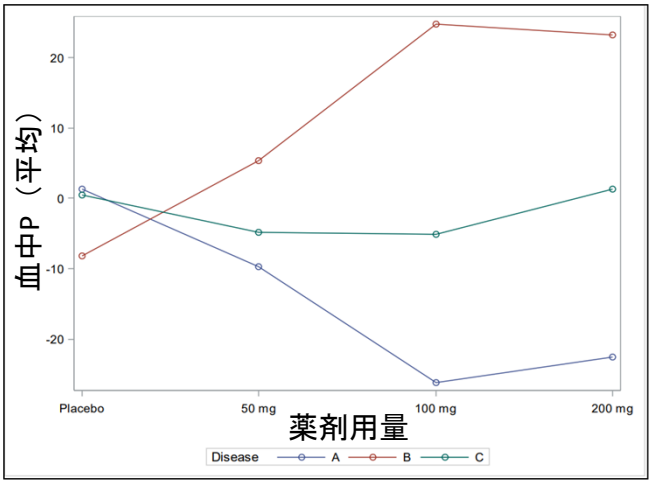
---住宅所有＝はい
—住宅所有=いいえ

ある予測変数が応答に及ぼす影響は、別の予測変数の値によって異なります。

問題：相互作用

この相互作用プロットを調べてください。次の記述のうち、正しいものはどれですか？

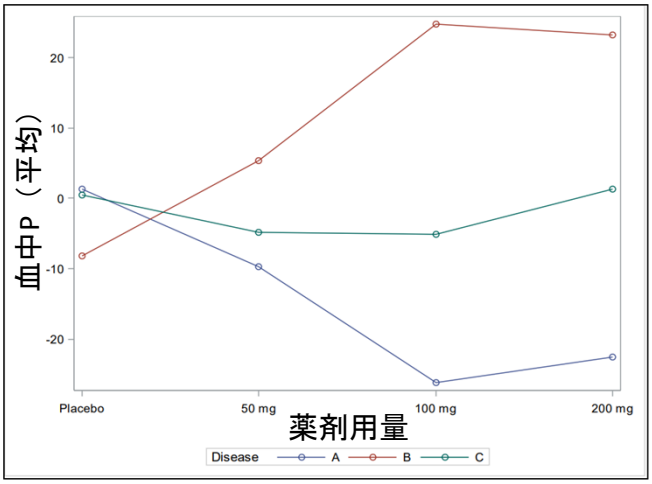
a. 薬の用量は血圧に影響を与えない
   血圧に影響を与えない。
b. 疾患は血圧に影響を与えない
   血圧に影響を与えない。
c. 薬の用量と疾患の間に
   相互作用がない。
d. 薬の用量と疾患の間に
   相互作用が存在する。



回答：相互作用

この相互作用プロットを調べてください。次の記述のうち、正しいものはどれですか？

a. 薬の用量は血圧に影響を与えない
   血圧に影響を与えない。
b. 疾患は血圧に影響を与えない
   血圧に影響を与えない。
c. 薬の用量と疾患の間に
   相互作用がない。
d. 薬の用量と疾患の間に
   相互作用が存在する。

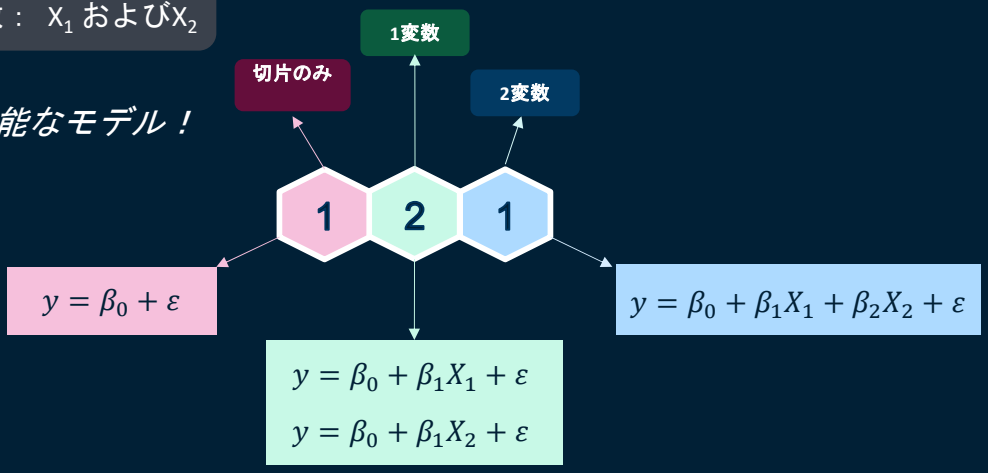


インタラクション

さまざまなモデルがあります！

応答変数： Y
2つの説明変数： X_1 およびX_2

2^2 = 4 つの可能なモデル！



多くのモデルが考えられます！

完全モデルにおける変数の数（k）	サブセットモデルの数 (2^k)
0	1
1	2
2	4
3	8
4	16
5	32

回帰モデルの比較

- R^2
- 調整済みR^2
- 情報基準

調整済みR^2

調整済みR^2 →

$$R^2_{ADJ} = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

i = 切片がある場合は1、ない場合は0
n = モデル推定に使用された観測値の数
p = モデル内のパラメータの数

情報量基準

- モデルの適合度を評価する

情報基準

- 赤池情報量基準（AIC）
- 補正赤池情報量基準（AICC）
- シュワルツのベイズ情報量基準（SBC）、別名：ベイズ情報量基準（BIC）

情報量基準

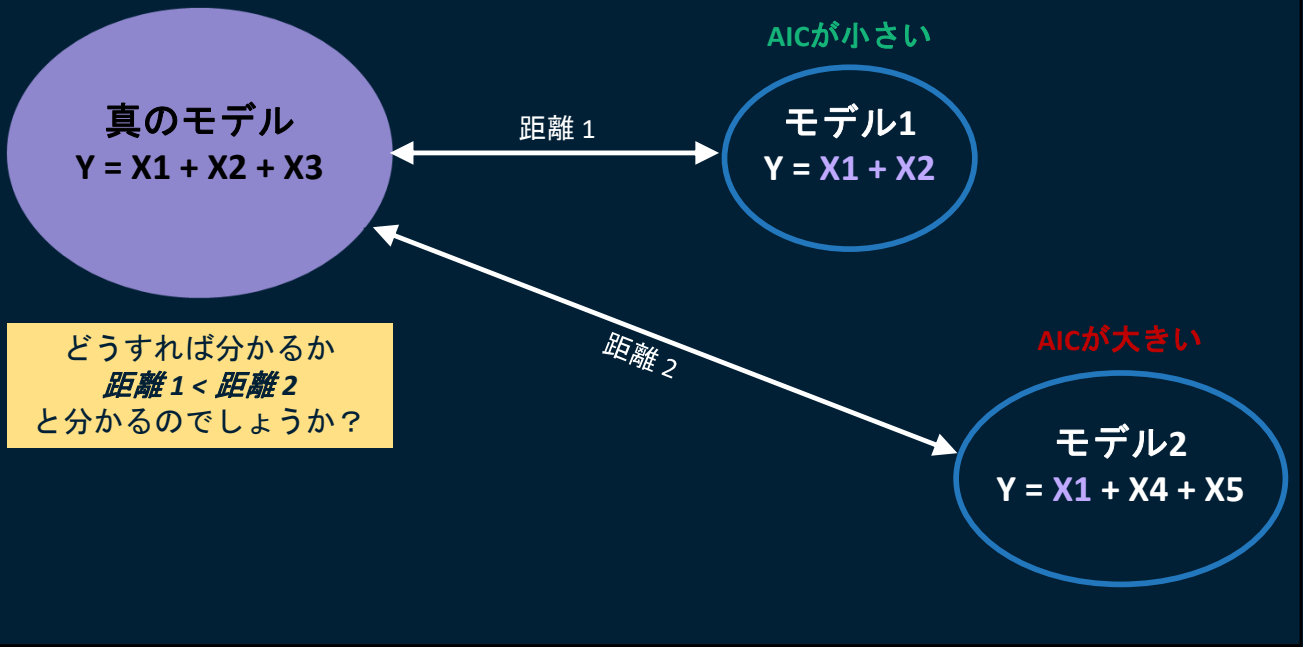
- モデルの適合度を評価する
- SSE、モデル内のパラメータ数、および標本サイズに関する情報を組み合わせる
- 最も単純なモデルを選択するために尤度関数にペナルティを課す
- 同じデータセットに適合させた代替モデルを比較する
- 情報量基準の値が低いモデルは、値が高いモデルよりも優れている

赤池情報量基準（AIC）

真のモデル
Y = X1 + X2 + X3

- *AIC – 候補モデルと真のモデルとの相対的な類似度を測る指標。*
- *類似性の定量化 – 候補モデルと真のモデルとの間の距離。*
- *距離* – 特定のモデルを用いて真のモデルを近似する際に、どれだけの情報が失われるかを示す。

赤池情報量基準（AIC）



赤池情報量基準（AIC）

AICが小さい

モデル 1
Y= X1 + X2

AICが大きい

モデル2
Y= X1 + X4 + X5

真のモデルは
式から外れ、
相対距離 1 < 相対距離 2 となる。

- AICはモデルの適合度を*相対的に*評価する指標で、数値が小さいほど良いとされます。
- 絶対的な意味で、モデルが良いか悪いかを判断することはできない

最もよく使われる情報基準

問題：AIC

モデルについて、

$$\widehat{Income} = 58.1 + 2.35 \cdot Age + 9.02 \cdot HrsWeek$$

について、R²やAICを含む適合度統計量が算出されました。AIC = 660という値から、どのような結論が導き出せますか？

a. AICの値が高いことは、モデルの適合性が良いことを示しています。
b. AICの値が非負とは、モデルの適合度が低いことを示しています。
c. AIC = 660は、R²が高いことを意味します。
d. AICは相対的な指標であるため、このモデルをAICに基づいて評価することはできない。

回答：AIC

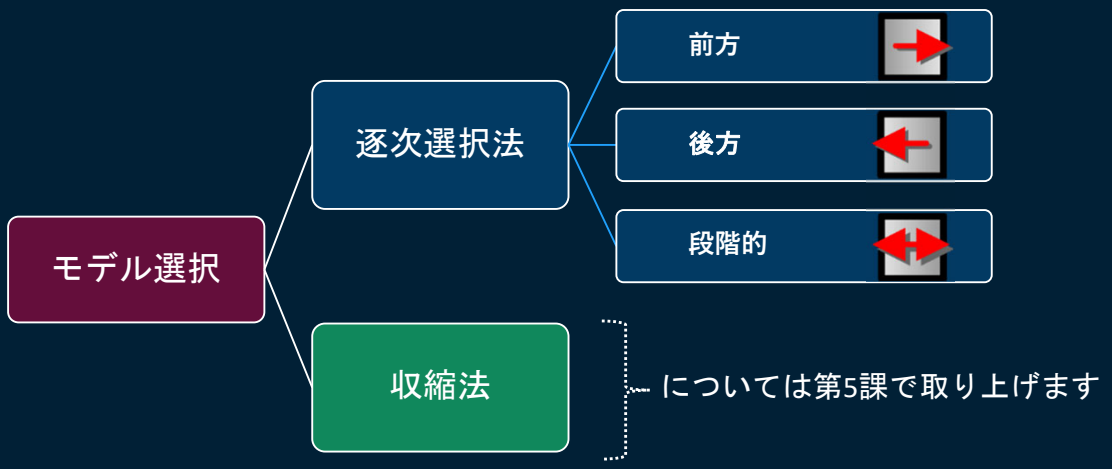
モデルについて適合度統計量が算出された

$$\widehat{Income} = 58.1 + 2.35 \cdot Age + 9.02 \cdot HrsWeek$$

について、R²やAICを含む適合度統計量が算出された。AIC = 660という値から、どのような結論が導き出せるか？

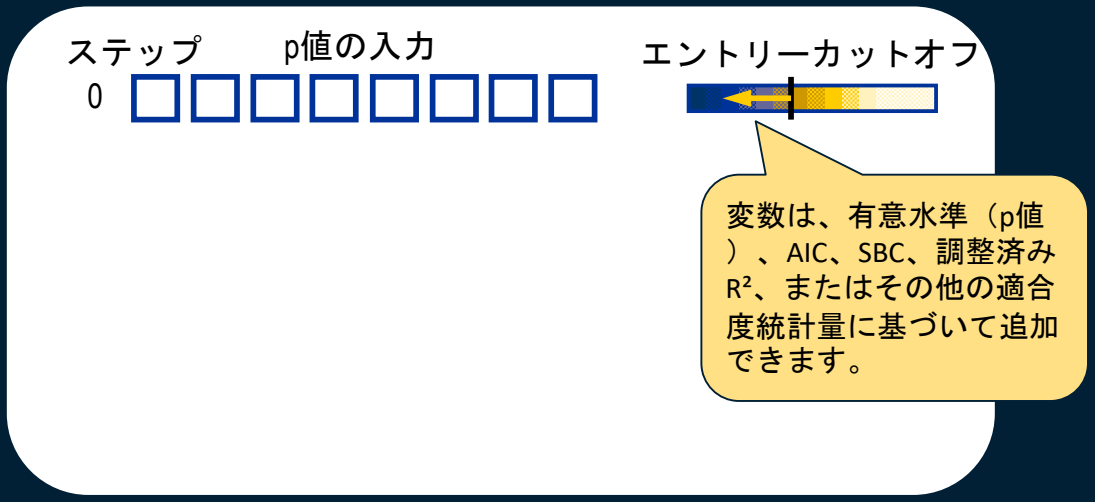
a. AICの値が高いことは、モデルの適合性が良いことを示している。
b. AICの値が非負とは、モデルの適合度が低いことを示しています。
c. AIC = 660は、R²が高いことを意味します。
d. AICは相対的な指標であるため、このモデルをAICに基づいて評価することはできない。

一般的な回帰モデルの選択手法

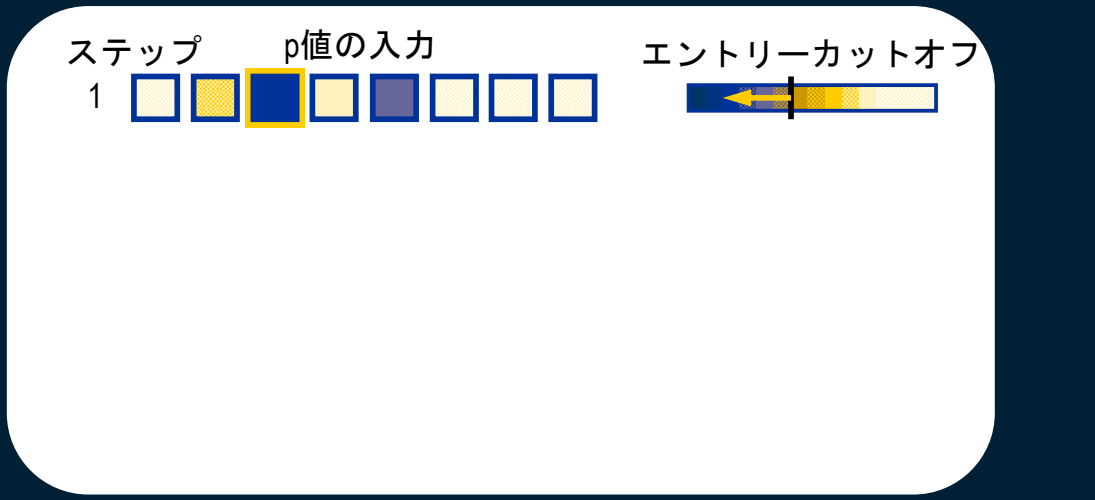


変数の選択／モデルの選択

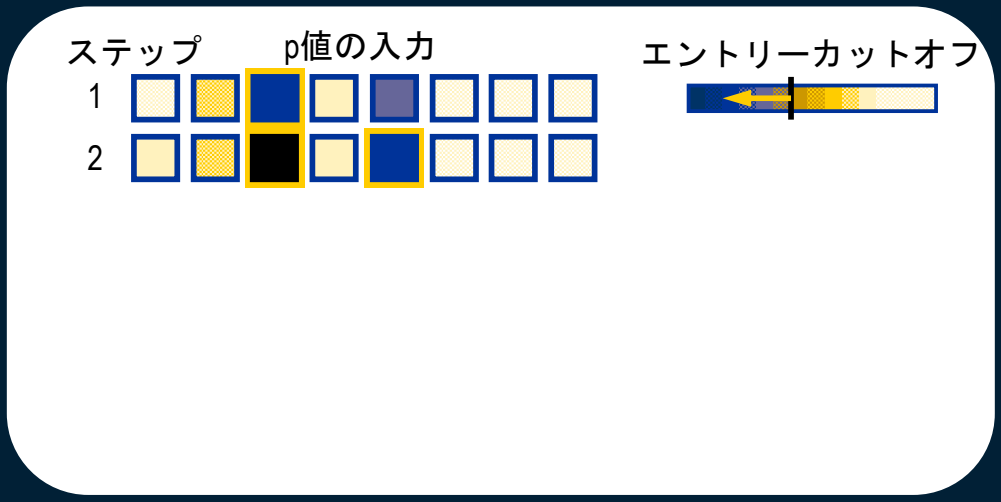
順次選択：前方



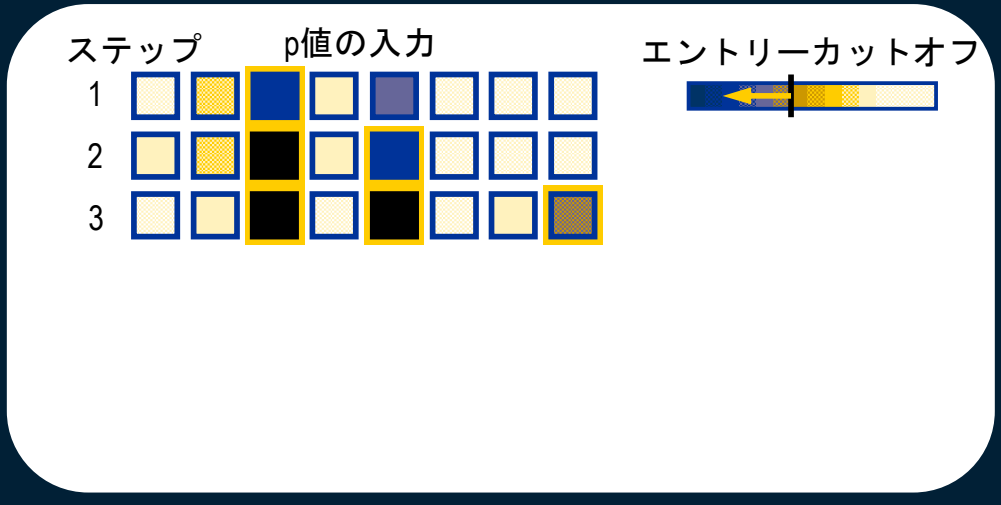
順次選択：前方



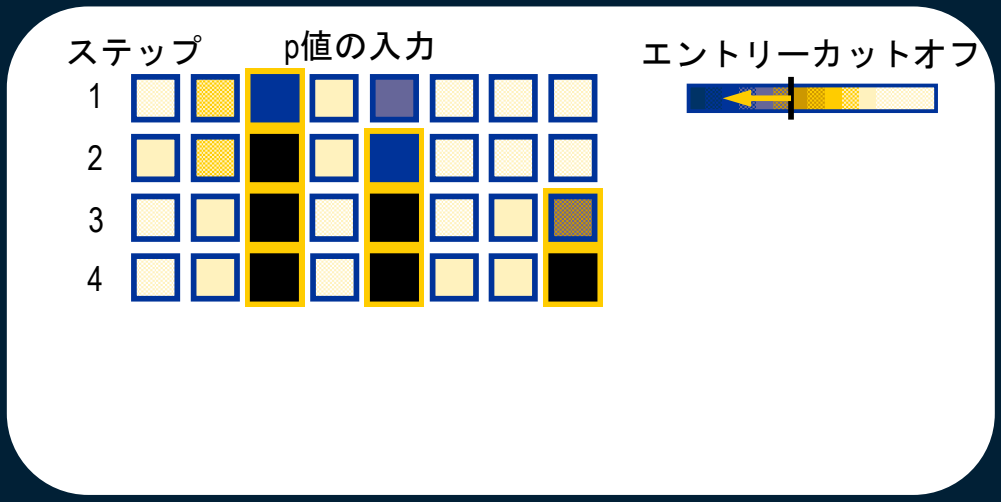
順次選択：前方



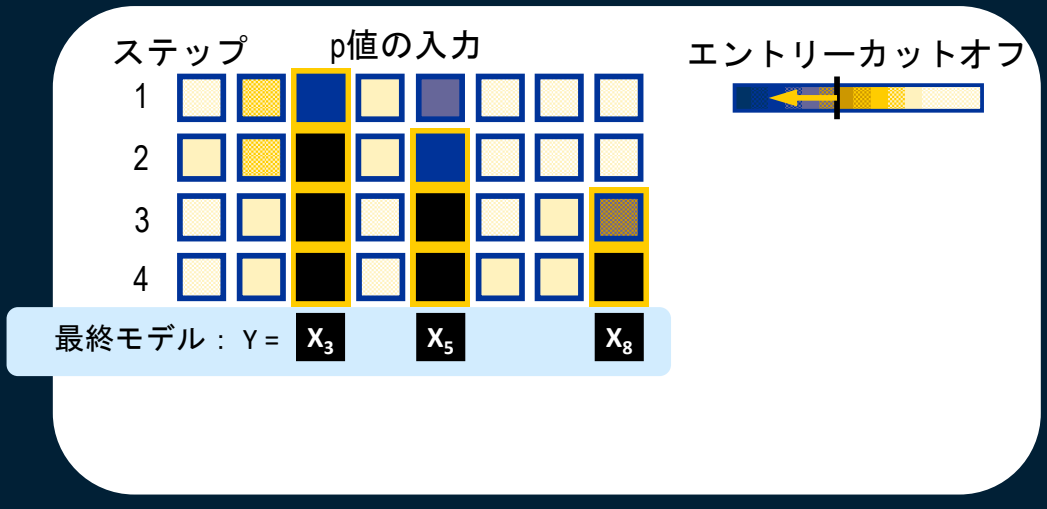
順次選択：前方



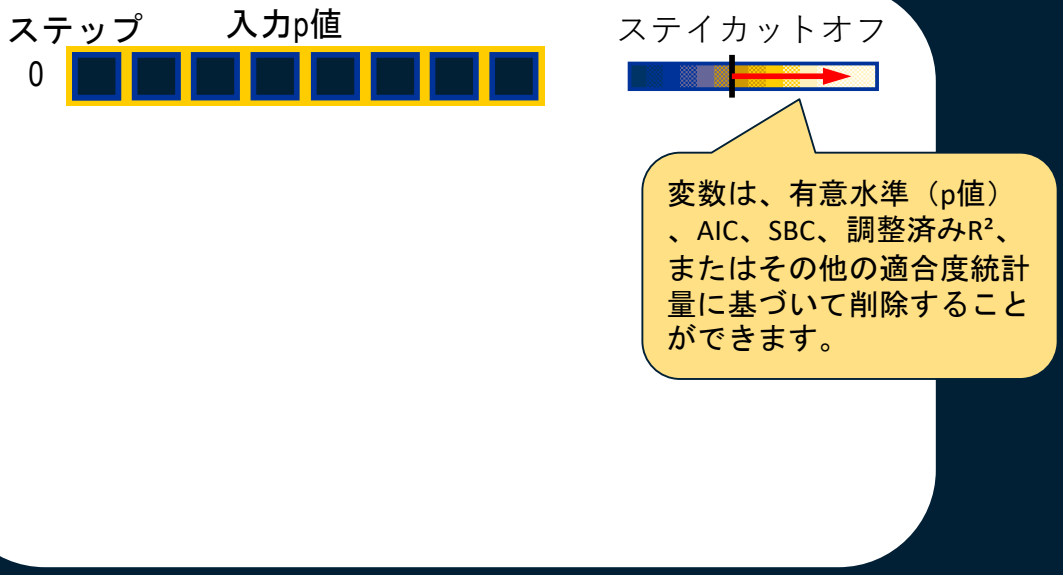
順次選択：前方



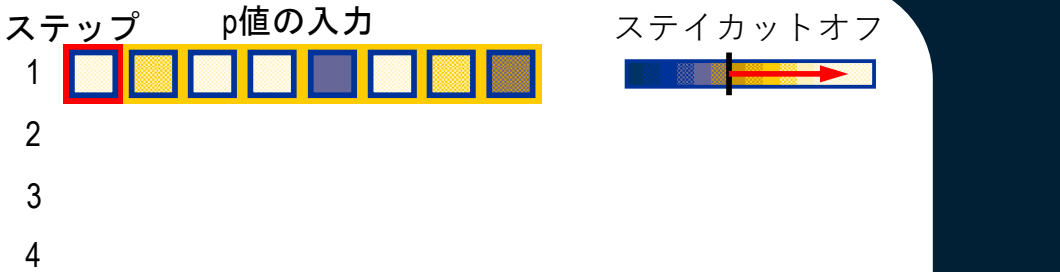
順次選択： 前方



順次選択：後方



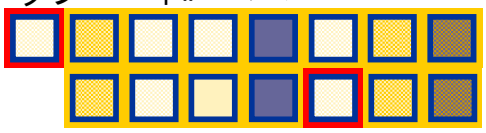
順次選択：後方



順次選択：後方

ステップ　　p値の入力　　ステイカットオフ

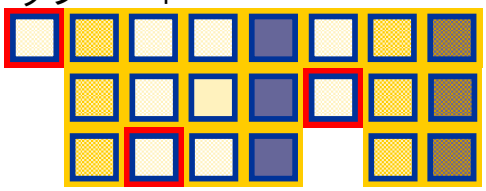
1
2
3
4



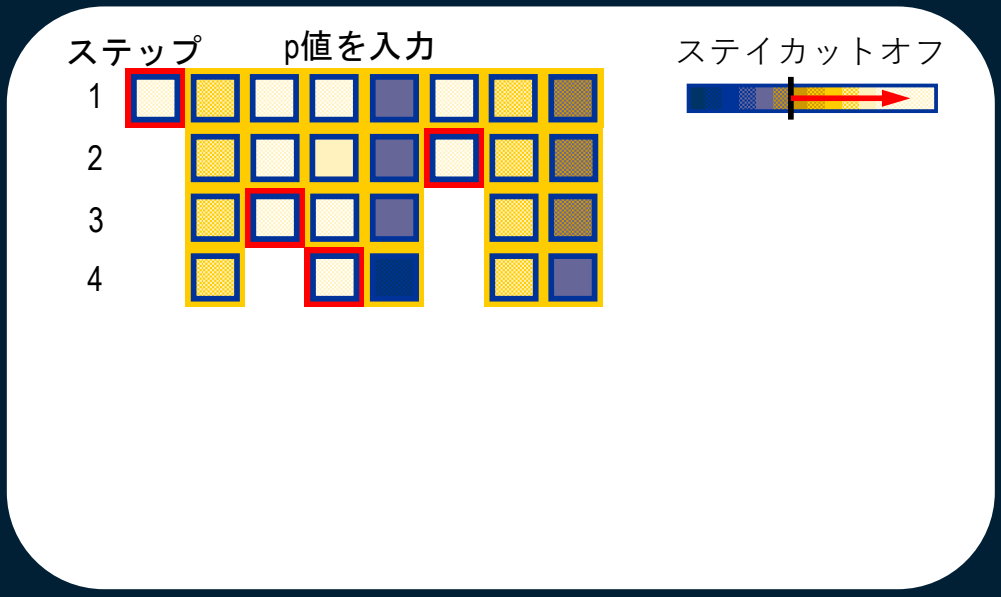
順次選択：後方

ステップ　　p値の入力　　ステイカットオフ

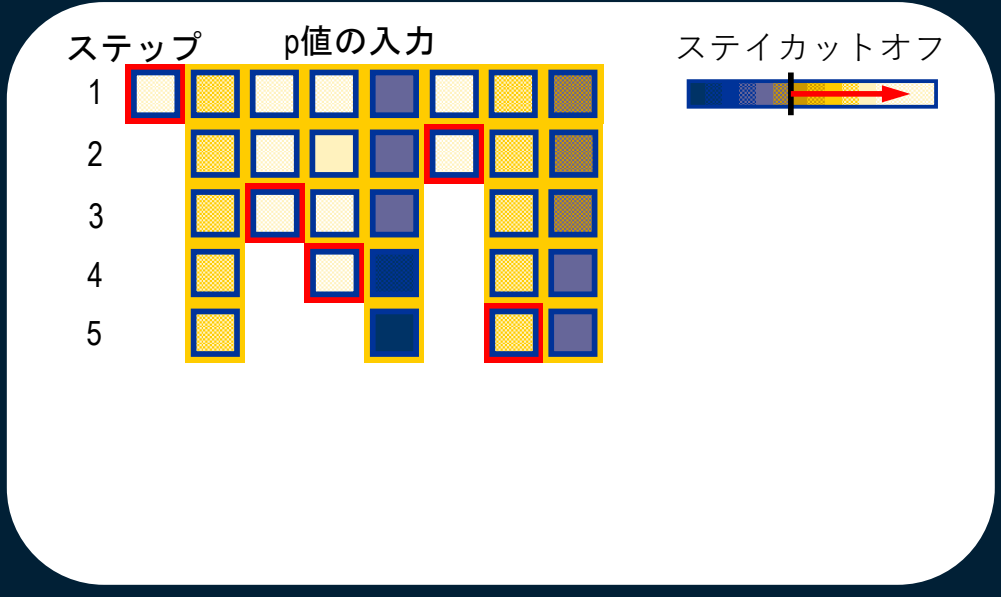
1
2
3
4



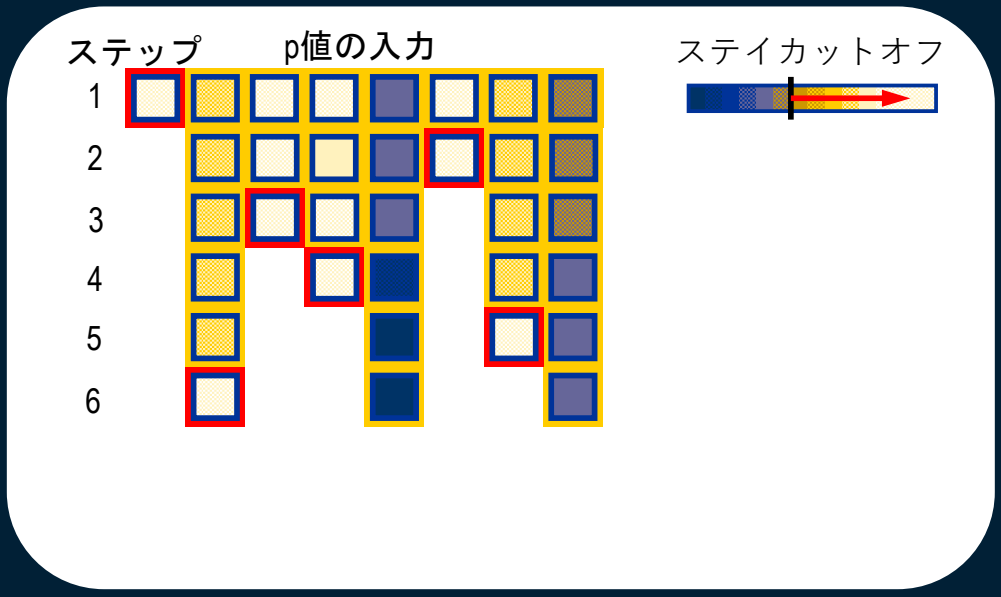
順次選択：後方



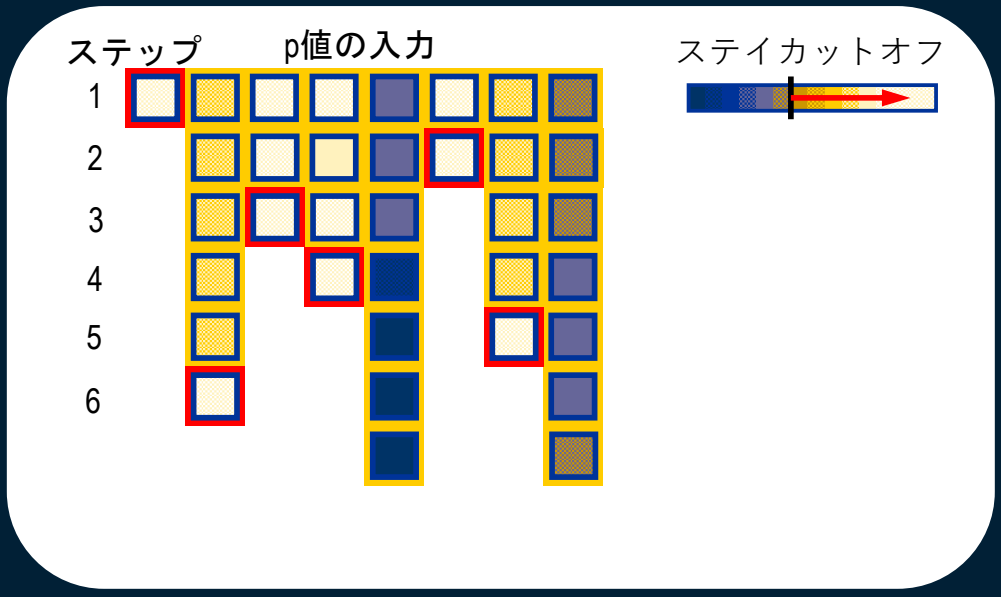
順次選択：後方



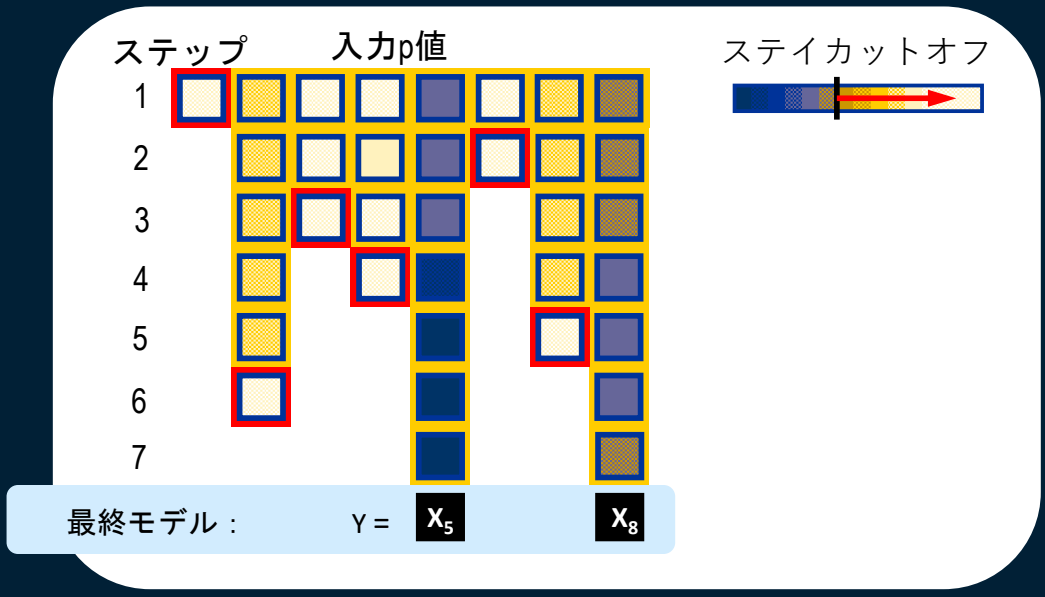
順次選択：後方



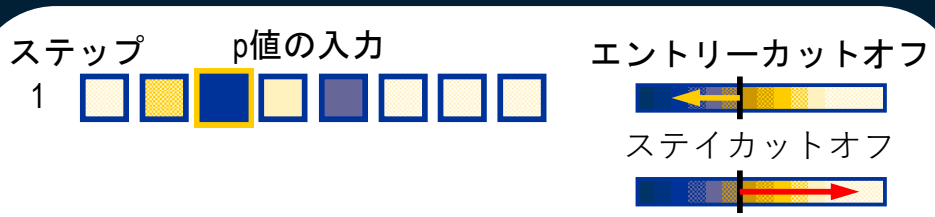
順次選択：後方



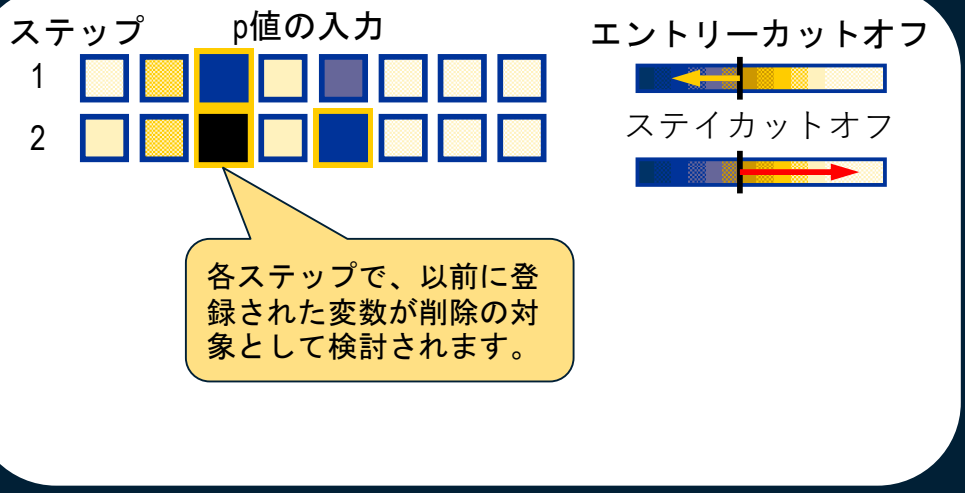
順次選択：後方



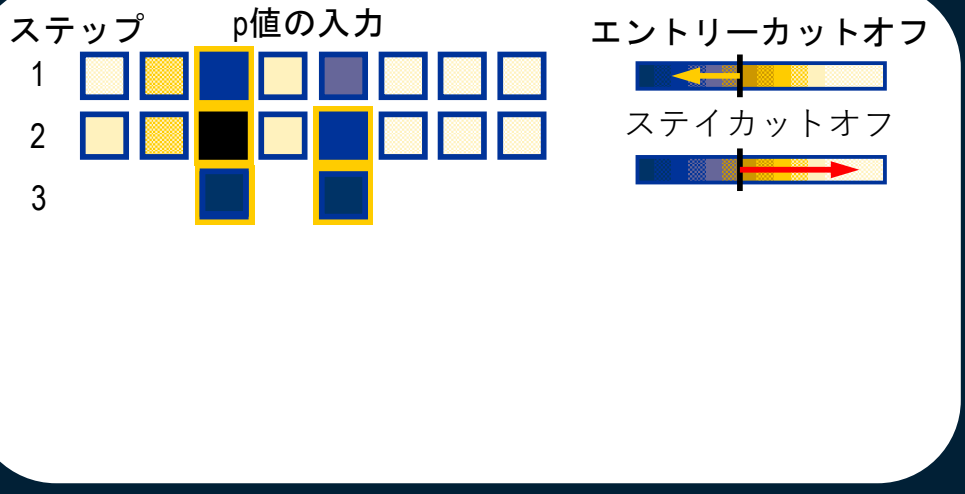
順次選択： 段階的（ステップワイズ法）



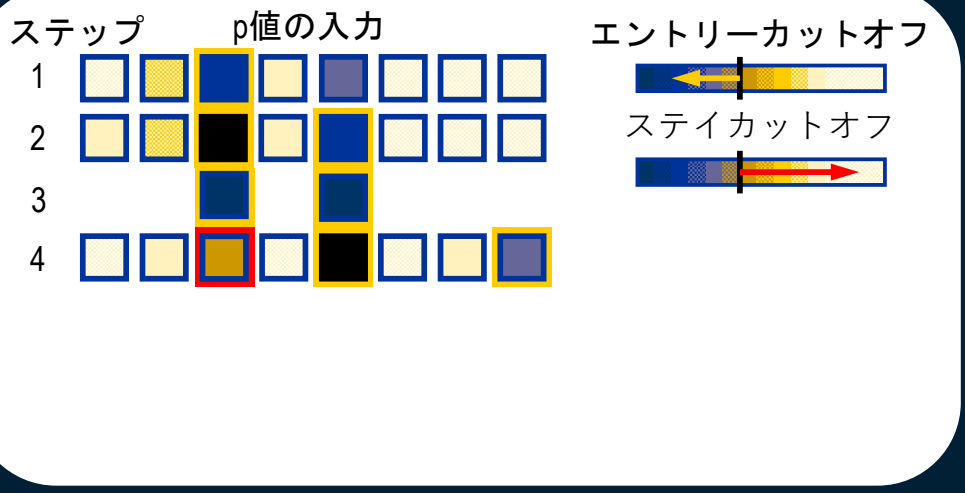
順次選択： 段階的（ステップワイズ法）



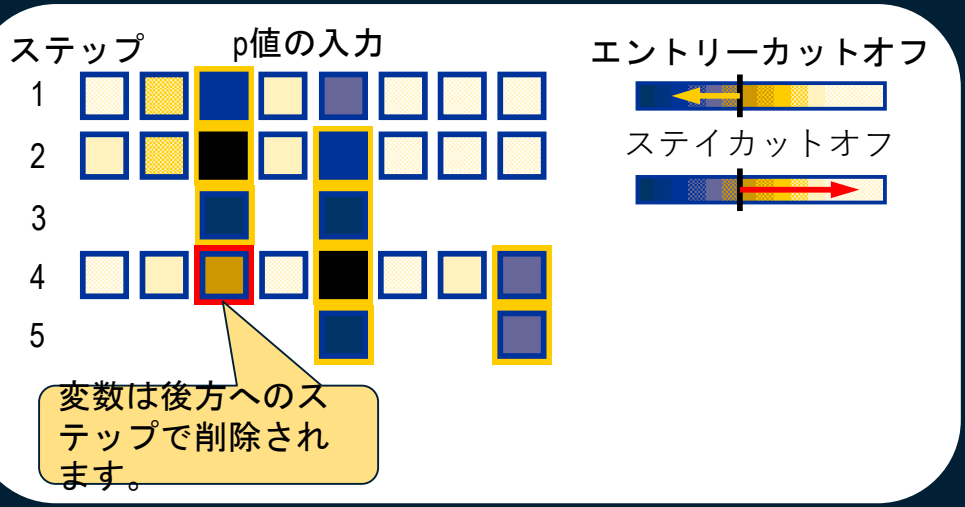
順次選択： 段階的（ステップワイズ法）



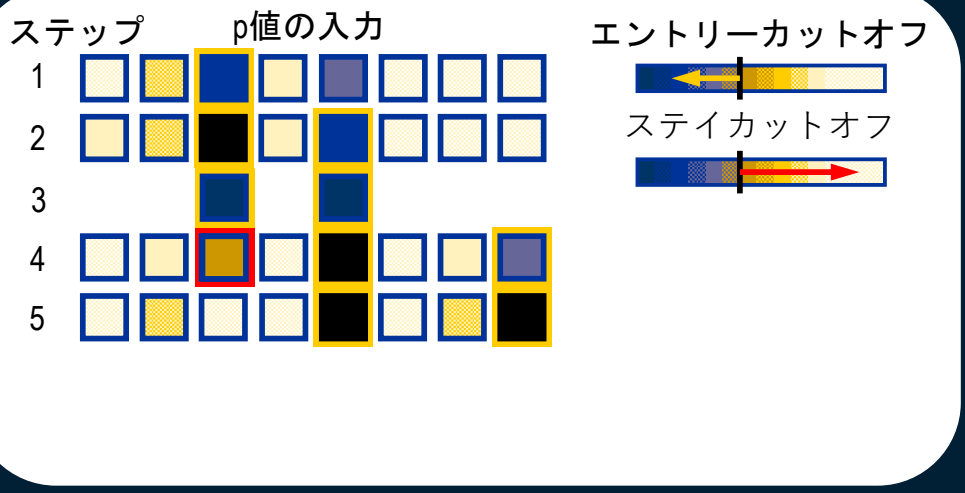
順次選択： 段階的（ステップワイズ法）



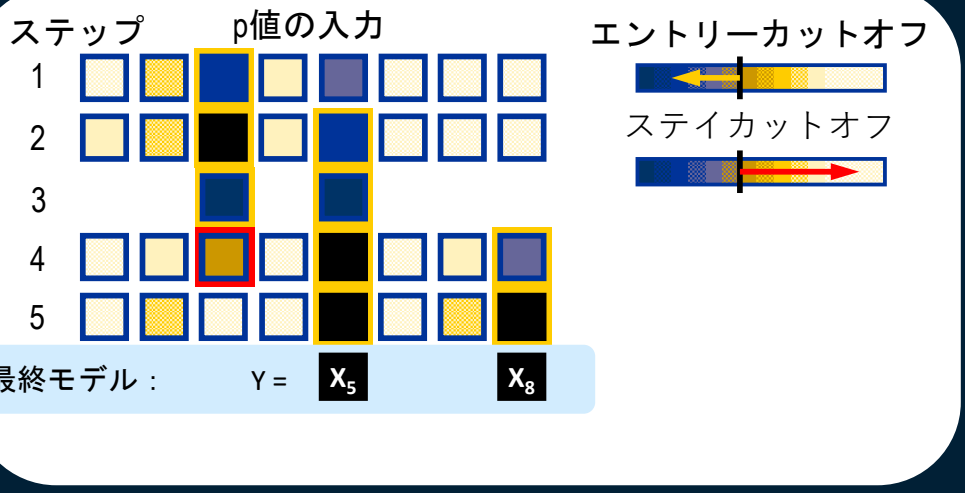
順次選択： 段階的（ステップワイズ法）



順次選択： 段階的（ステップワイズ法）



逐次選択： 段階的（ステップワイズ法）



問題：逐次選択法

STEPWISE、BACKWARD、およびFORWARDの各戦略は、これら3つすべてで同じ有意水準が使用された場合、同じ最終モデルが得られる。

a. 正しい
b. 誤り

回答：逐次選択法

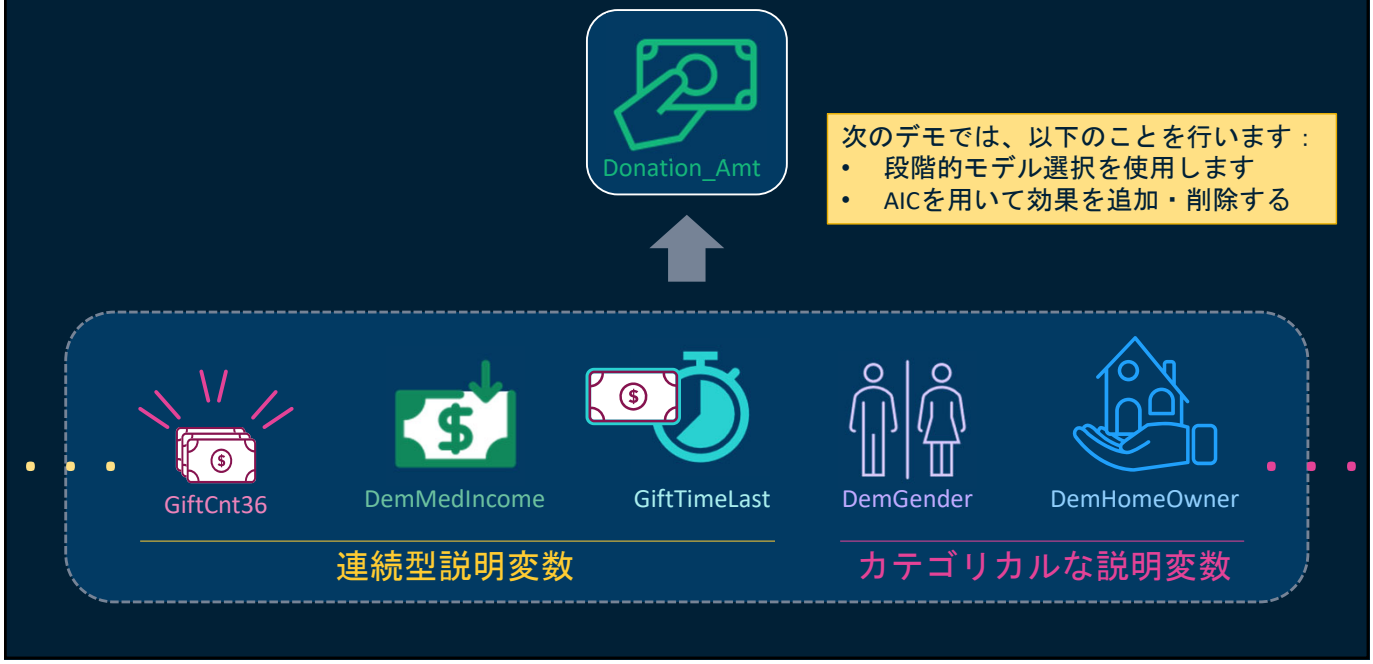
STEPWISE、BACKWARD、およびFORWARDの各戦略は、これら3つすべてで同じ有意水準が使用された場合、同じ最終モデルが得られる。

a. 正しい
b. 誤り

段階的選択法の問題点

収縮方法

多重回帰とモデル選択



デモ：多重回帰とモデル選択

このデモでは、多重回帰における段階的モデル選択法の使用方法を解説します

候補モデル

どのモデルを使用すべきか？

モデル診断

モデル診断の理解

Lesson 03, Section 03

モデルの診断

- 回帰分析の仮定
- 共線性
- 高レバレッジ観測値

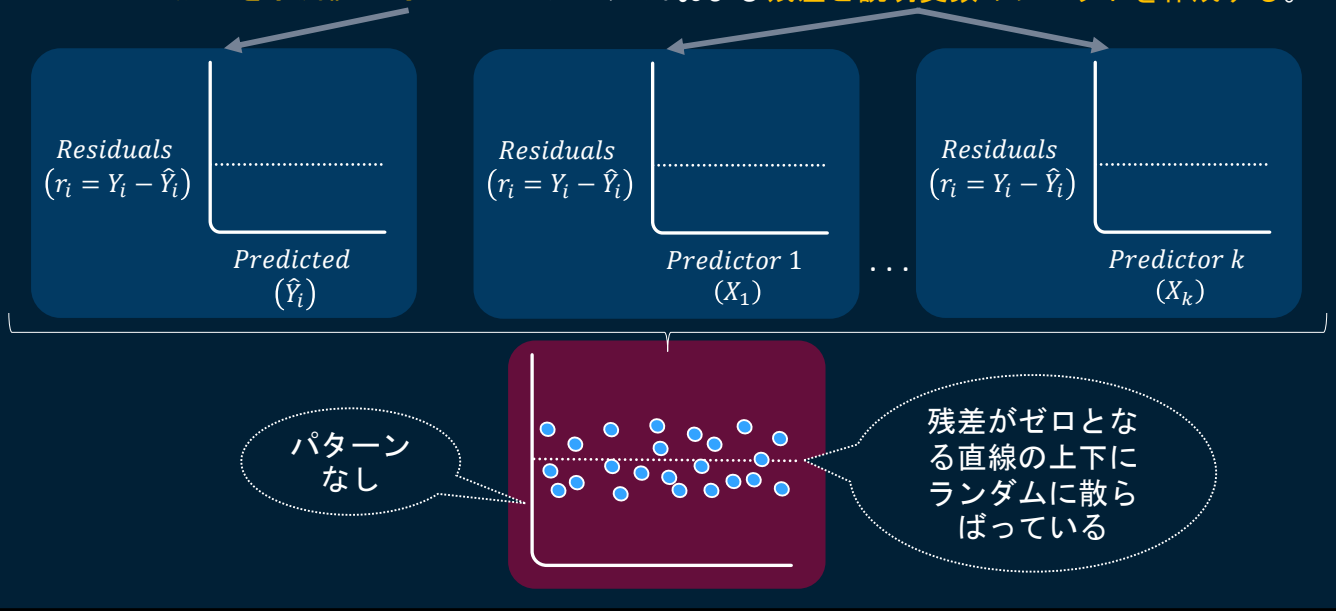
線形回帰の仮定

パラメータが線形であること	*（線形性）*
誤差の独立性	*（自己相関なし）*
誤差の正規分布	*（正規性）*
誤差の分散が等しい	*（等分散性）*

Linear in the parameters	*(Linearity)*
Independent errors	*(No Autocorrelation)*
Normally distributed errors	*(Normality)*
Equal error variance	*(Homoscedasticity)*

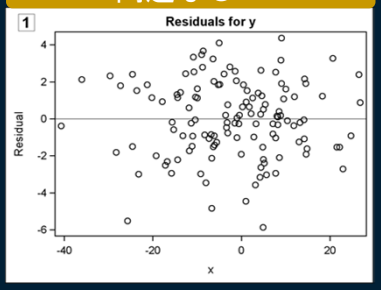
残差プロットを用いた仮定の検証

残差を予測値に対してプロットする。および残差と説明変数のプロットを作成する。

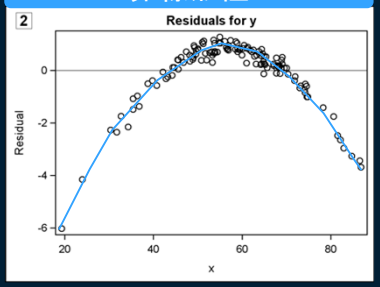


不規則なパターンは問題の兆候

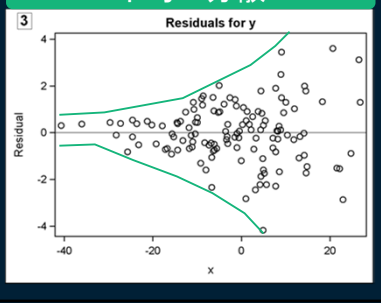
問題なし！



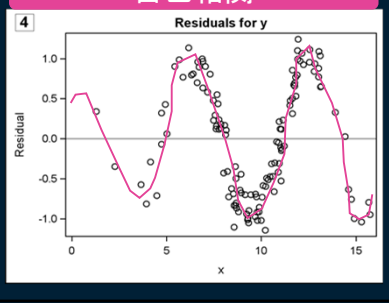
非線形性



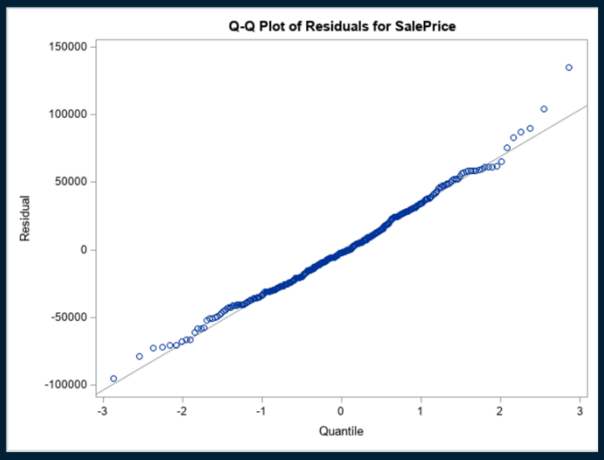
不均一分散



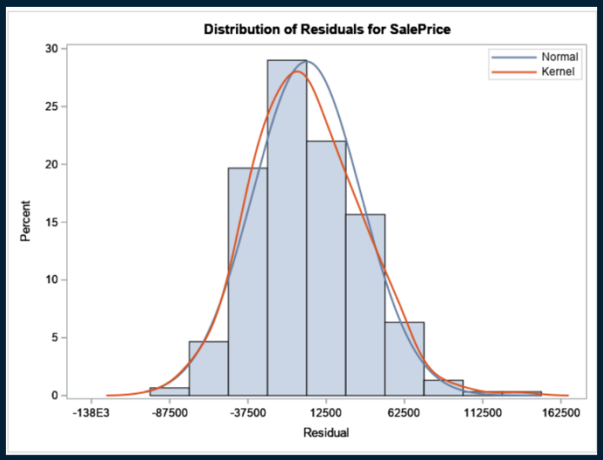
自己相関



他のプロットを用いて正規性を確認する



正規分布の
クアンタイル・クアンタイル図



残差のヒストグラム

問題：回帰分析の仮定

線形回帰モデルでは、説明変数は正規分布に従うと仮定されます。

a. 正しい
b. 誤り

回答：回帰分析の仮定

線形回帰モデルでは、説明変数は正規分布に従うと仮定される。

a. 正しい
b. 誤り

潜在的な問題：共線性

応答変数は関与しない

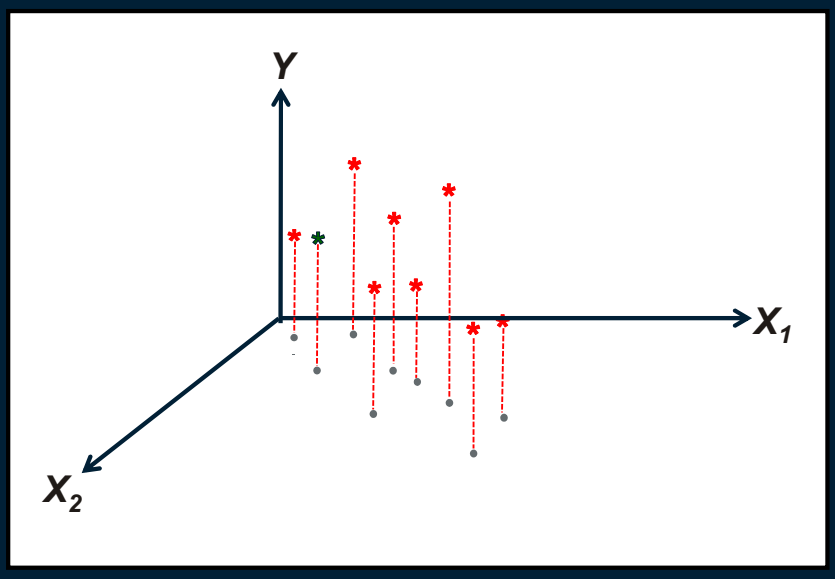
共線性：
複数の説明変数間で強い線形相関

モデルの仮定に違反するものではありませんが、それでも問題となります！

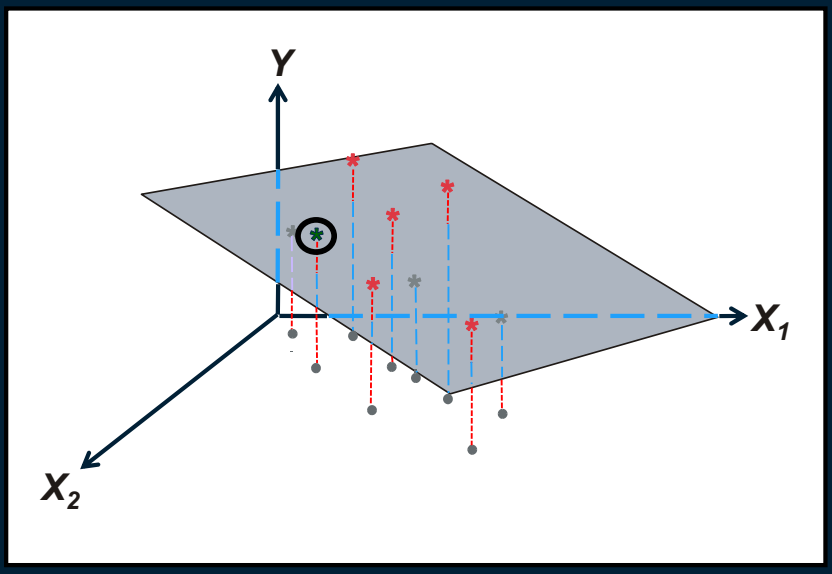
影響：

- パラメータ推定値の分散を過大評価する
- 予測値の分散が過大になる
- パラメータが不安定になる
- モデルの解釈を困難にする
- 自動変数選択の精度を低下させる

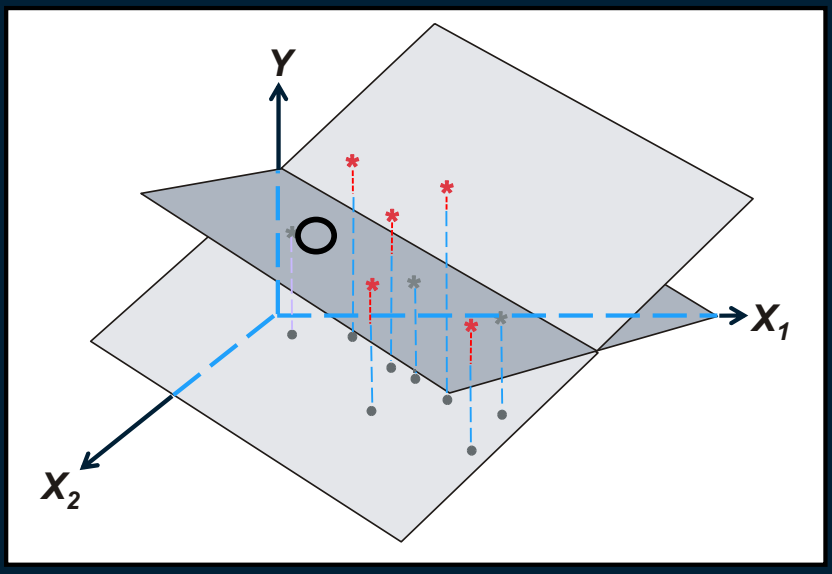
共線性の図解



共線性の図解



共線性の図解



共線性の検出方法

分散拡大係数（VIF）

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

共線性は
応答変数Yには
影響しない。

最初の説明変数を
残りの説明変数に
対して回帰分析する。

$X_1 = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$

モデルから
「R^2」を保存する。

各Xについて
繰り返す。

VIFの計算式
に代入する

$$\mathrm{VIF} = \frac{1}{1 - R^2}$$

$R^2 = 0.9$ → VIF = 10

VIFが10以上の場合、共線性の問題があることを示す。

共線性の除去

問題：共線性

共線性は、以下のどの仮定に反していますか？

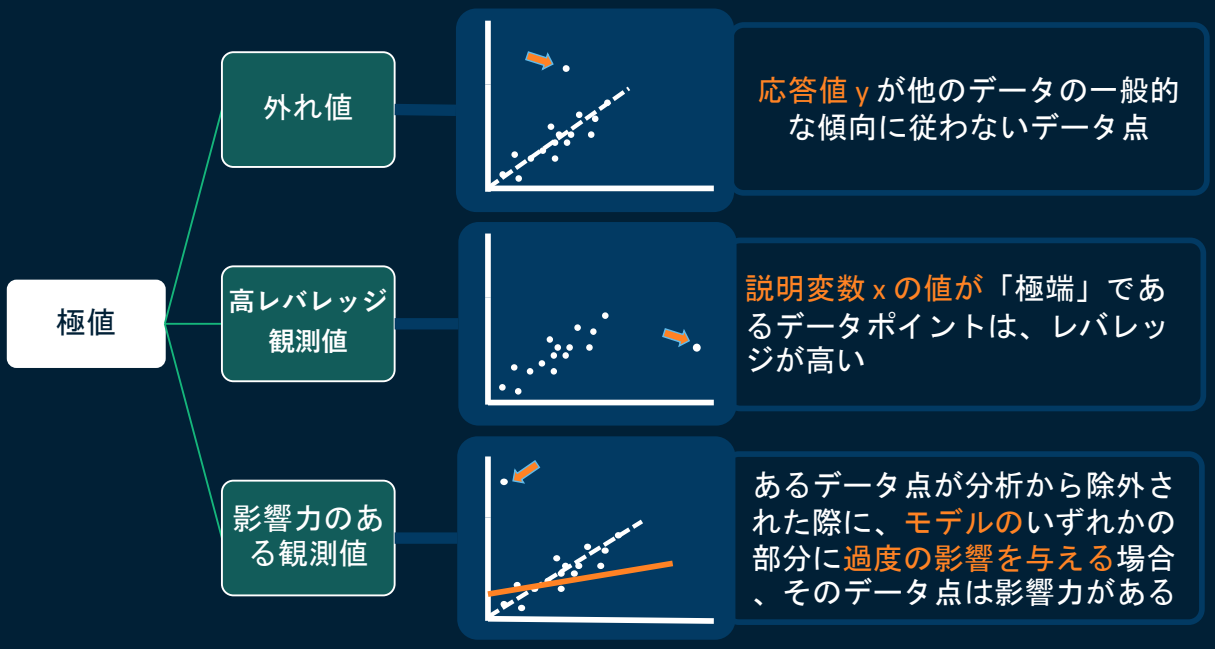
a. 誤差項の独立性
b. 分散が一定であること
c. 誤差の正規分布
d. 上記のいずれでもない

回答：共線性

共線性は、以下のどの仮定に反していますか？

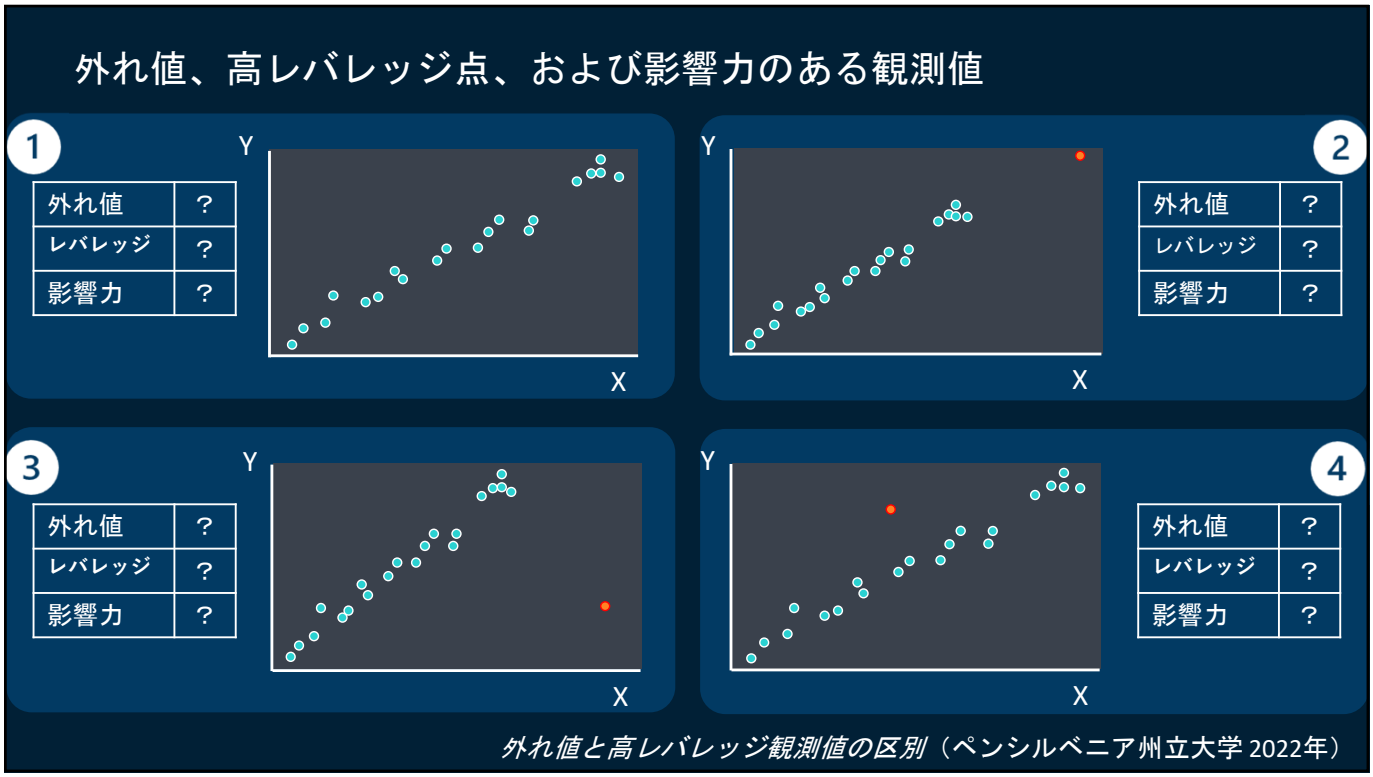
a. 誤差の独立性
b. 分散が一定であること
c. 誤差の正規分布
d. いずれも該当しない

潜在的な問題：外れ値



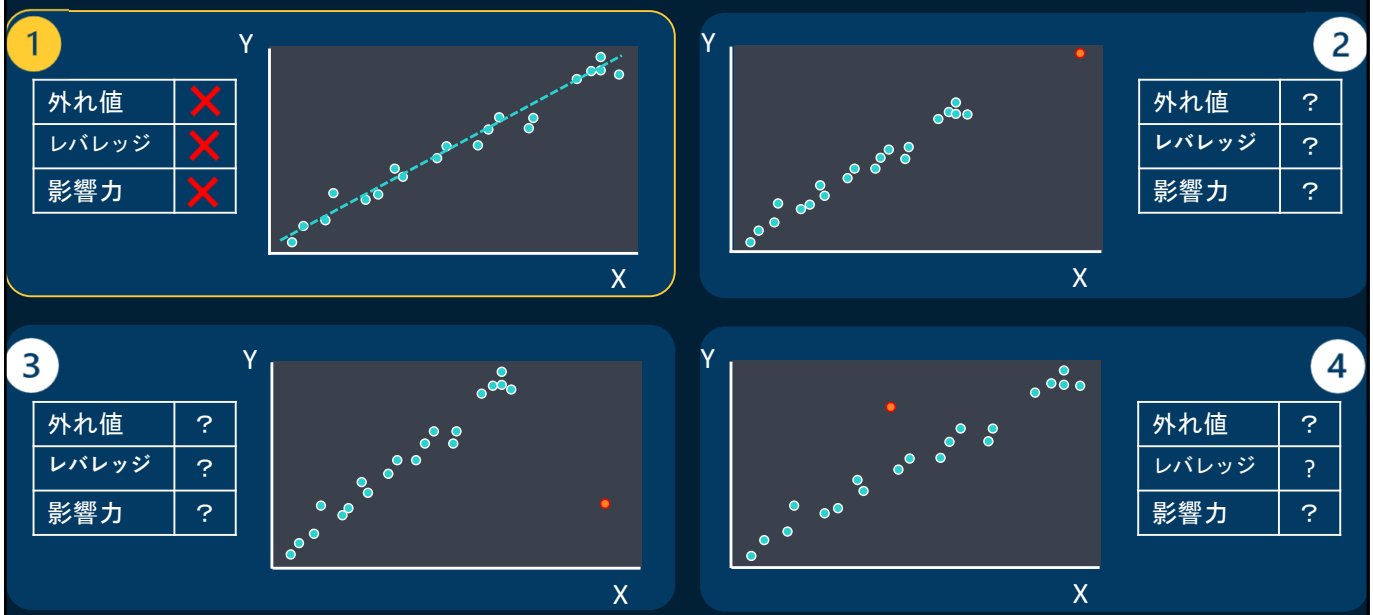
外れ値とレバレッジの高い観測値の区別が重要となる（ペンシルベニア州立大学 2022）

外れ値、高レバレッジ点、および影響力のある観測値

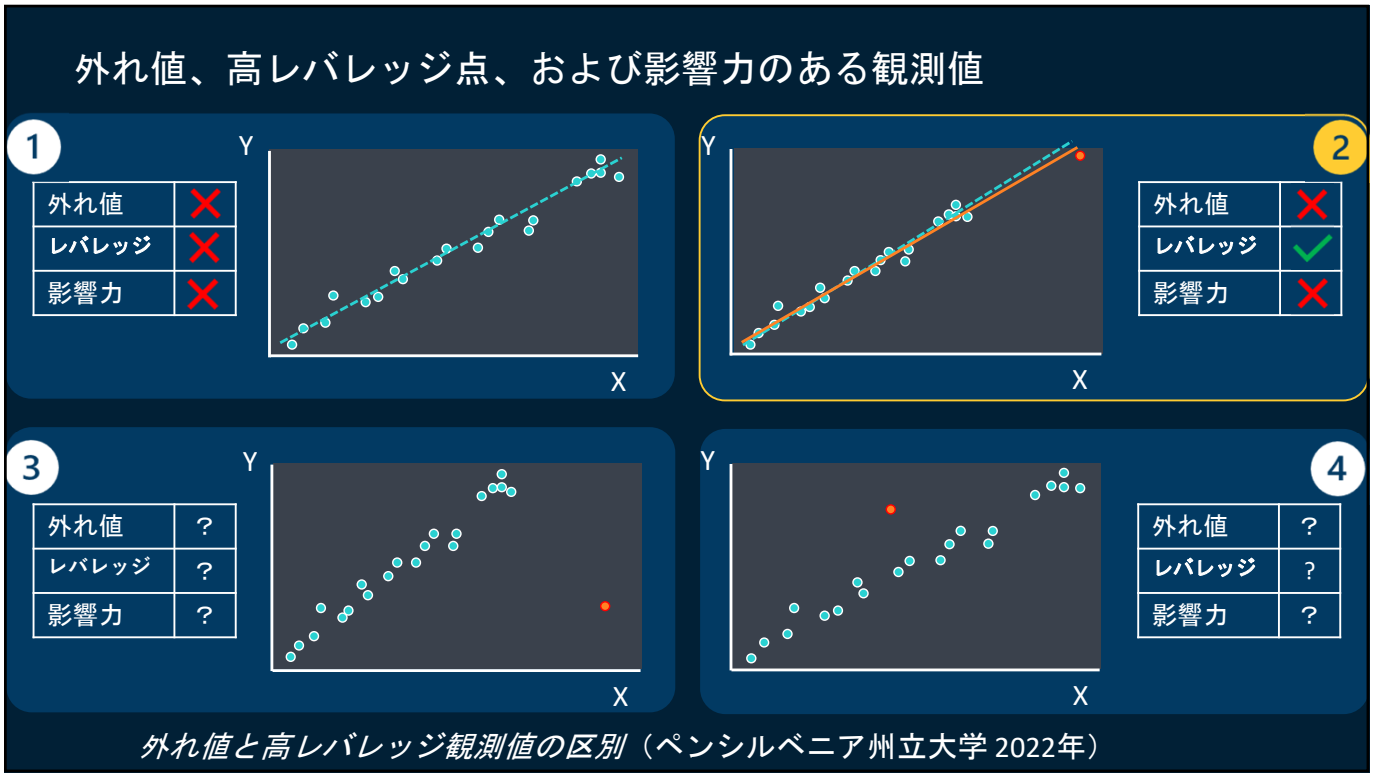


外れ値と高レバレッジ観測値の区別（ペンシルベニア州立大学 2022年）

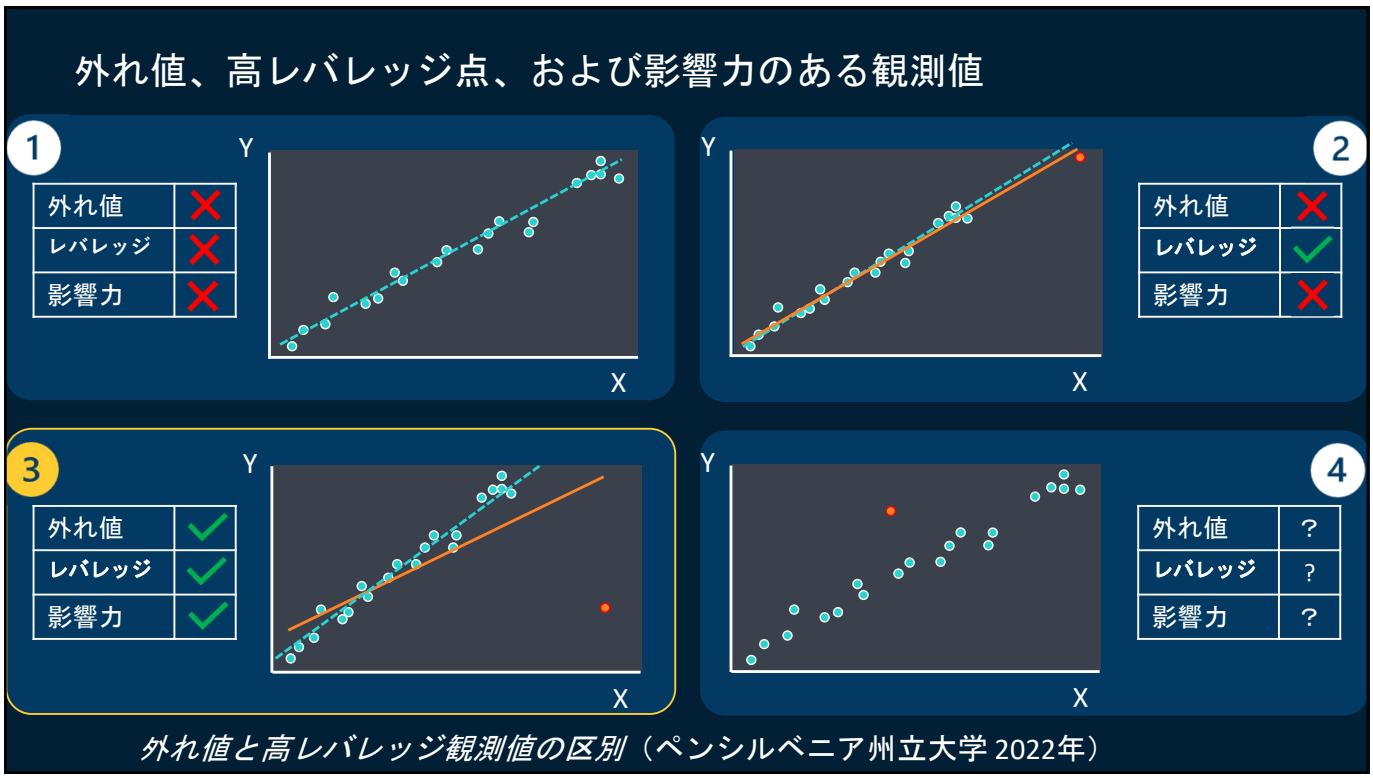
外れ値、高レバレッジ点、および影響力のある観測値



外れ値と高レバレッジ観測値の区別（ペンシルベニア州立大学 2022年）

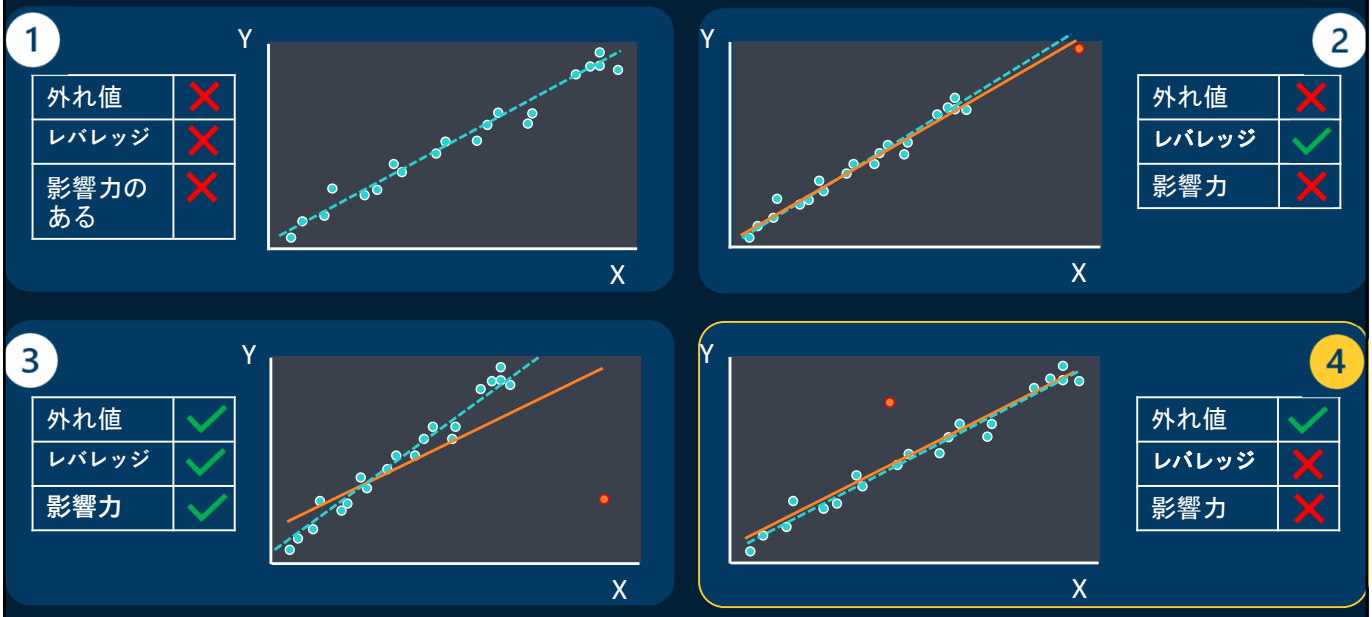


外れ値と高レバレッジ観測値の区別（ペンシルベニア州立大学 2022年）



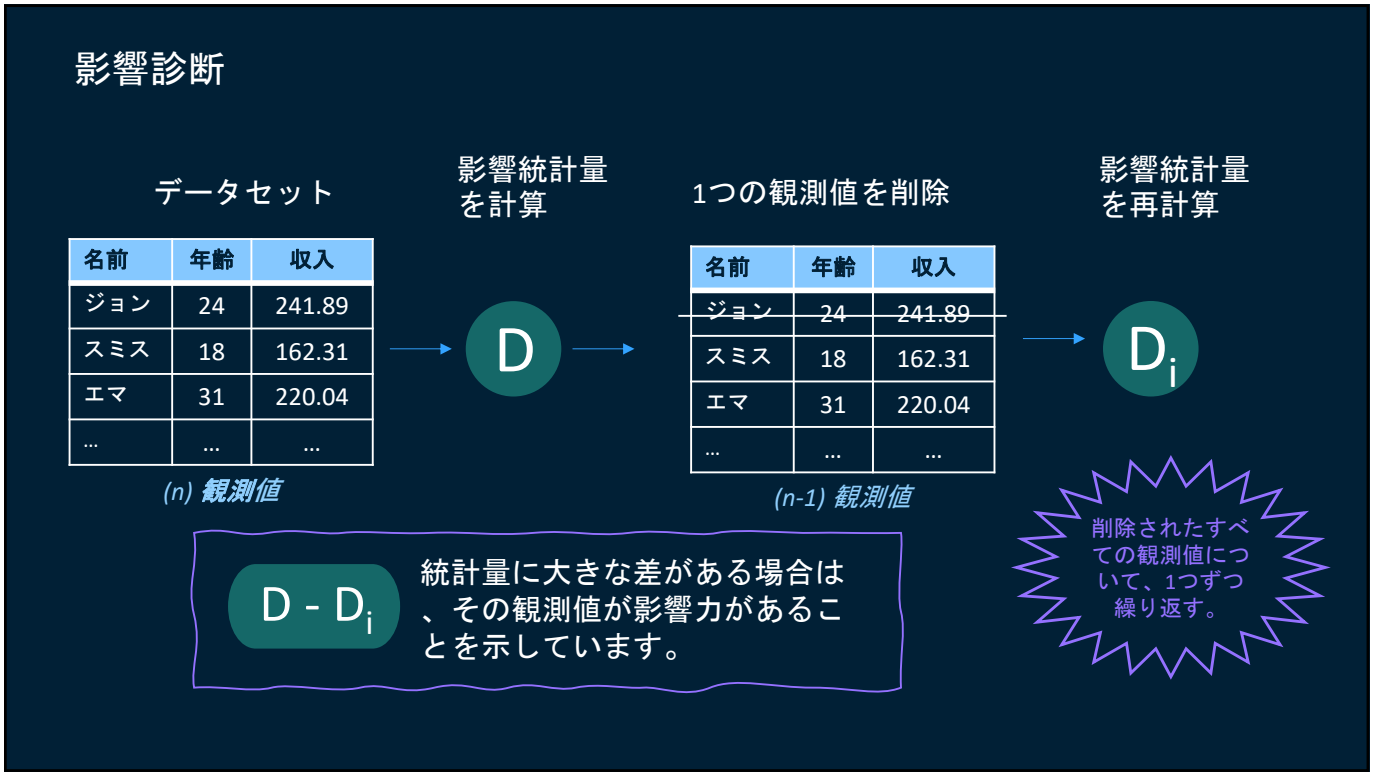
外れ値と高レバレッジ観測値の区別（ペンシルベニア州立大学 2022年）

外れ値、高レバレッジ点、および影響力のある観測値



外れ値と高レバレッジ観測値の区別（ペンシルベニア州立大学 2022年）

影響診断



外れ値と高レバレッジ観測値の検出

- スチューデント化残差
- レバレッジ
- クックのD
- 共分散比
- DFFITS

外れ値の検出

外れ値と高レバレッジ観測値の検出

- スチューデント化残差
- レバレッジ
- クックのD
- 共分散比
- DFFITS

影響力のある観測値の検出

外れ値および高レバレッジ観測値の検出

- スチューデント化残差
- レバレッジ
- クックのD
- 共分散比
- DFFITS

パラメータ推定値への影響

標準誤差への影響

予測への影響

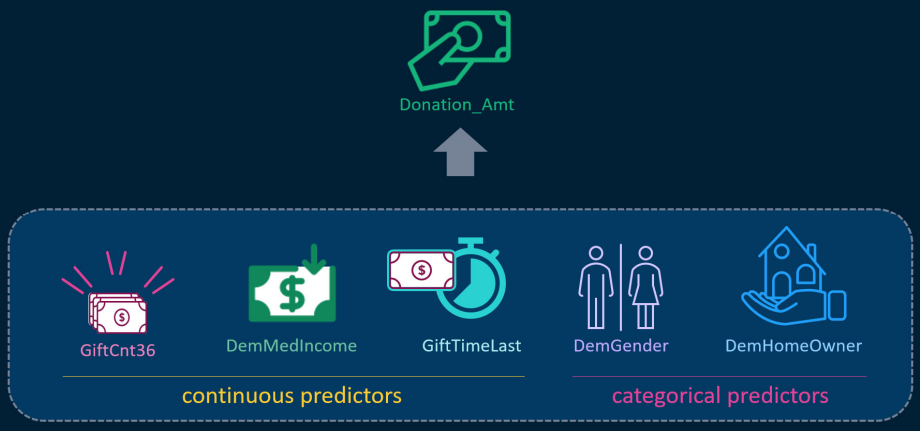
外れ値と高レバレッジ観測値の検出

- スチューデント化残差
- レバレッジ
- クックのD
- 共分散比
- DFFITS

外れ値や高レバレッジ観測値への対処法は、研究の目的によって異なります。

機械学習モデルは、一般的に外れ値に対して頑健です。

モデルの評価



- 回帰の仮定は満たされていたか？
- 説明変数間に共線性は見られるか？
- 影響力のある観測値や外れ値は存在するか？

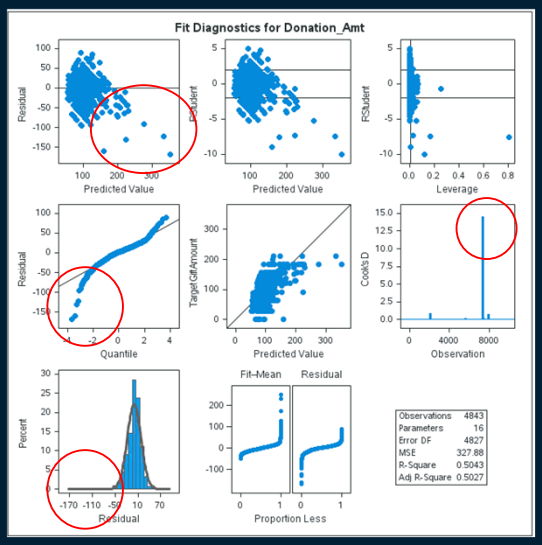
デモ：モデルの診断

このデモでは、多重回帰モデルの仮定の評価、および共線性や高レバレッジ観測値の確認について解説します。

このモデルについて何が分かったか？

負の残差が見られ、これはモデルの仕様が不適切であることを示唆している

残差の分布が歪んでいる



高レバレッジ観測値が少なくとも1つある

$R^2 = 0.50$

次のステップは？

- 予測変数や交互作用項を追加する。
- 説明変数を変換する。
- 他のモデルを試す。

ご質問はありますか？

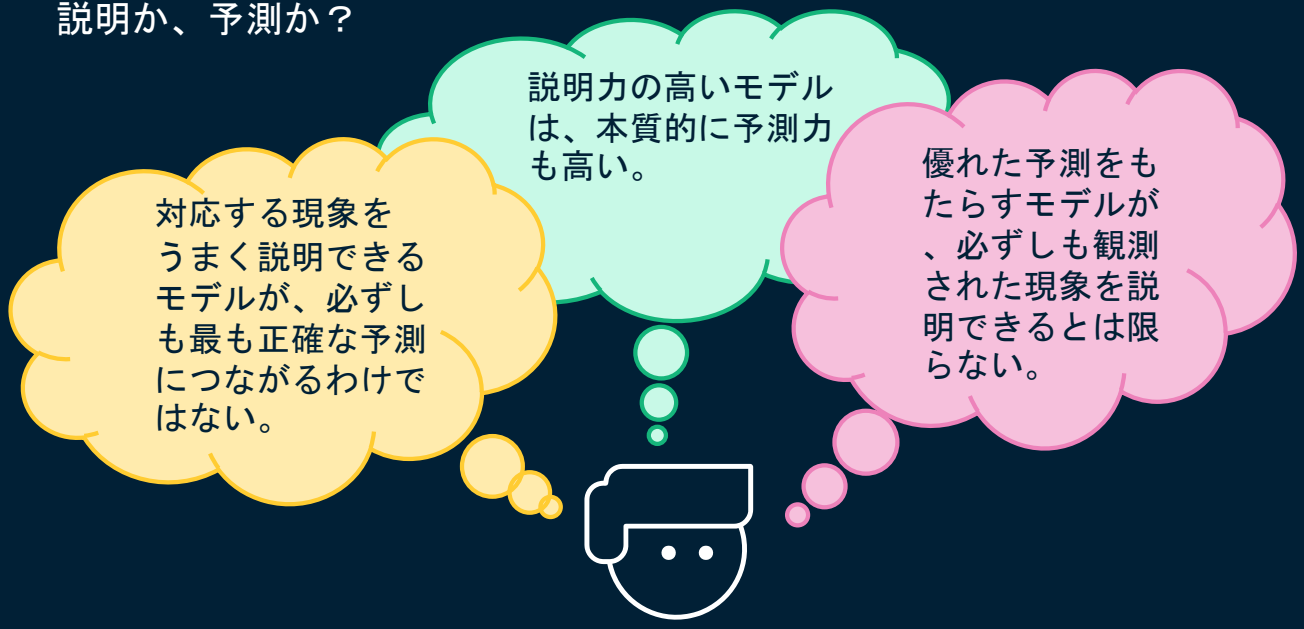
レッスンのまとめ

Lesson 4: ロジスティック回帰による予測モデリング

予測モデリング入門

Lesson 04, Section 01

説明か、予測か？



最良の予測モデルは、最良の説明モデルとは異なる（Sriboonchitta et al. 2019）

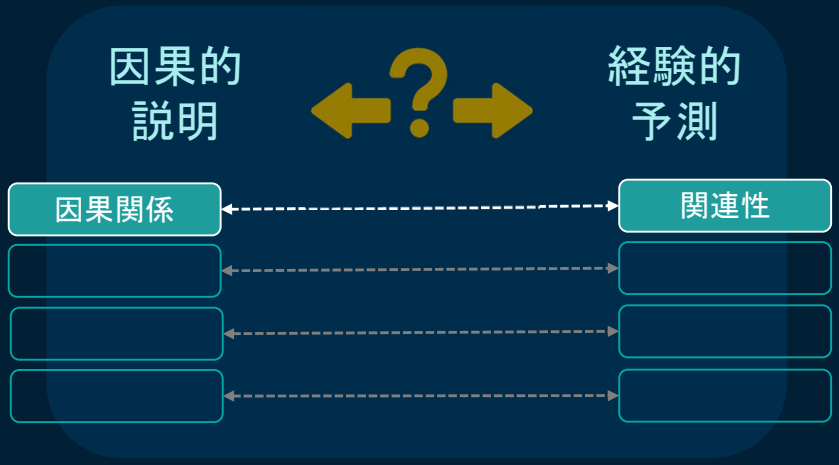
説明か、予測か？

説明モデリング	予測モデリング
因果的説明	経験的予測
• 因果仮説を検証する • 関係が統計的に有意であるかどうかを判断する	• 関連性を利用して将来の観測値を予測する • 予測精度に焦点を当てる
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$

(Sainani 2014)

説明か、予測か？

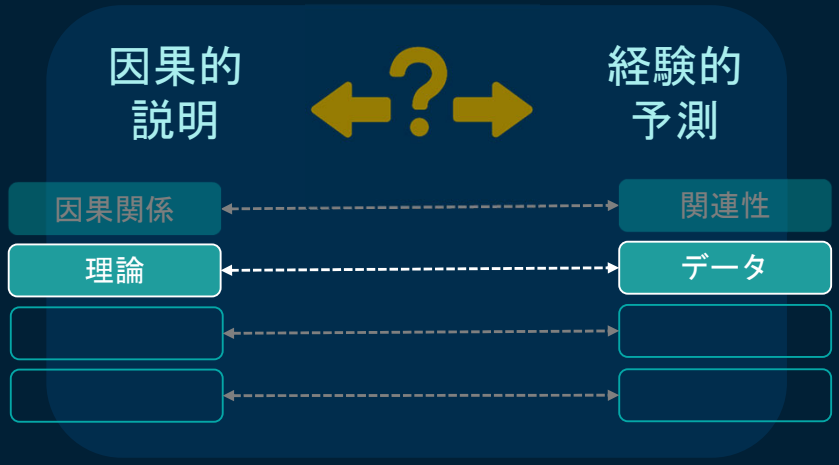
その違いとは？



(Shmueli, 2010)

説明か、予測か？

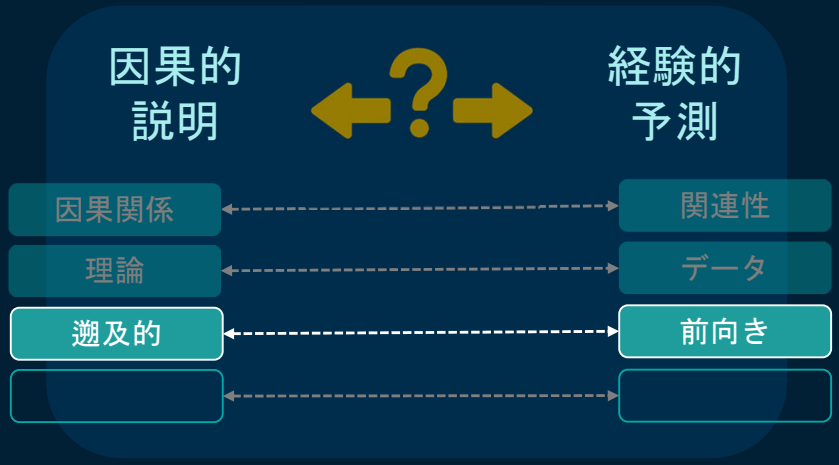
その違いとは？



(Shmueli, 2010)

説明か、予測か？

その違いとは？



(Shmueli, 2010)

説明か、予測か？

説明モデリング

回顧
（遡及的）

説明か、予測か？

予測モデリング

説明か、予測か？

違いは何ですか？

因果的説明		経験的予測
因果関係	⟷	関連性
理論	⟷	データ
遡及的	⟷	前向き
バイアス	⟷	バイアス－分散

(Shmueli 2010)

問題：説明と予測

ある通信会社は、過去のサービス利用状況や人口統計学的データを用いて、どの顧客が解約する可能性が高いかを特定したいと考えています。同社は回帰モデルを構築しましたが、以下の記述のうち正しいものはどれですか？

（該当するものをすべて選択してください）

a. これは本質的に説明モデリングの問題である。
b. これは本質的に予測モデリングの問題である。
c. おそらく、このモデルは、どの顧客が解約する可能性が高いかを予測するだけでなく、なぜ顧客が解約したのかを説明するためにも使用できる。
d. いずれも正しくない。

回答：説明と予測

ある通信会社は、過去のサービス利用状況や人口統計学的データを用いて、どの顧客が解約する可能性が高いかを特定したいと考えています。同社は回帰モデルを構築しましたが、以下の記述のうち正しいものはどれですか？

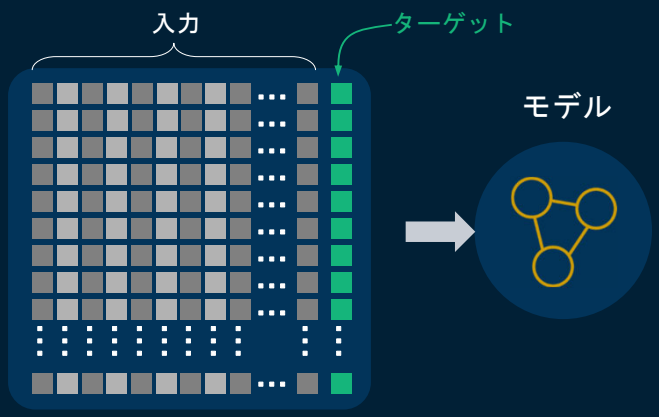
（該当するものをすべて選択してください）

a. これは本質的に説明モデリングの問題である。
b. これは本質的に予測モデリングの問題である。
c. おそらく、このモデルは、どの顧客が解約する可能性が高いかを予測するだけでなく、なぜ顧客が解約したのかを説明するためにも使用できる。
d. いずれも正しくない。

予測モデリング

予測モデル

過去データ



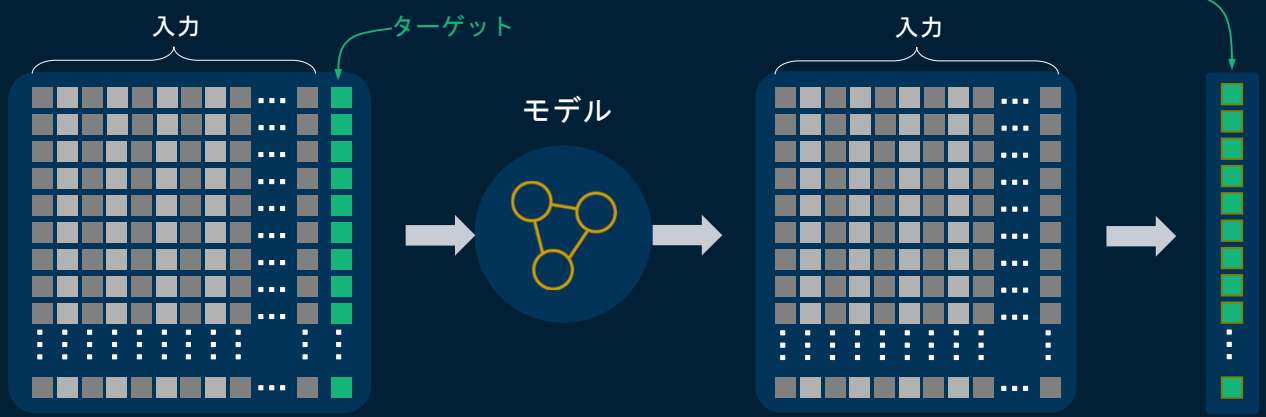
予測モデリング

スコアリング

過去データ

スコアリングデータ

予測



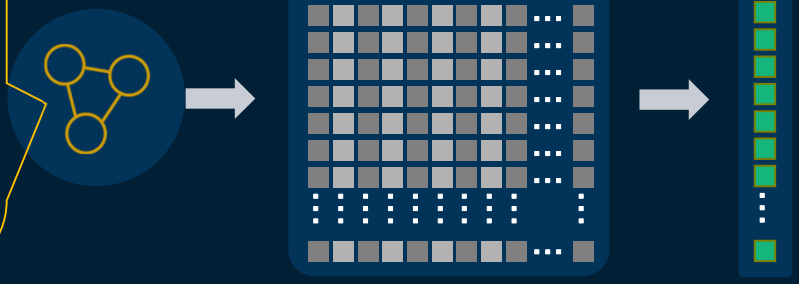
予測モデリング

スコアリング・レシピ

モデル
- 計算式 / 配分ルール

データの変更
- 派生入力
- 変換
- 欠損値の補完

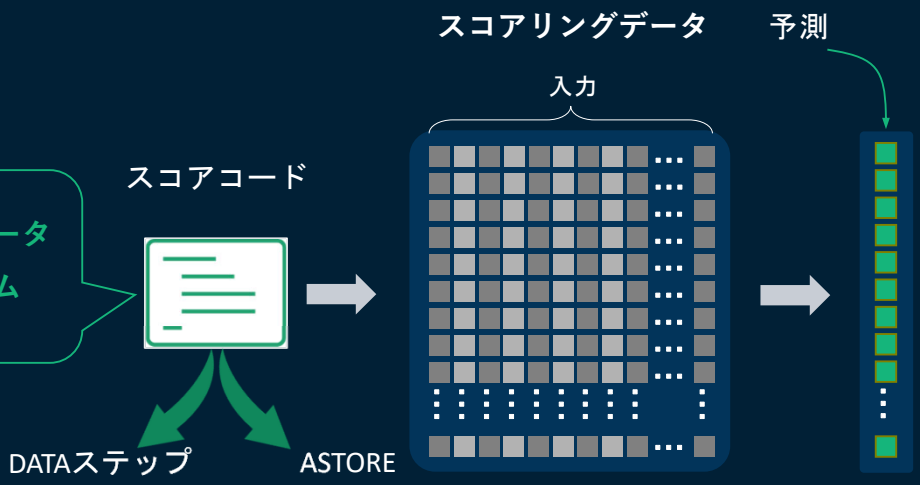
など



予測モデリング

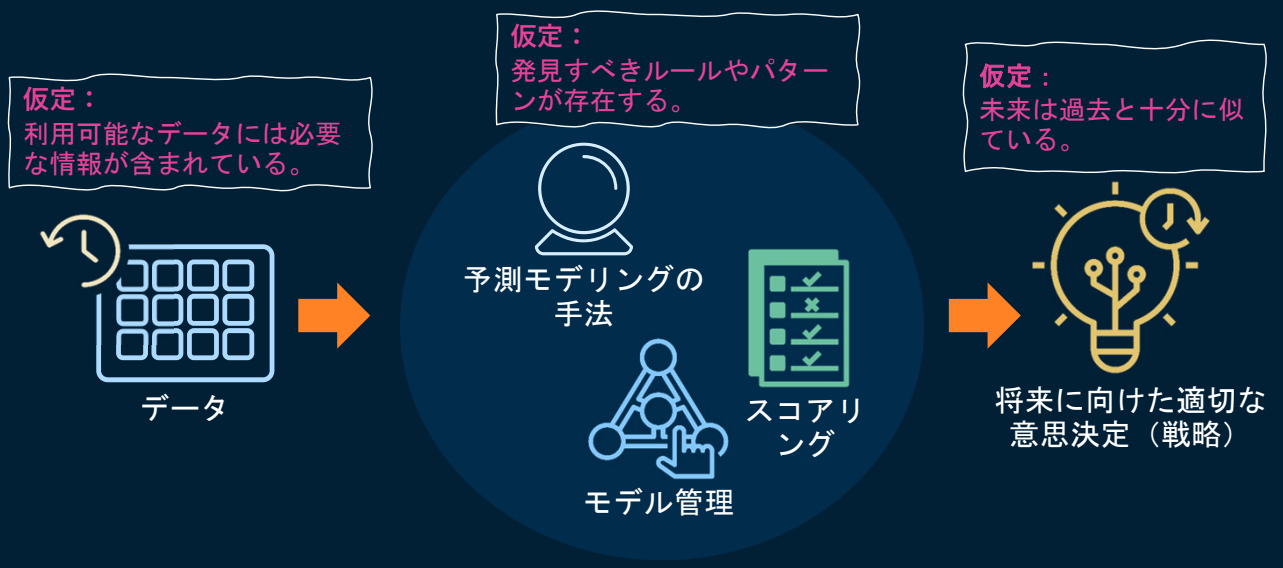
スコアリング・レシピ

- ≠スコアリング済みデータ
- ≠元の計算アルゴリズム



予測モデリング

予測モデリングの仮定



予測モデリングのための分析手法

問題：SAS Viyaにおけるスコアコード

SAS Viyaのスコアコードに関して、次の記述のうち正しいものはどれですか？（該当するものをすべて選択してください）

a. スコアコードとは、データ前処理を含むすべてのモデリングタスクを実装するSASコードである。
b. スコアコードは、元のモデリングアルゴリズムと厳密に同等である。
c. スコアコードは予測推定値を生成します。
d. スコアコードを使用して、データセットのスコアリングを行うことができます。

回答：SAS Viyaにおけるスコアコード

SAS Viyaのスコアコードに関して、次の記述のうち正しいものはどれですか？（該当するものをすべて選択してください）

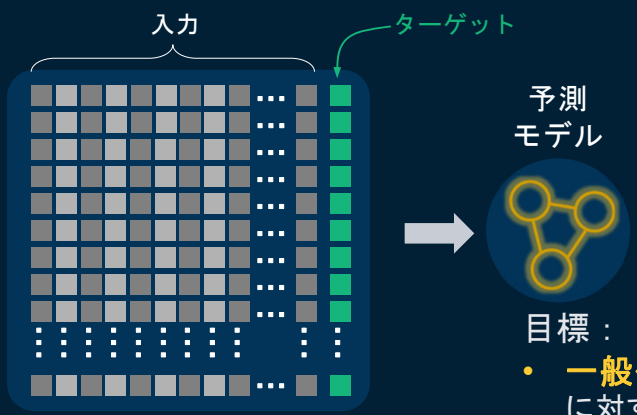
a. スコアコードとは、データ前処理を含むすべてのモデリングタスクを実装するSASコードである。
b. スコアコードは、元のモデリングアルゴリズムと厳密に同等である。
c. スコアコードは予測推定値を生成します。
d. スコアコードを使用して、データセットのスコアリングを行うことができます。

モデリングの期間

公正な評価

偽りの希望：楽観バイアス

過去データ



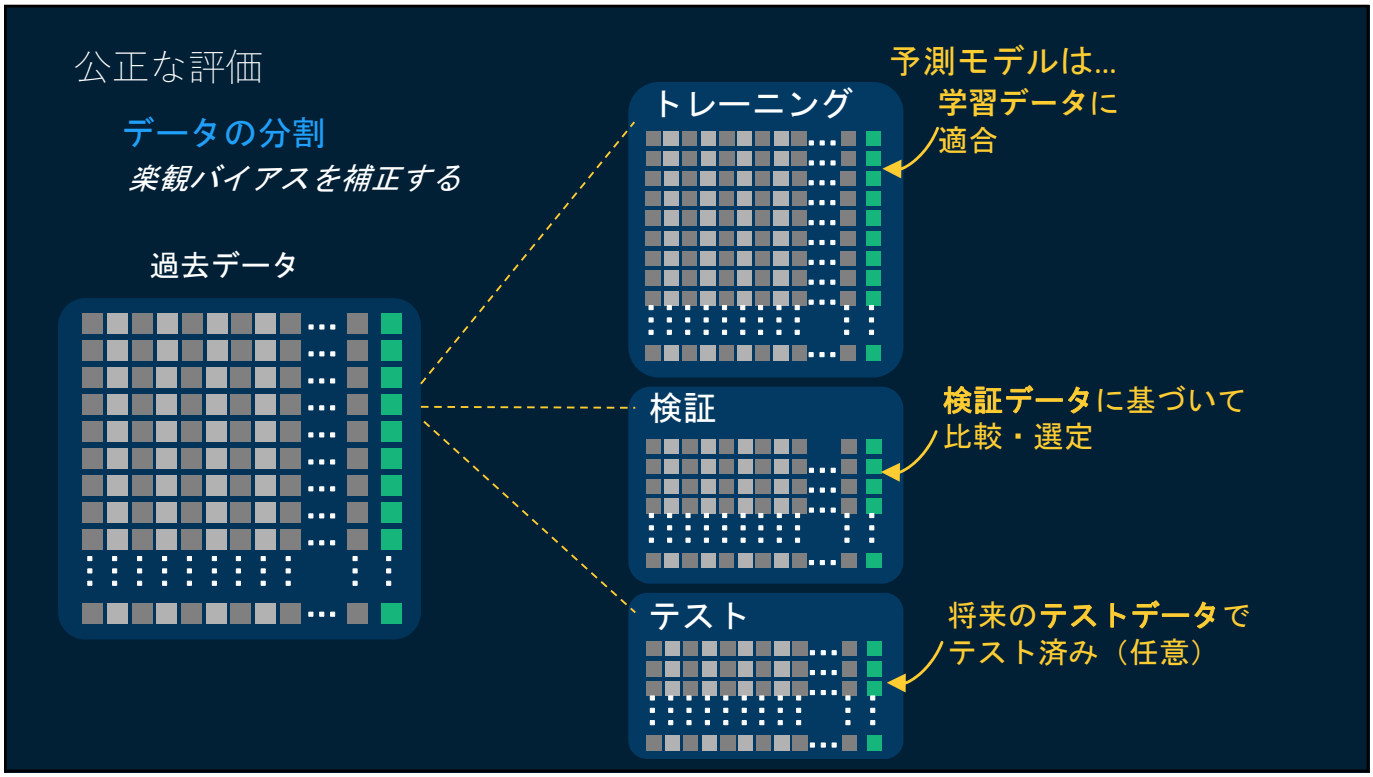
予測
モデル

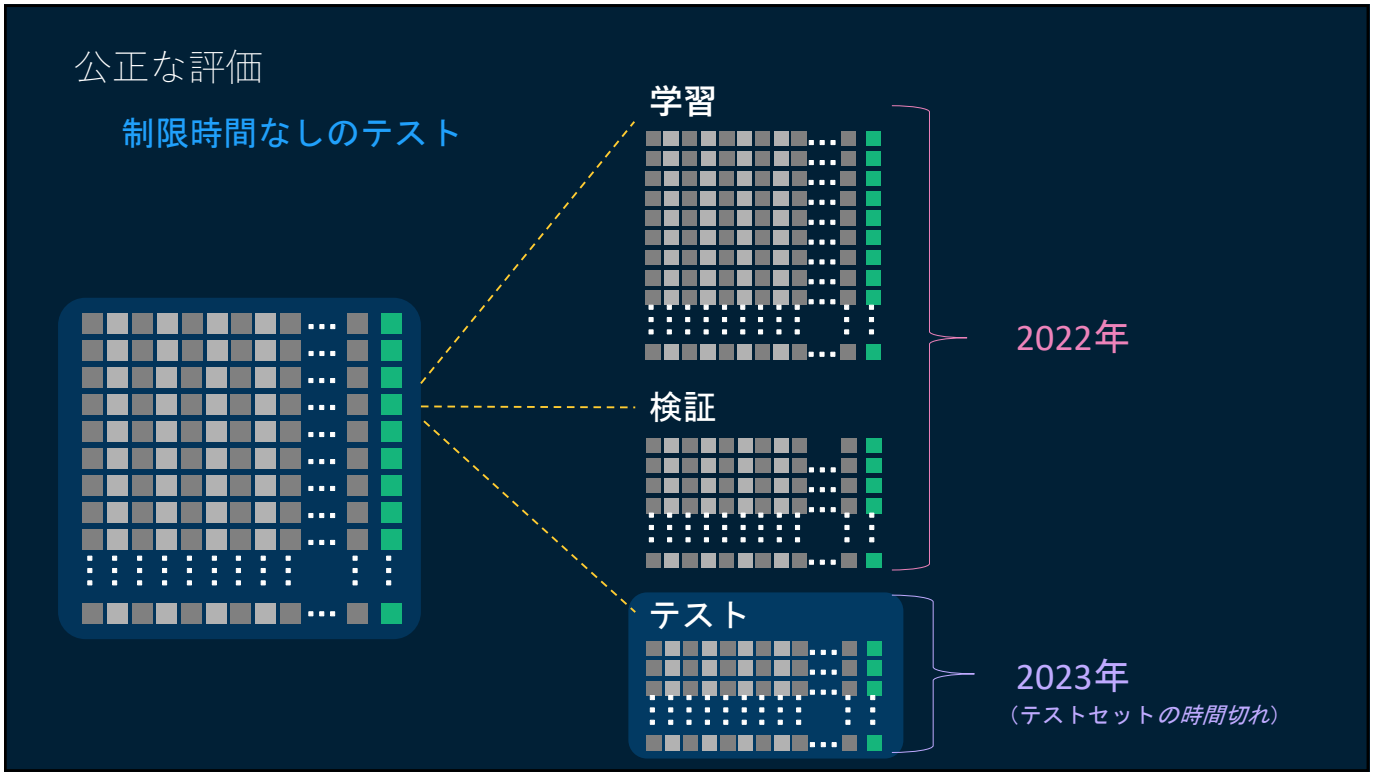
目標：

- **一般化**（新しいデータに対する予測性能）
- **偏りのない評価**

私のモデルは
データに完璧に
適合しています
…
大当たりだ！

（学習データのあらゆる特異性を最大限に活用した、Mosteller and Tukey 1977）


公正な評価
データの分割
楽観バイアスを補正する
過去データ
予測モデルは...
トレーニング
学習データに
適合
検証
検証データに基づいて
比較・選定
テスト
将来のテストデータで
テスト済み（任意）


公正な評価
制限時間なしのテスト
学習
検証
テスト
2022年
2023年
（テストセットの時間切れ）

問題：公正な評価

正直な評価に関して、以下の記述のうち正しいものはどれか。（該当するものをすべて選択してください）

a. 学習データ、検証データ、およびテストデータは、過去のデータから導出される。
b. テストデータは、学習データよりも必ず大きい。
c. 検証データおよびテストデータには、必ずしもターゲット変数が含まれるとは限りません。
d. モデルは学習データに基づいて構築され、検証データまたはテストデータを用いて評価される。

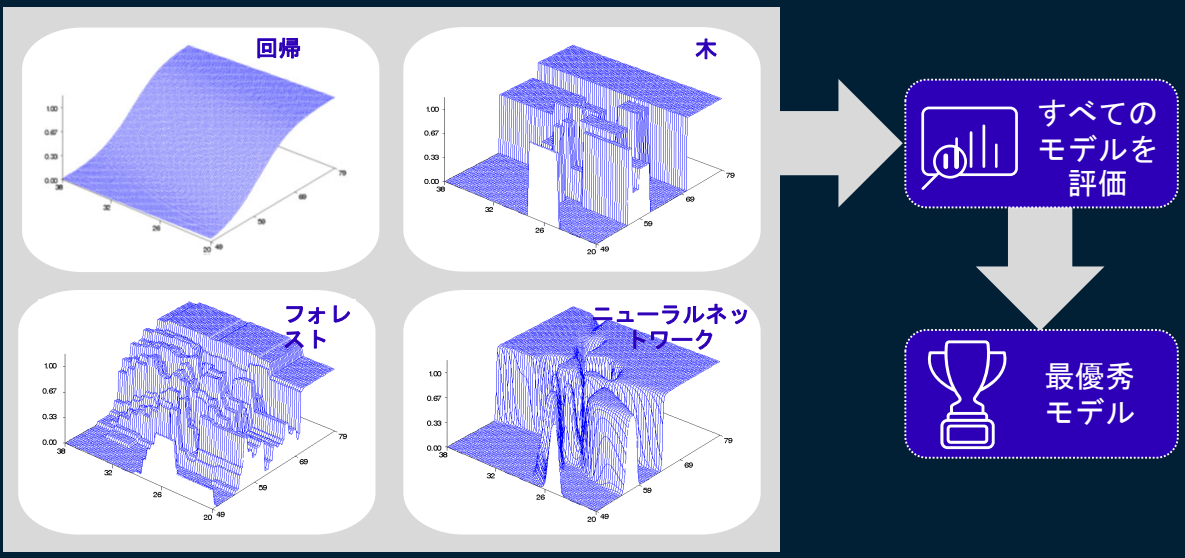
回答：公正な評価

正直な評価に関して、以下の記述のうち正しいものはどれか。（該当するものすべてを選択）

(a.) 学習データ、検証データ、およびテストデータは、過去のデータから導出される。
b. テストデータは、学習データよりも必ず大きい。
c. 検証データおよびテストデータには、必ずしもターゲット変数が含まれるとは限りません。
(d.) モデルは学習データに基づいて構築され、検証データまたはテストデータを用いて評価される。

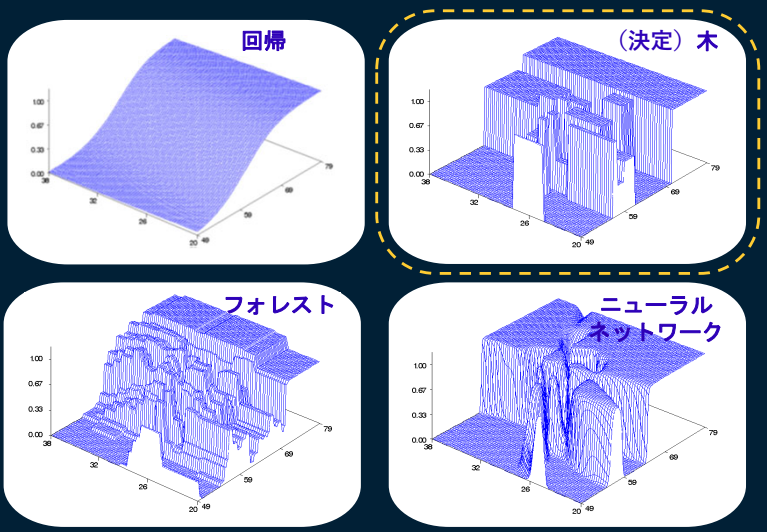
候補モデル

アルゴリズムの多様性



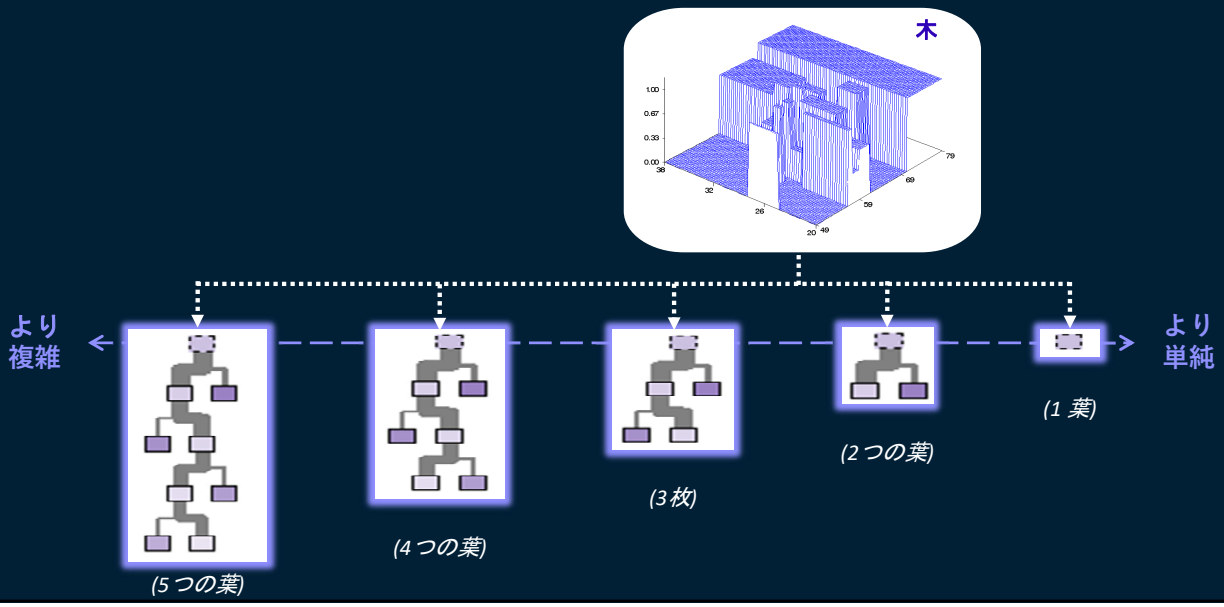
候補モデル

アルゴリズム内の多様性



候補モデル

モデルの複雑さ



問題：モデルの多様性

データマイニングや機械学習はデータ駆動型であるため、複数の候補モデルを作成し、その中から最適なものを選択するのが常に良い方法です。

a. 正しい
b. 誤り

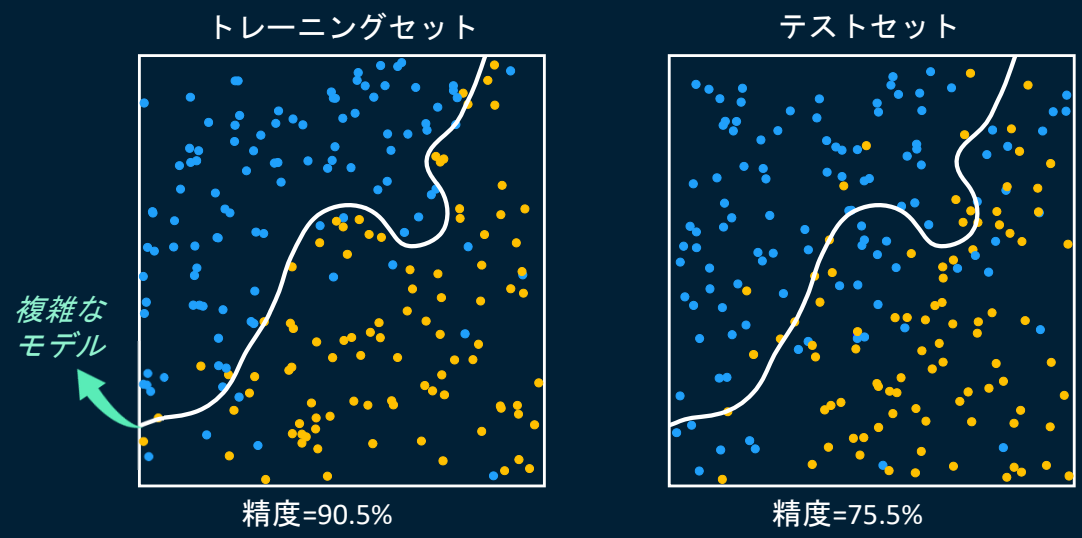
回答：モデルの多様性

データマイニングや機械学習はデータ駆動型であるため、複数の候補モデルを作成し、その中から最適なものを選択するのが常に良い方法です。

a. 正しい
b. 誤り

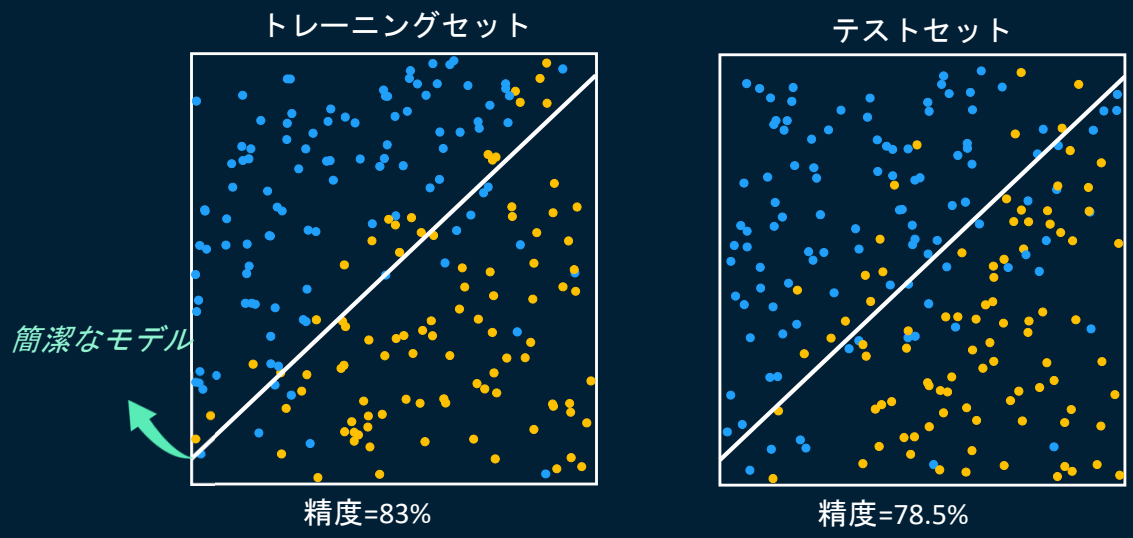
モデルの複雑さの最適化

適合不良の例 ☹



モデルの複雑さの最適化

優れた適合の例 ☺



モデルの複雑さの最適化

どのモデルが最適か？

モデルの複雑さの最適化

予測誤差の分解

予測誤差

バイアス

モデル予測の平均値と真値との間の誤差

分散

与えられたデータに対するモデル予測の平均的な変動

モデルの複雑さの最適化

バイアス・分散のトレードオフ

予測誤差

モデルの柔軟性または複雑さ

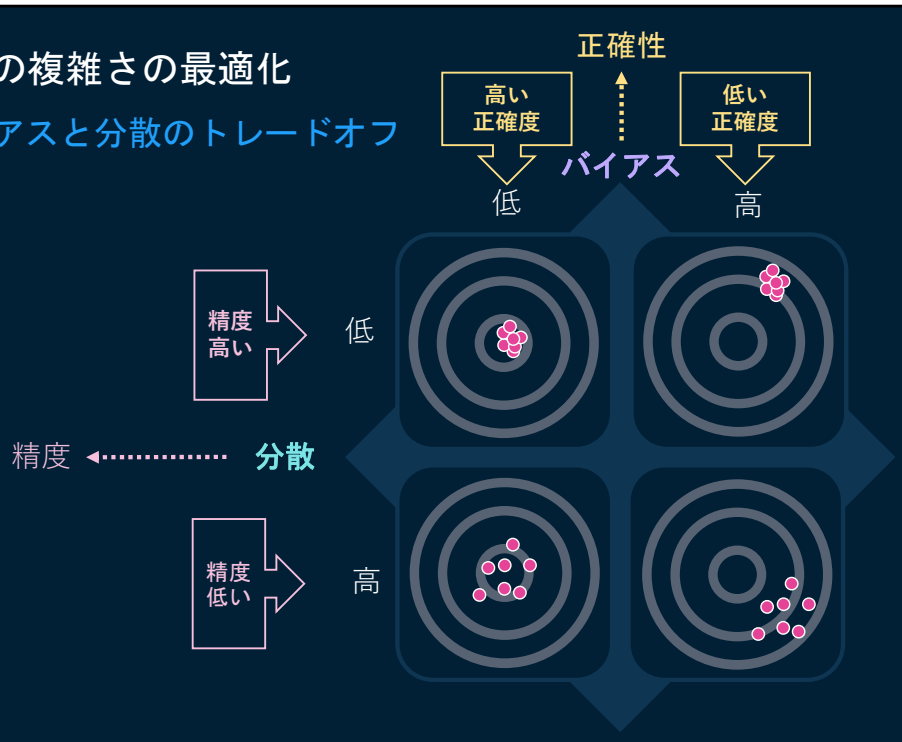
バイアス

分散

モデルの感度

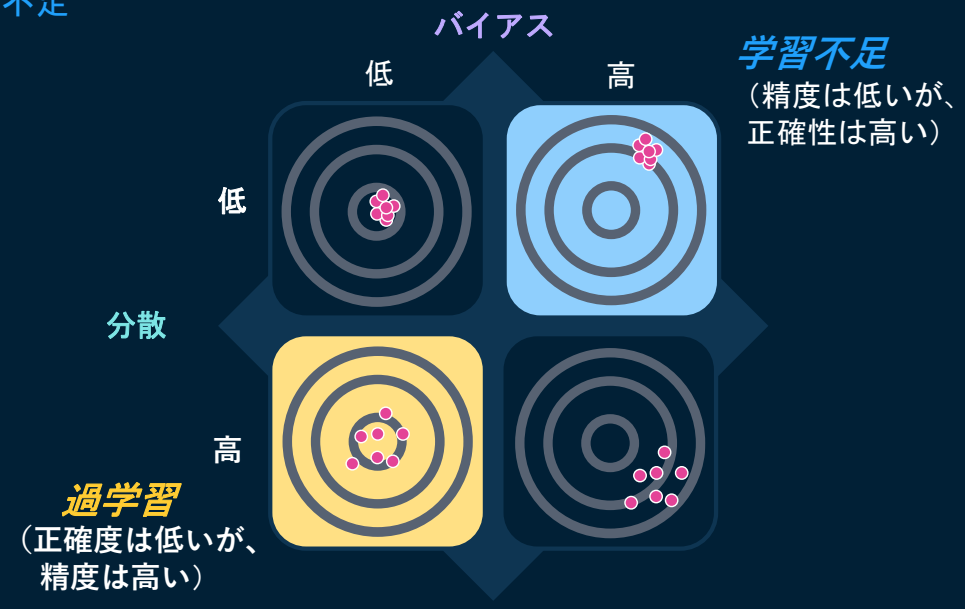
モデルの複雑さの最適化

バイアスと分散のトレードオフ



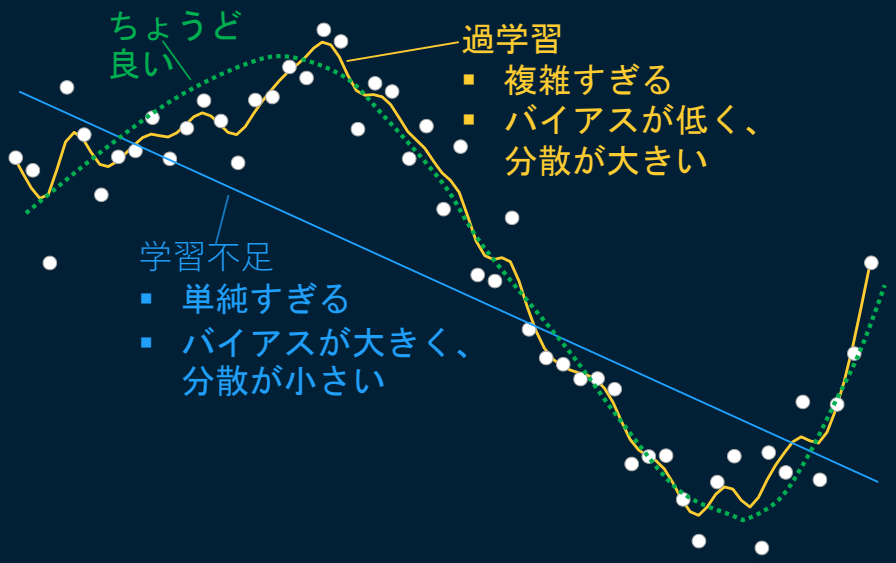
モデルの複雑さの最適化

過学習と学習不足



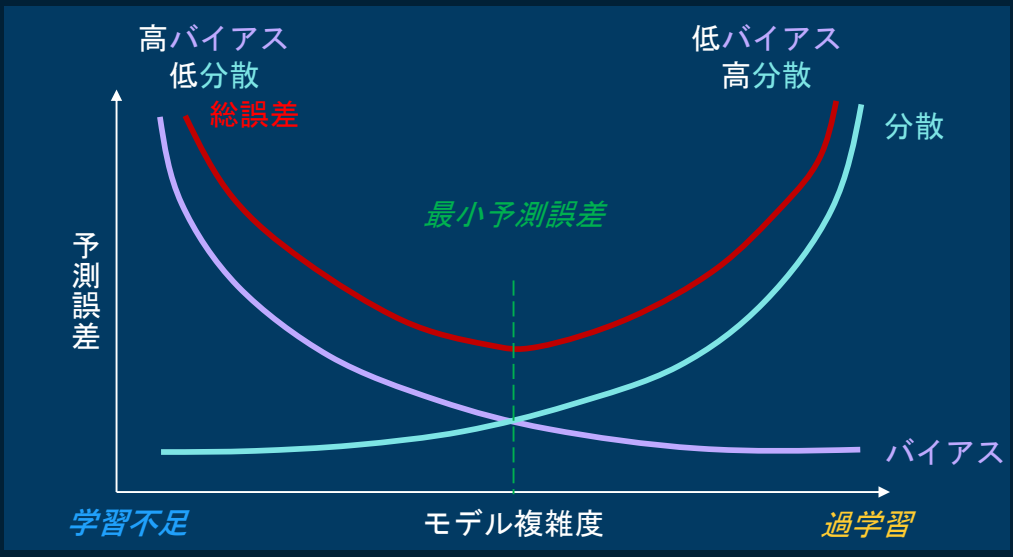
モデル複雑度の最適化

適切なモデルの複雑さとは何か？



モデルの複雑さの最適化

バイアスと分散の最適バランス

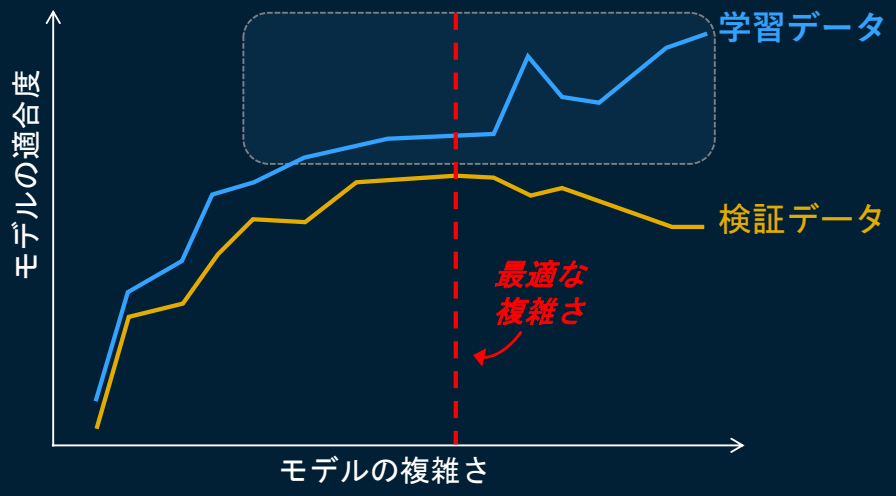


(Hastie, Tibshirani, and Friedman 2001)

モデルの複雑さの最適化

適合度と複雑さの関係

モデルがデータ内の関連する傾向を捉え、かつ訓練データセットのばらつきを特定するにつれて、適合度は向上する



予測誤差の分解

問題：予測モデル

予測モデルは広く利用されており、多種多様なものが存在します。しかし、どのようなモデルであっても、以下の重要な役割を果たす必要があります：

- 入力測定値から予測値へと変換するためのルールを提供すること
- 過学習と過小学習のバランスをとるためにその複雑さを調整できること。

a. 正しい
b. 誤り

回答：予測モデル

予測モデルは広く利用されており、多種多様なものが存在します。しかし、どのようなモデルであっても、以下の重要な役割を果たす必要があります：

- 入力測定値から予測値へと変換するためのルールを提供すること
- 過学習と過小学習のバランスをとるためにその複雑さを調整できること。

a. 正しい
b. 誤り

カテゴリカルな関連性

カテゴリカルな関連性の理解

Lesson 04, Section 02

問題：カテゴリ変数

PVA_Donorsデータにおいて、以下のうちどれがカテゴリ変数とみなされる可能性が高いでしょうか？（該当するものをすべて選択してください。）

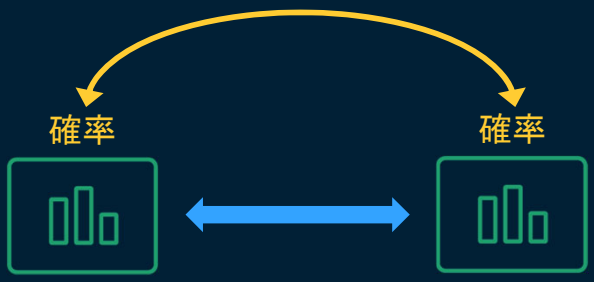
a. 年齢 (DemAge)
b. 性別 (DemGender)
c. 前回の寄付額 (GiftAvgLast)
d. 前回の寄付からの期間 (GiftTimeLast)
e. 住宅所有者 (DemHomeOwner)

回答：カテゴリ変数

PVA_Donorsデータにおいて、以下のうちどれがカテゴリ変数とみなされる可能性が高いでしょうか？（該当するものをすべて選択してください。）

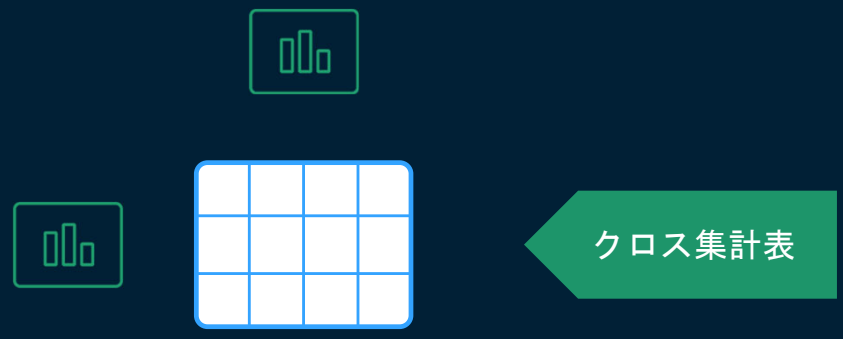
a. 年齢 (DemAge)
b. 性別 (DemGender)
c. 前回の寄付額 (GiftAvgLast)
d. 前回の寄付からの期間 (GiftTimeLast)
e. 住宅所有者 (DemHomeOwner)

カテゴリ変数間の関連性



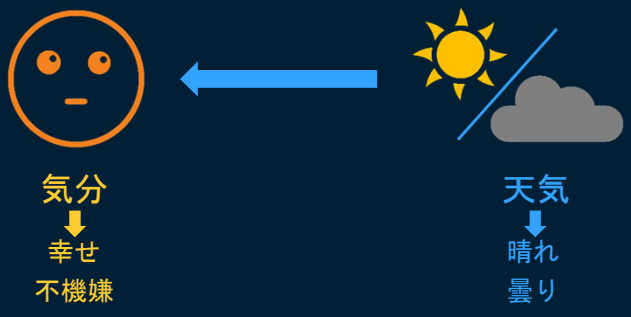
- ある変数の値が変わると、もう一方の変数の分布も変化する。
- ある変数の確率は、もう一方の変数の確率に依存する。

カテゴリ変数間の関連性

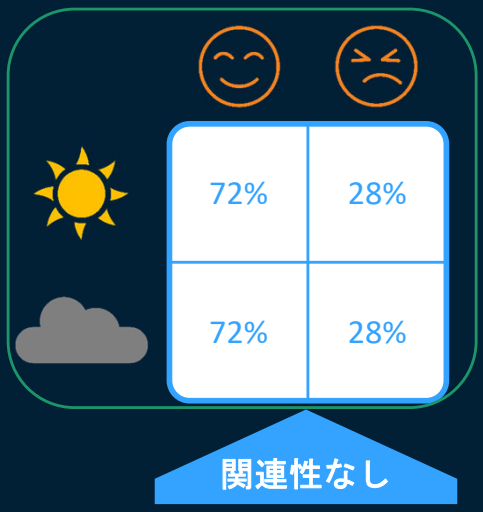


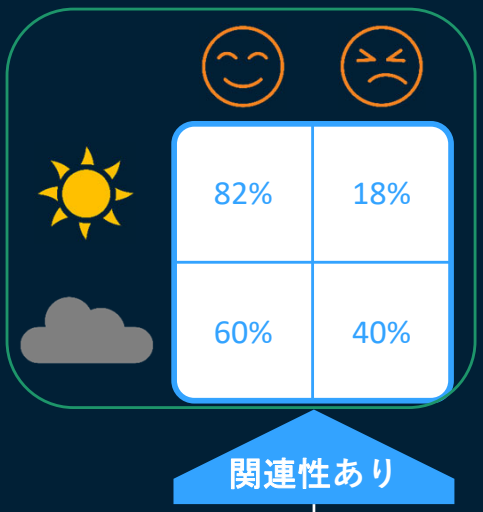
変数の組み合わせごとの値の頻度

カテゴリ変数間の関連性



カテゴリ変数間の関連性





統計的に有意か？

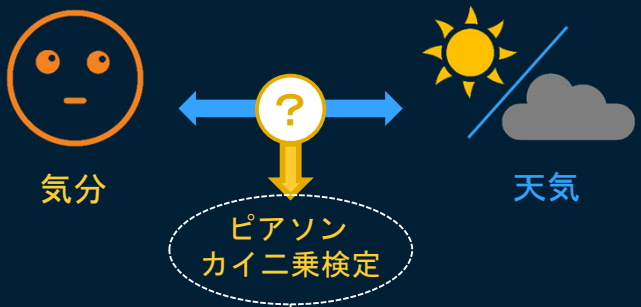
カテゴリ変数間の関連性

H_0 : 関連なし

H_a : 関連あり

確率は同じ、
天候にかかわらず

確率は同じでない、
天候にかかわらず



機械学習における一般的な特徴量選択手法

カテゴリ変数間の関連性

H_0 : 関連なし　　H_a : 関連あり

応答 ⟷ 予測変数

ピアソン
カイ二乗検定

	1	2	合計
1	E O	E O	R
2	E O	E O	R
合計	C	C	T

$E=R*C/T$

O → 観測度数

E → 期待度数

カテゴリ変数間の関連性

H_0 : 関連なし　　H_a : 関連あり

応答 ⟷ 予測変数

ピアソン
カイ二乗検定

	1	2	合計
1	E O	E O	R
2	E O	E O	R
合計	C	C	T

χ^2 → 有意

カテゴリ変数間の関連性

H_0 : 関連なし　　H_a : 関連あり

応答 ⟷ 予測変数

ピアソン
カイ二乗検定

	1	2	合計
1	E O	E O	R
2	E O	E O	R
合計	C	C	T

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

カテゴリ変数間の関連性

H_0 : 関連なし　　H_a : 関連あり

応答 ⟷ 予測変数

ピアソン
カイ二乗検定

	1	2	合計
1	E O	E O	R
2	E O	E O	R
合計	C	C	T

$$\sum\sum \left(\frac{(\text{observed}_{rc} - \text{expected}_{rc})^2}{\text{expected}_{rc}} \right)$$

カテゴリ変数間の関連性

H_0 : 関連なし　　H_a : 関連あり

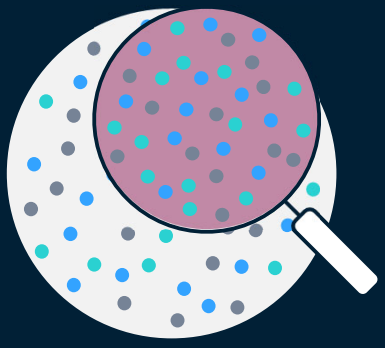
応答 ⟷ 予測変数

ピアソン
カイ二乗検定

関連性が存在するという確
信度を示すに過ぎない

カイ二乗
統計量 χ^2

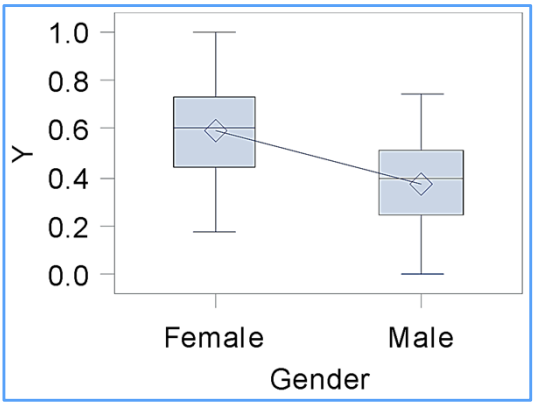
p値 p



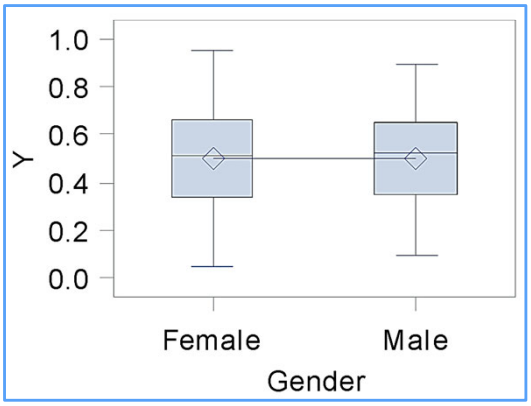
問題：関連性

2つの連続変数間の関連性（第3課）と、2つのカテゴリ変数間の関連性（本課）については、すでに学習しました。以下は、連続変数とカテゴリ変数の関係を示す図です。次のうち、関連性を表しているのはどれですか？

a.



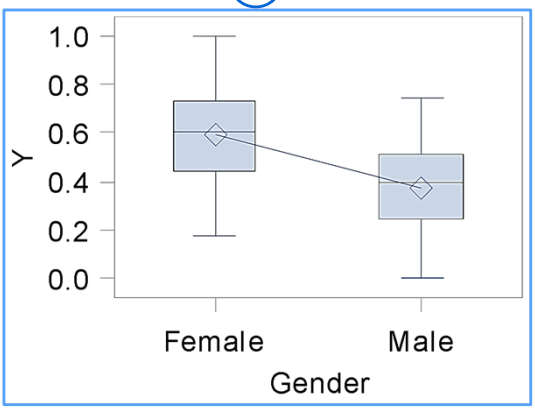
b.



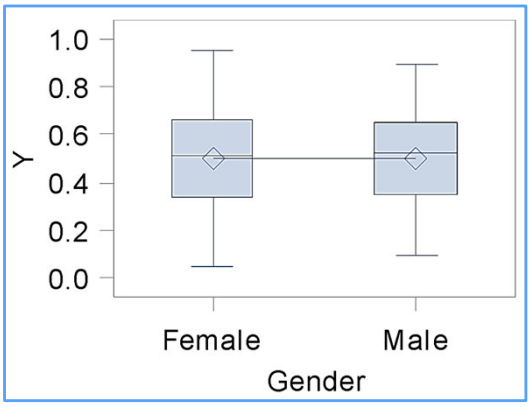
回答：関連性

2つの連続変数間の関連性（第3課）と、2つのカテゴリ変数間の関連性（本課）については、すでに学習しました。以下は、連続変数とカテゴリ変数の関係を示す図です。次のうち、関連性を表しているのはどれですか？

a.

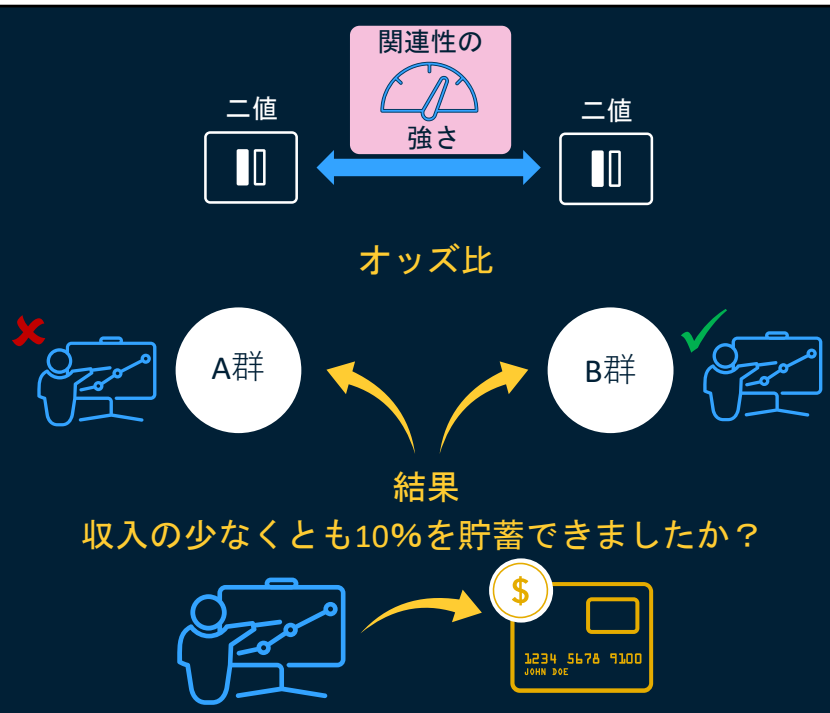


b.



クレイマーのV統計量

オッズ比



オッズ比

$$\text{Odds} = \frac{P_{event}}{1 - P_{event}}$$

群	結果		
	はい	いいえ	合計
A	60	20	80
B	90	10	100
合計	150	30	180

B群における「はい」の確率

$90/100 = .90 = 90\%$

A群における「はい」の確率

$60/80 = .75 = 75\%$

B群における「いいえ」の確率

$10/100 = .10 = 10\%$

A群における「いいえ」の確率

$20/80 = .25 = 25\%$

オッズ比

B対A

$$\text{Odds} = \frac{P_{event}}{1 - P_{event}}$$

群	結果		
	はい	いいえ	合計
A	60 .75	20 .25	80
B	90 .90	10 .10	100
合計	150	30	180

B群における「はい」のオッズ

$$.90 / .10 = 9 = 9:1$$

A群における「はい」のオッズ

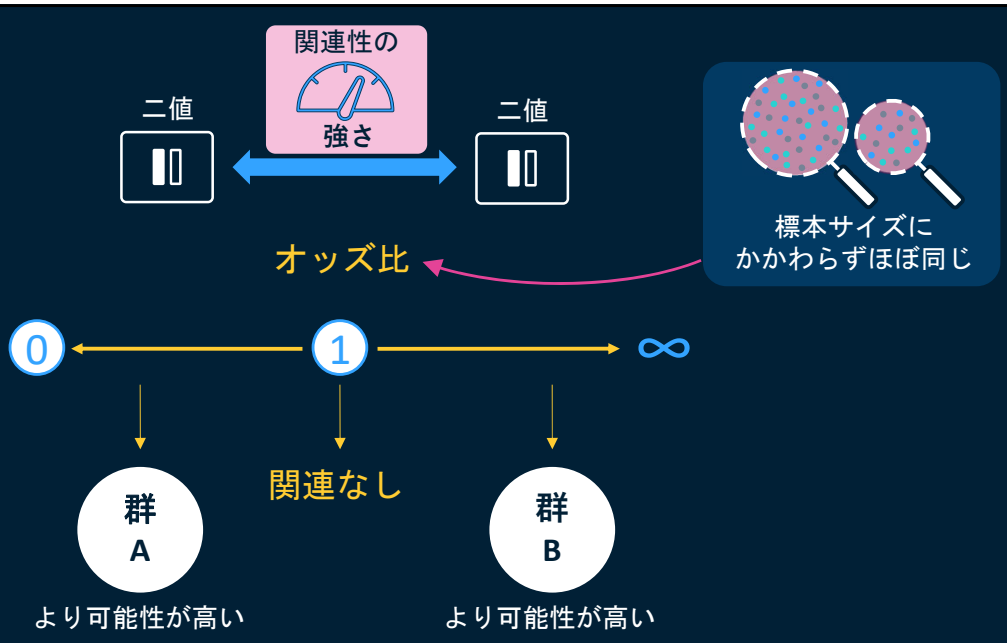
$$.75 / .25 = 3 = 3:1$$

$$9 / 3 = 3 \Rightarrow$$

オッズ比、A対B = 1/[オッズ比、B対A]
= 1/3 = 0.3333

オッズ比

B対A



問題：関連性の強さ

ある研究者が、2つの二値変数間の関連性の強さを測定したいと考えています。彼女はどの統計量を使用すればよいでしょうか？

a. ピアソン相関係数
b. オッズ比
c. 決定係数（R^2）
d. 上記のいずれでもない

回答：関連性の強さ

ある研究者が、2つの二値変数間の関連性の強さを測定したいと考えています。彼女はどの統計量を使用できますか？

a. ピアソン相関係数
b. オッズ比
c. 決定係数（R^2）
d. 上記のいずれでもない

デモ：カテゴリ変数の関連性の検証

このデモでは、カテゴリ変数 StatusCatStarAll と Response の間の関連性を検定する方法を示します。

ロジスティック回帰モデル

ロジスティック回帰の使い方

Lesson 04, Section 03

問題：回帰モデル

回帰モデルに関して、次の記述のうち正しいものはどれか。（該当するものすべてを選択してください。）

a. 線形回帰は説明モデルにのみ使用できる。
b. ロジスティック回帰は、予測モデリングにのみ使用できる。
c. 線形回帰は、説明モデリングだけでなく予測モデリングにも使用できる。
d. ロジスティック回帰は、説明モデリングだけでなく、予測モデリングにも使用できる。

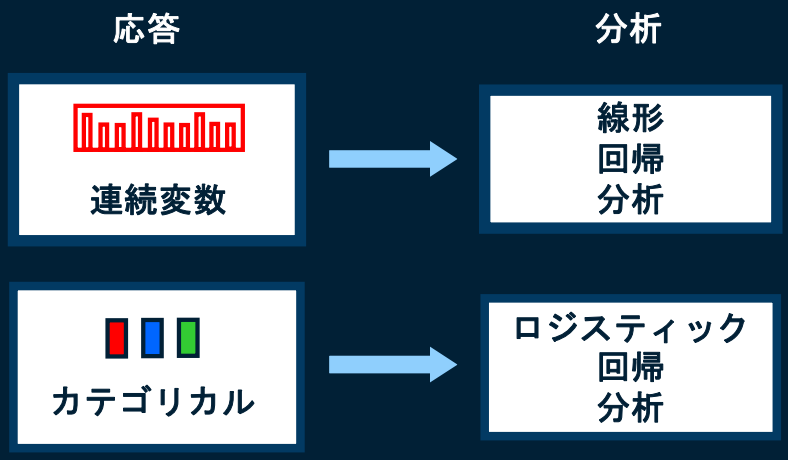
回答：回帰モデル

回帰モデルに関して、次の記述のうち正しいものはどれか。（該当するものすべてを選択してください。）

a. 線形回帰は説明モデルにのみ使用できる。
b. ロジスティック回帰は、予測モデリングにのみ使用できる。
c. 線形回帰は、説明モデリングだけでなく予測モデリングにも使用できる。
d. ロジスティック回帰は、説明モデリングだけでなく、予測モデリングにも使用できる。

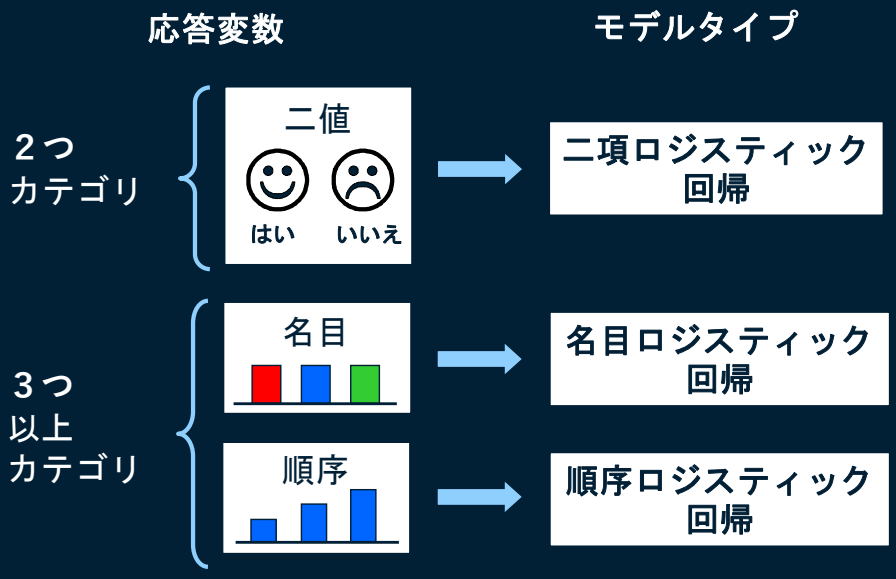
ロジスティック回帰

概要

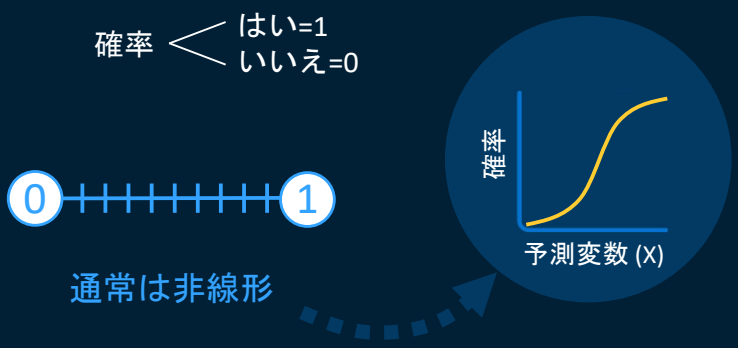


ロジスティック回帰

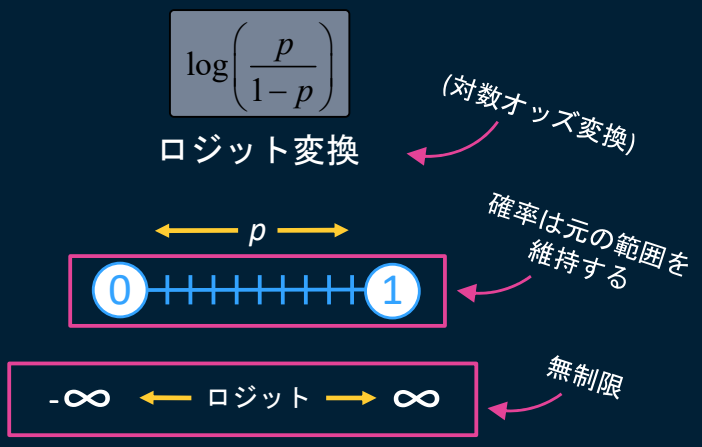
概要



ロジスティック回帰



ロジスティック回帰



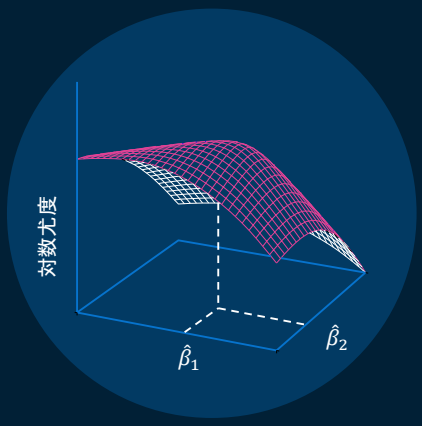
ロジスティック回帰

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

ロジスティック回帰

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

ロジスティック回帰



$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

最尤法

問題：ロジットの上限と下限

ロジットの上限と下限はどれくらいですか？

a. 下限＝0、上限＝1
b. 下限＝0、上限なし
c. 下限なし、上限なし
d. 下限なし、上限＝1

$$logit(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

回答：ロジットの上限と下限

ロジットの上限と下限はどれくらいですか？

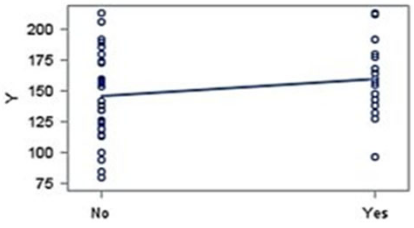
a. 下限＝0、上限＝1
b. 下限＝0、上限なし
c. 下限なし、上限なし
d. 下限なし、上限＝1

$$logit(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

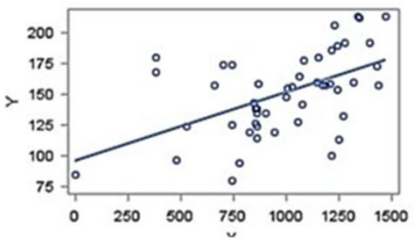
問題：ロジスティック回帰モデル

Y軸に応答変数、X軸に予測変数をプロットした場合、次のうちロジスティック回帰モデルはどれか。

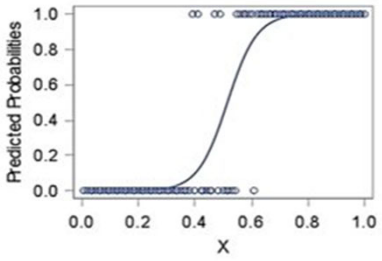
a.



b.



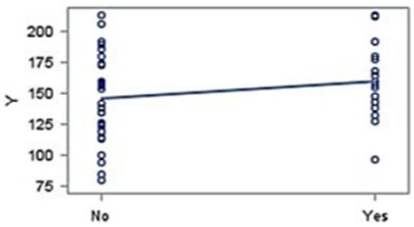
c.



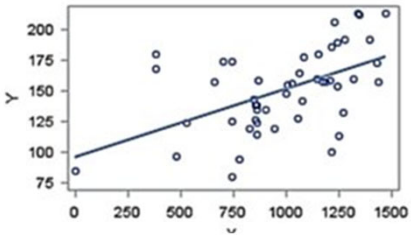
回答：ロジスティック回帰モデル

Y軸に応答変数を、X軸に予測変数をプロットした場合、次のうちロジスティック回帰モデルはどれか。

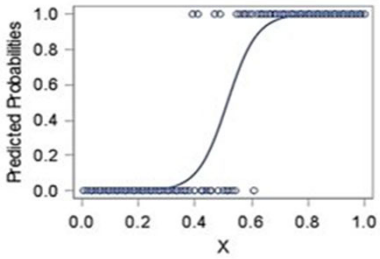
a. .



b.



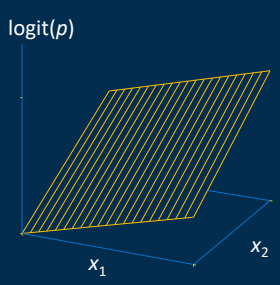
c.



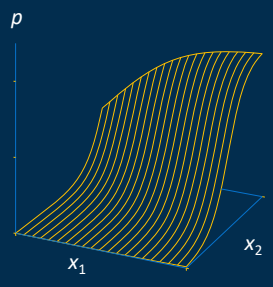
オッズ比の解釈

$$logit(\hat{p}) = log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2$$

x_2の単位変化$_2$ ⇒



ロジットにおける
$\hat{\beta}_2$変化量



オッズにおける
$100\ (exp(\hat{\beta}_2) - 1)$%変化

$\hat{\beta}_2$ = x_2が1単位変化した際の
ロジットの期待変化量

$e^{\hat{\beta}_2}$ = x_2 のオッズ比

オッズ比の解釈

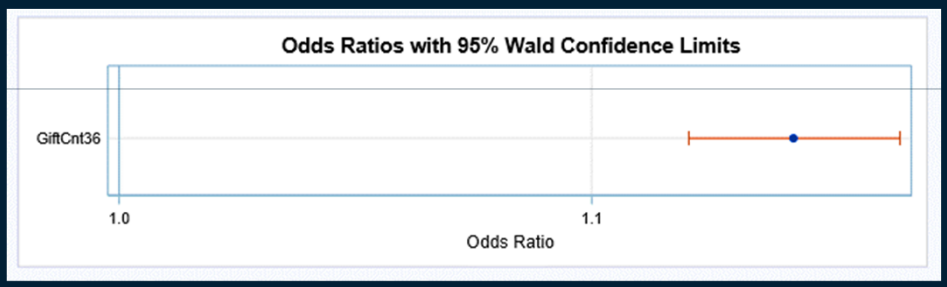
ロジスティック回帰モデル

$$\text{logit}(p) = \log(\text{odds}) = \beta_0 + \beta_1 * \text{GiftCnt36}$$

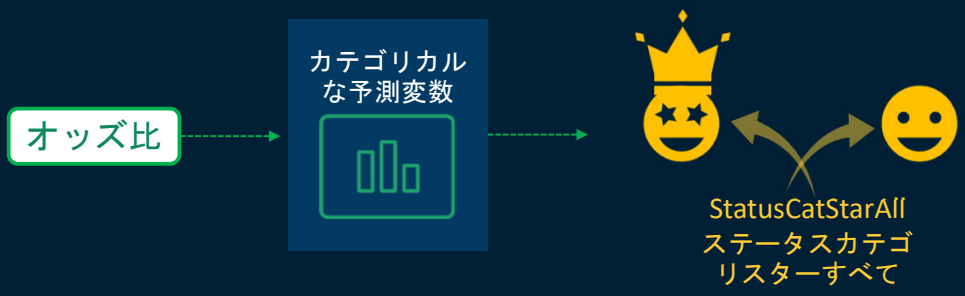
オッズ比の解釈

オッズ比の推定値				
効果	単位	点推定値	95% ウォルド信頼区間	
GiftCnt36	1	1.143	1.121	1.165

大標本 ← ウォルド信頼区間 対 プロファイル尤度信頼区間 → 小標本



オッズ比の解釈



オッズ比の解釈

ロジスティック回帰モデル

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{StatusCatStarAll}$$

スター無し – 参照レベル

$$\ln\left(\frac{p_i}{(1-p_i)}\right)$$

オッズ比の解釈

ロジスティック回帰モデル

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{StatusCatStarAll}$$

スター無し=0
スター=1

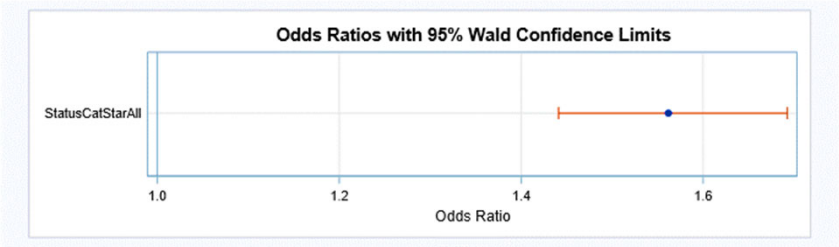
$$\text{odds}_{\text{star}} = e^{\beta_0+\beta_1}$$

$$\text{odds}_{\text{non-star}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

問題：オッズ比（設定）

オッズ比の推定値				
効果	**単位**	**点推定値**	**95% ウォルド信頼区間**	
StatusCatStarAll (スター vs スターなし)	1	1.562	1.441	1.693



問題：オッズ比

StatusCatStarAll（スターあり対スターなし）のオッズ比1.562は、以下のうちどれを示していますか？（ヒント：1/1.562=0.6402）

a. スターカテゴリーを持つ個人は、スターカテゴリーを持たない個人に比べて、回答する確率が1.562倍高かった。
b. スターカテゴリーに属する個人は、スターカテゴリーに属さない個人に比べて、回答する確率が56.2%高かった。
c. スターカテゴリーに属さない個人の回答確率は、スターカテゴリーに属する個人よりも35.97%低かった。
d. 上記のすべて。

回答：オッズ比

StatusCatStarAll（スターあり対スターなし）のオッズ比1.562は、以下のうちどれを示していますか？（ヒント：1/1.562=0.6402）

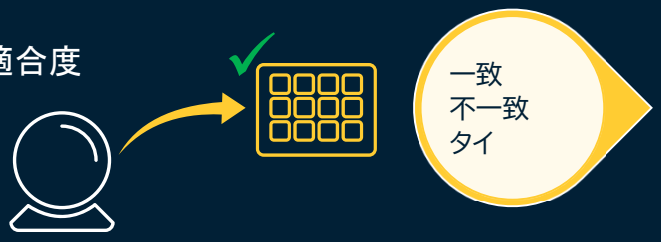
a. スターカテゴリーを持つ個人は、スターカテゴリーを持たない個人に比べて、回答する確率が1.562倍高かった。
b. スターカテゴリーを持つ個人は、スターカテゴリーを持たない個人に比べて、回答する確率が56.2%高かった。
c. スターカテゴリーに属さない個人の回答確率は、スターカテゴリーに属する個人よりも35.97%低かった。
d. 上記のすべて。

モデルの適合度の評価

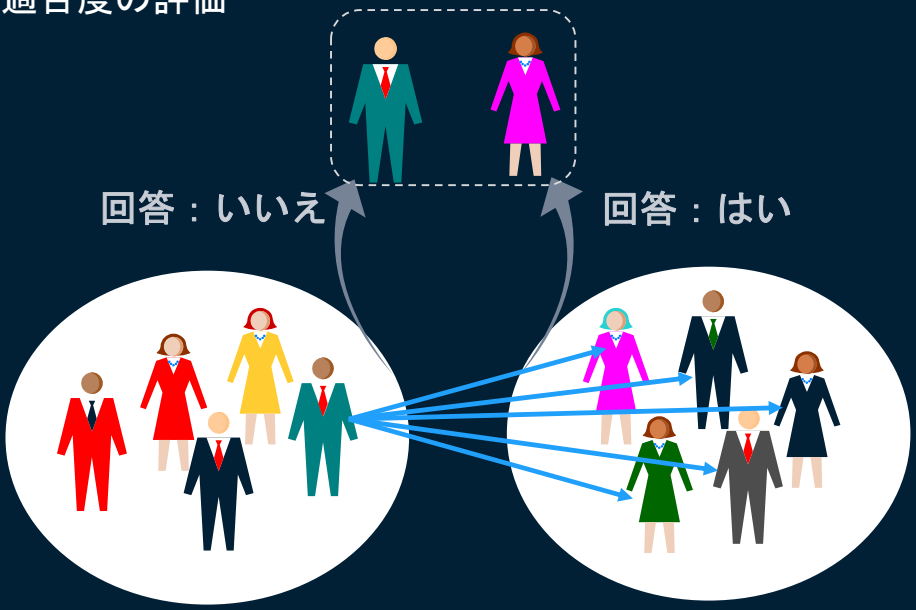
ロジスティック回帰モデルの適合度の評価

ロジスティック回帰モデル

$$\text{logit}(p_i) = \ln(\text{odds}) = \beta_0 + \beta_1 X_i$$



モデルの適合度の評価



モデルの適合度の評価

一致ペア

事象が発生した観測の予測確率が、事象が発生しなかった観測の予測確率より高い場合、そのペアは*一致*している。

回答：いいえ

P(回答) = 0.32

回答：はい

P(回答) = 0.42

実際の分類結果はモデルと一致している。
これは**一致する**ペアである。

モデル適合度の評価

不一致なペア

事象が発生した観測の予測確率が、事象が発生しなかった観測の予測確率より低い場合に、そのペアは*不一致*である。

回答：いいえ

P(回答) = 0.42

回答：はい

P(回答) = 0.32

実際の分類結果はモデルと一致しない。
これは**不一致**ペアです。

モデル適合度の評価

タイのペア

一致も不一致もどちらでもない場合、そのペアは*引き分け*となります。予測される結果の確率は同じである。

回答：いいえ

P(回答) = 0.42

回答：はい

P(回答) = 0.42

モデルはこれら2つを区別できません。
これは**タイ**ペアです。

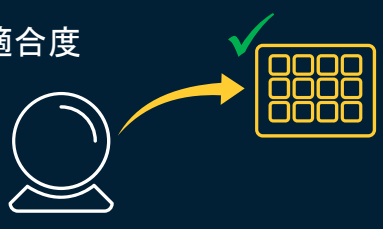
モデル適合度の評価

ロジスティック回帰モデルの適合度の評価

ロジスティック回帰モデル

$$\text{logit}(p_i) = \ln(\text{odds}) = \beta_0 + \beta_1 X_i$$

適合度



$$c = \frac{(\#Concordant\ Pairs) + \frac{1}{2}(\#Tied\ Pairs)}{Total\ Pairs}$$

問題：モデル評価のためのペア比較

モデル評価でペアを比較する際、次の記述のうち正しいものはどれか。

a. 一致するペアは、観測結果と一致する予測を示しています。
b. 不一致のペアは、予測が観測結果と一致しないことを示しています。
c. 同値なペアは、モデルが事象と非事象を十分に区別できていないことを示唆します。
d. すべて正しい。

回答：モデル評価のためのペア比較

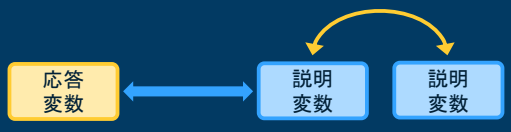
モデル評価でペアを比較する際、次の記述のうち正しいものはどれか。

a. 一致するペアは、観測結果と一致する予測を示しています。
b. 不一致のペアは、予測が観測結果と一致しないことを示しています。
c. 同値なペアは、モデルが事象と非事象を十分に区別できていないことを示唆します。
d. すべて正しい。

多重ロジスティック回帰モデル

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ \ldots + + \beta_k X_{ki}$$

交互作用

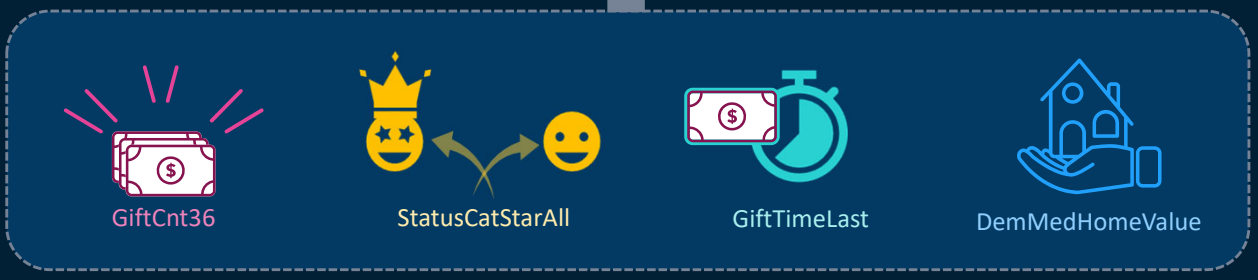


逐次選択法

- フォワード
- バックワード
- ステップワイズ

カテゴリカル予測変数

多重ロジスティック回帰モデル

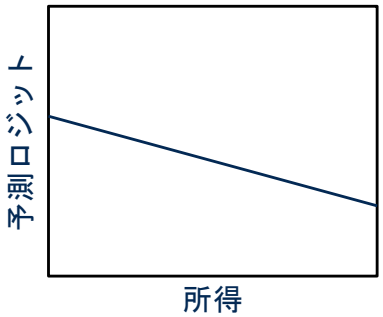


デモ：多重ロジスティック回帰モデルの推定

このデモでは、ロジスティック回帰モデルを推定する方法を示します。

問題：債務不履行と所得

以下は、顧客の所得別のローン債務不履行予測確率のロジット値を示すグラフです。

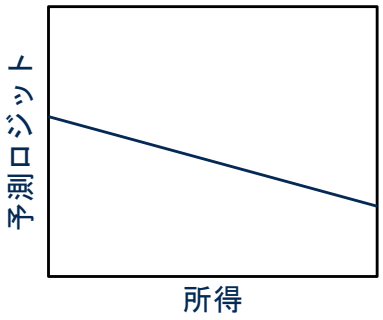


顧客の所得が増えるにつれて、ローンのデフォルト確率は低下します。

a. 正しい
b. 誤り

回答：債務不履行と所得

以下は、顧客の所得別のローン債務不履行予測確率のロジット値を示すグラフです。



顧客の所得が増えるにつれて、ローンのデフォルト確率は低下します。

a. 正しい
b. 誤り

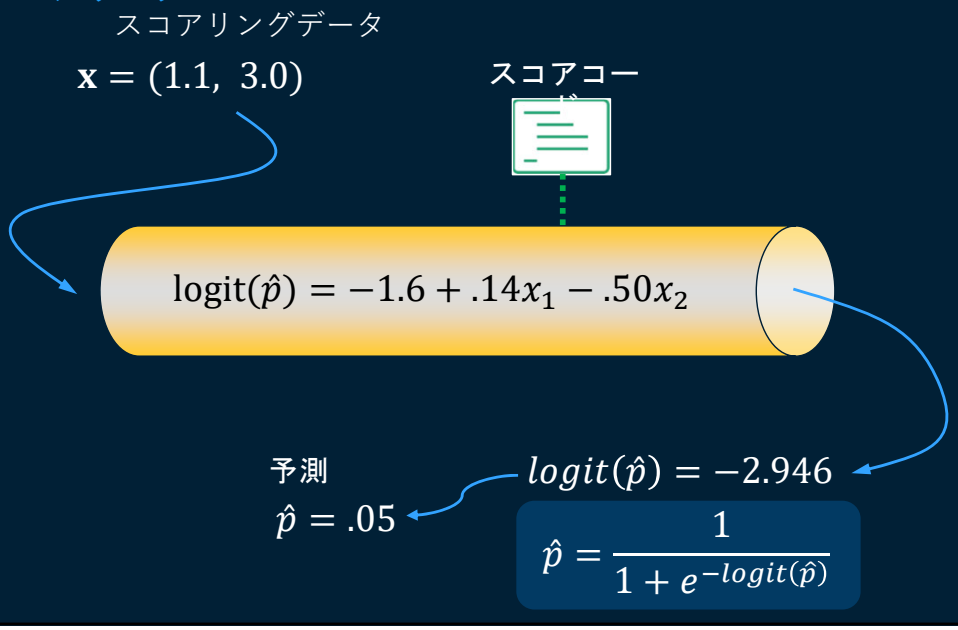
モデルのデプロイ

モデルのデプロイについて

Lesson 04, Section 04

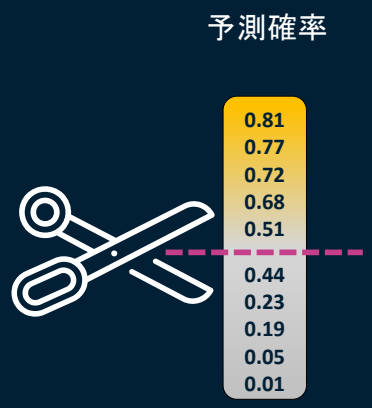
モデルのデプロイ

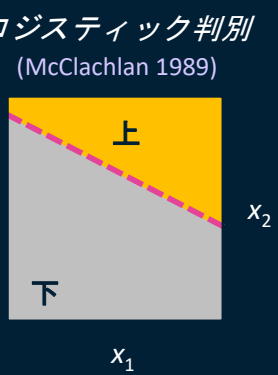
新規症例のスコアリング

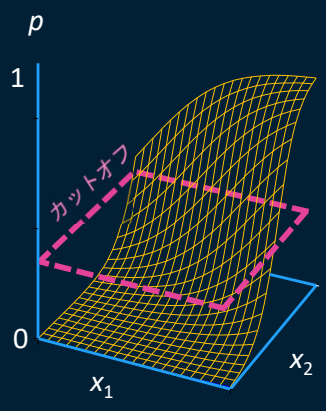


モデルの展開

ロジスティック判別



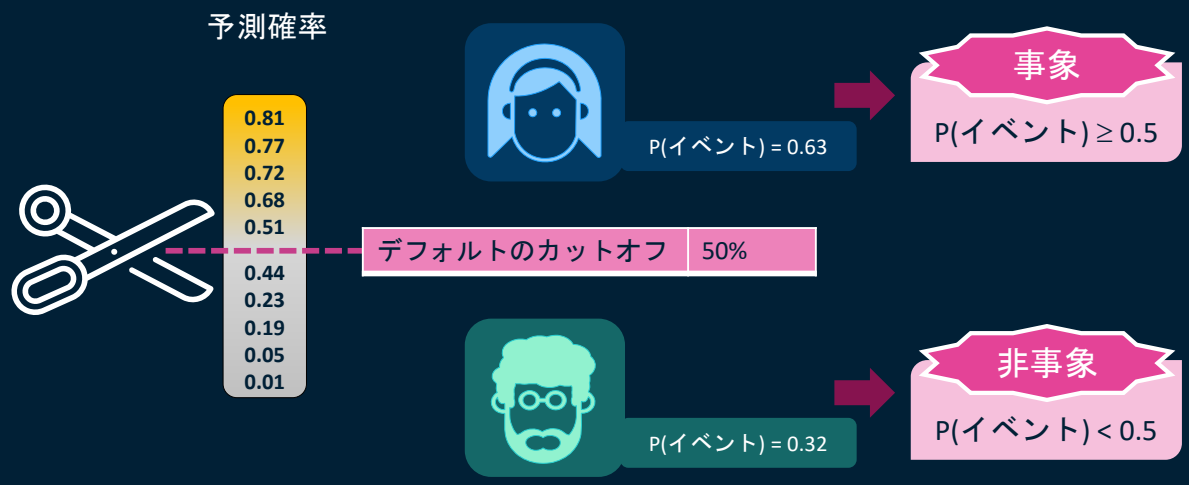




デフォルトの閾値	50%
一般的な閾値	イベント発生率

モデルのデプロイ

割り当てルール



デモ：ロジスティック回帰モデルのスコアリング

このデモでは、**SAS プログラムステップを使用して**、選択したモデルからの予測値を Score データセットに追加する方法を示しています。

問題：配分ルール

前のデモにおいて、0.45をカットオフ値とする割り当てルールを設定した場合、以下のどのケースが回答として予測されるでしょうか？（該当するものをすべて選択してください。）

	DemMedHomeValue	P_Response
a.	$114,300	0.4543596915
b.	$51,100	0.4359083123
c.	$65,000	0.4175748543
d.	$59,200	0.6024474572

回答：割り当てルール

前のデモにおいて、0.45をカットオフ値とする割り当てルールを採用するとした場合、以下のどのケースが回答として予測されるでしょうか？（該当するものをすべて選択してください。）

	DemMedHomeValue	P_Response
a.	$114,300	0.4543596915
b.	$51,100	0.4359083123
c.	$65,000	0.4175748543
d.	$59,200	0.6024474572

ご質問はありますか？

レッスンのまとめ

Lesson 5: 機械学習の統計学的基礎

sas

機械学習の概要

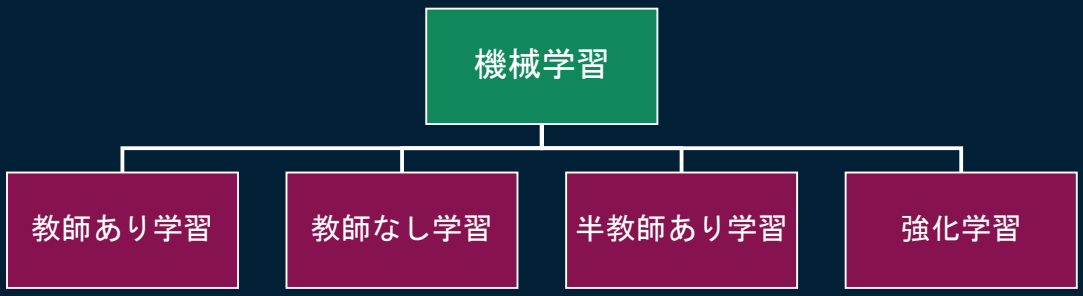
機械学習の概要

Lesson 05, Section 01

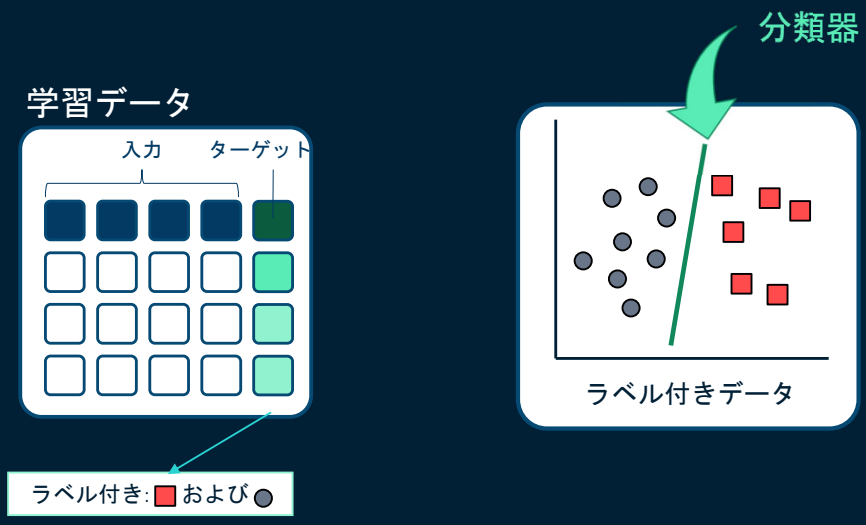
sas

機械学習

- ビジネス上の課題の性質は？
- データのタイプと量



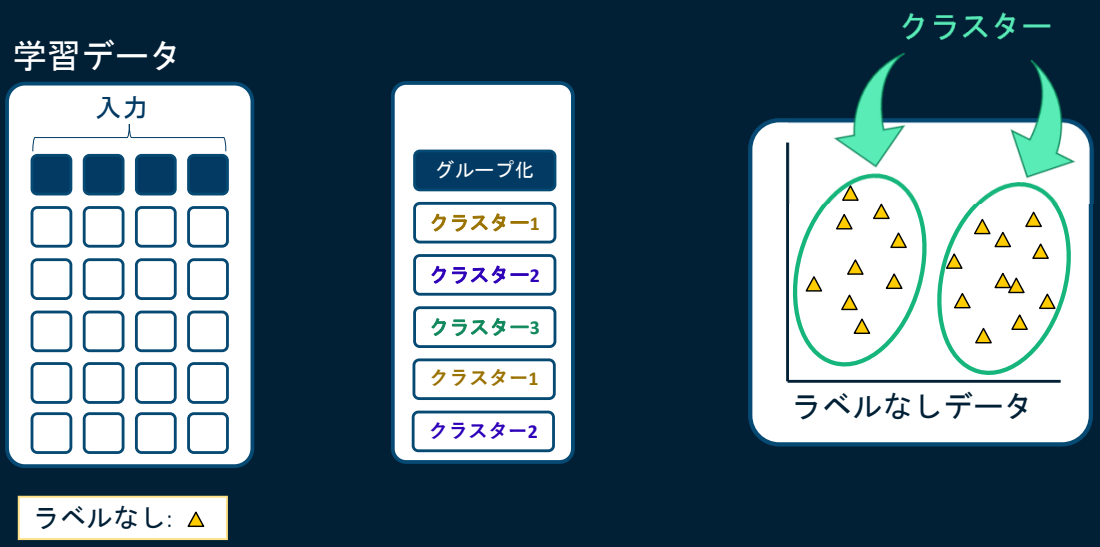
教師あり学習



教師あり学習

過去のデータに基づいて将来の出来事を予測する場合に使用される

教師なし学習



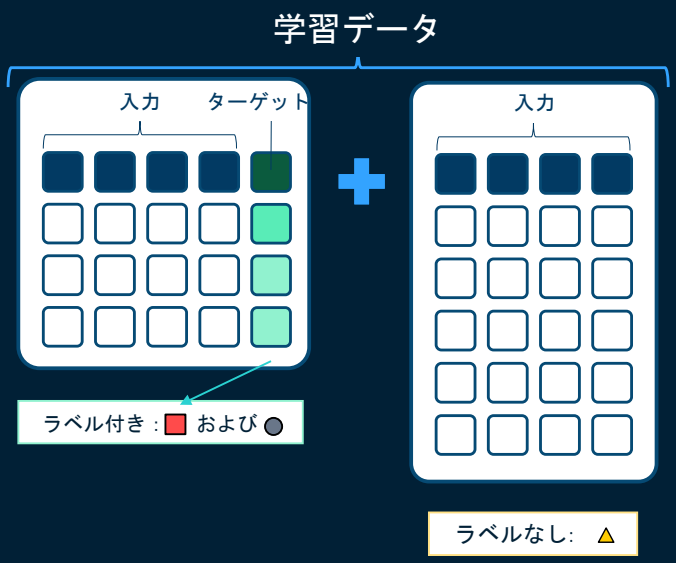
教師なし学習

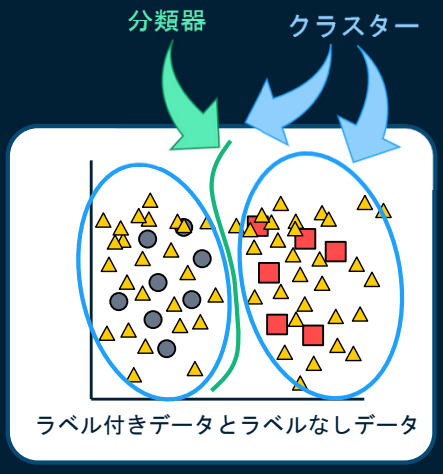
トランザクションデータに対して効果的

よく使われる手法：

- 自己組織化マップ
- 最近傍マッピング
- k-meansクラスタリング
- 特異値分解

半教師あり学習





クラスタの仮定：
同じ構造（クラスター）上の点は、同じラベルを持つ可能性が高い。

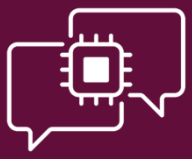
半教師あり学習

ラベル付けにかかるコストが高すぎる場合に有用

ウェブページの分類

画像認識

医療画像



自然言語処理

動作認識

強化学習

逐次的意思決定問題に有用

ロボティクス

ゲーム

ナビゲーション（自動運転車）

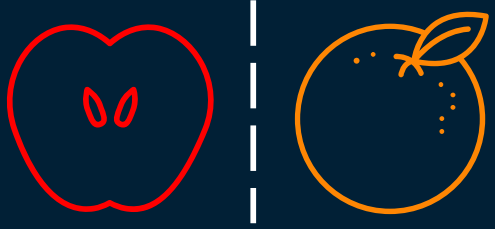
このアルゴリズムは、試行錯誤を通じて、
どの行動が最大の報酬をもたらすかを発見します。

強化学習についてさらに詳しく

機械学習はどのように機能するのでしょうか？

自動化され、
適応する

属性（独立変数）：
色、大きさ、香り、質感.....



リンゴとオレンジを見分ける

機械学習はどのように機能するのでしょうか？

自動化され、
適応する

より高度な学習を経て…

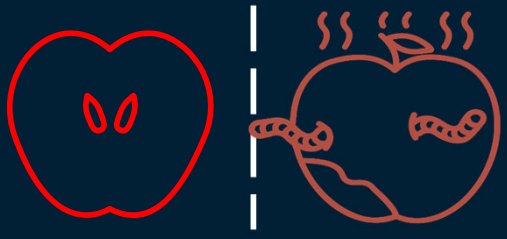
フジリンゴとグラニースミスリンゴを見分ける

機械学習はどのように機能するのでしょうか？

自動化され、
適応する

より多くの情報（柔らかさ、水分量、放出されるエチレンガスなど）を用いて

……



熟したリンゴと腐ったリンゴを見分ける

問題：機械学習の手法

機械学習の手法に関して、次の記述のうち正しいものはどれか。（該当するものすべてを選択）

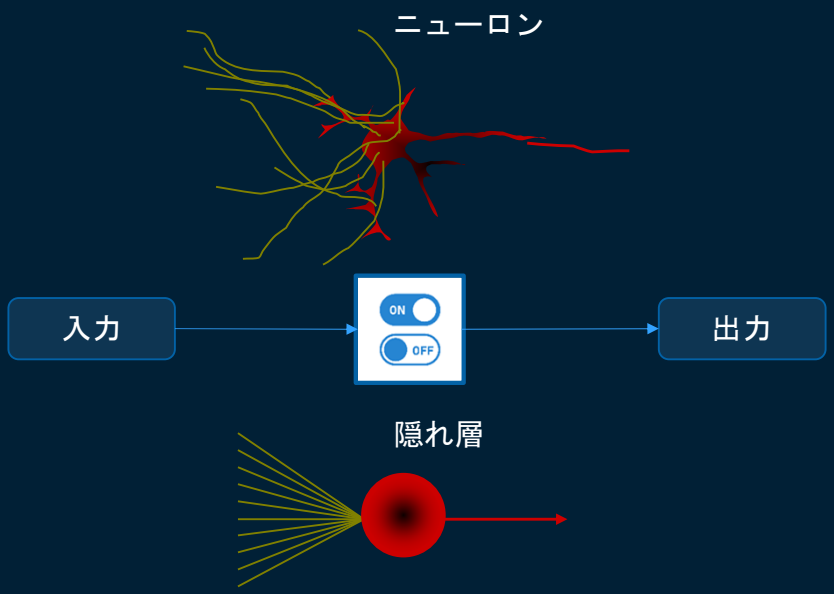
a. 教師なし学習は、入力の関連性に基づいて潜在的なパターンを発見する。
b. 教師あり学習は、入力とターゲットの関連性に基づいて出力を予測する。
c. 教師あり学習では入力データにラベルが付いていないのに対し、教師なし学習ではデータにラベルが付いている。
d. 強化学習は、試行錯誤の手法に従い、データはあらかじめ定義されていない。

回答：機械学習の手法

機械学習の手法に関して、次の記述のうち正しいものはどれか。（該当するものすべてを選択）

a. 教師なし学習は、入力の関連性に基づいて潜在的なパターンを発見する。
b. 教師あり学習は、入力とターゲットの関連性に基づいて出力を予測する。
c. 教師あり学習では入力データにラベルが付いていないのに対し、教師なし学習ではデータにラベルが付いている。
d. 強化学習は、試行錯誤の手法に従い、データはあらかじめ定義されていない。

ニューラルネットワーク



ニューラルネットワーク

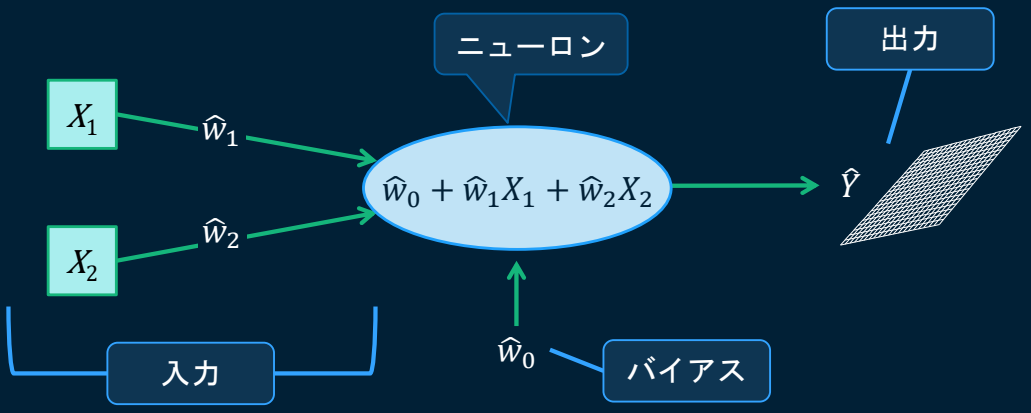
ニューロンとは何か？

ニューロン
（隠れ層）

ニューラルネットワーク

ニューラルネットワークとしての線形回帰モデル

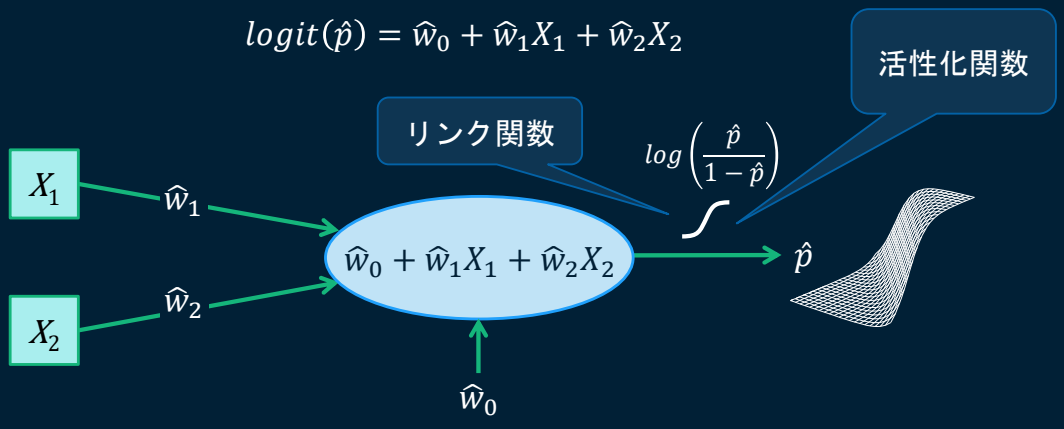
$$\hat{Y} = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2$$



ニューラルネットワーク

ニューラルネットワークとしてのロジスティック回帰モデル

$$logit(\hat{p}) = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2$$

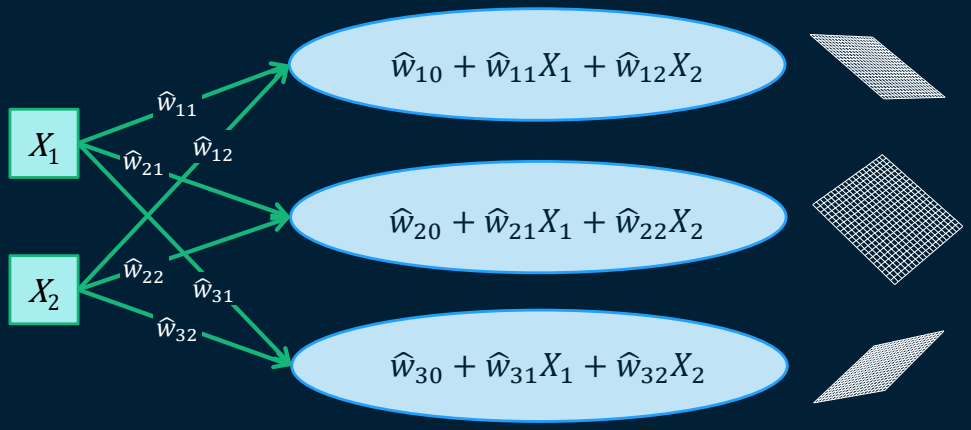


ニューラルネットワーク

汎用近似器

ニューラルネットワークモデル

線形結合関数を用いた3つのニューロン

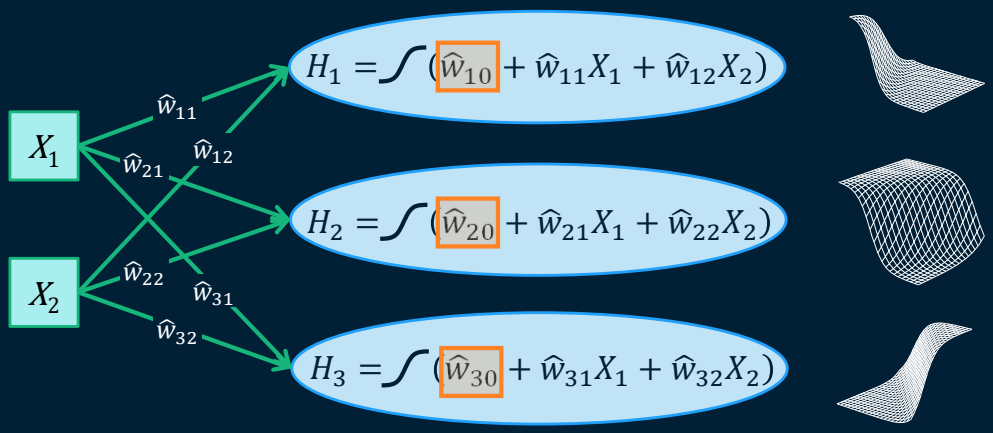


ニューラルネットワーク

線形結合を変換し、
入力として使用する

ニューラルネットワークモデル

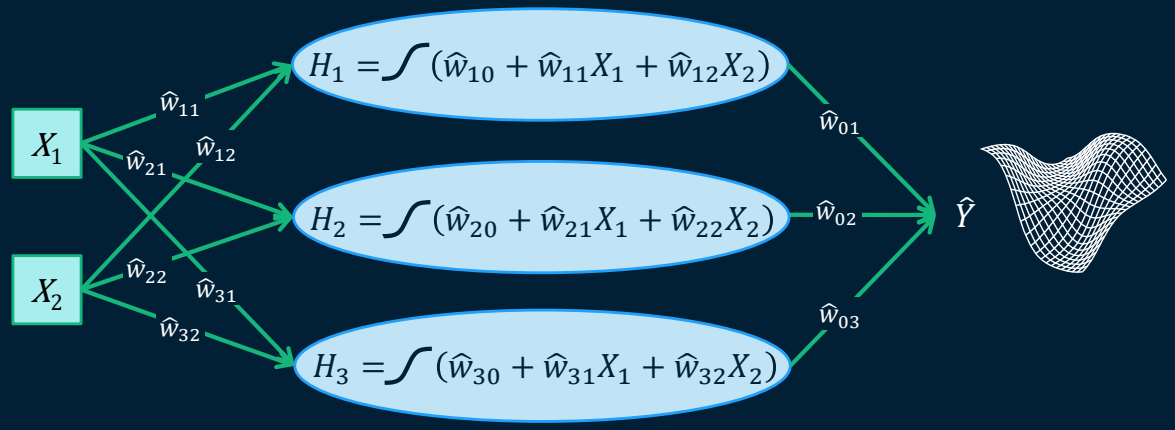
非線形活性化関数が適用された3つのニューロン



ニューラルネットワーク

ニューラルネットワークモデル

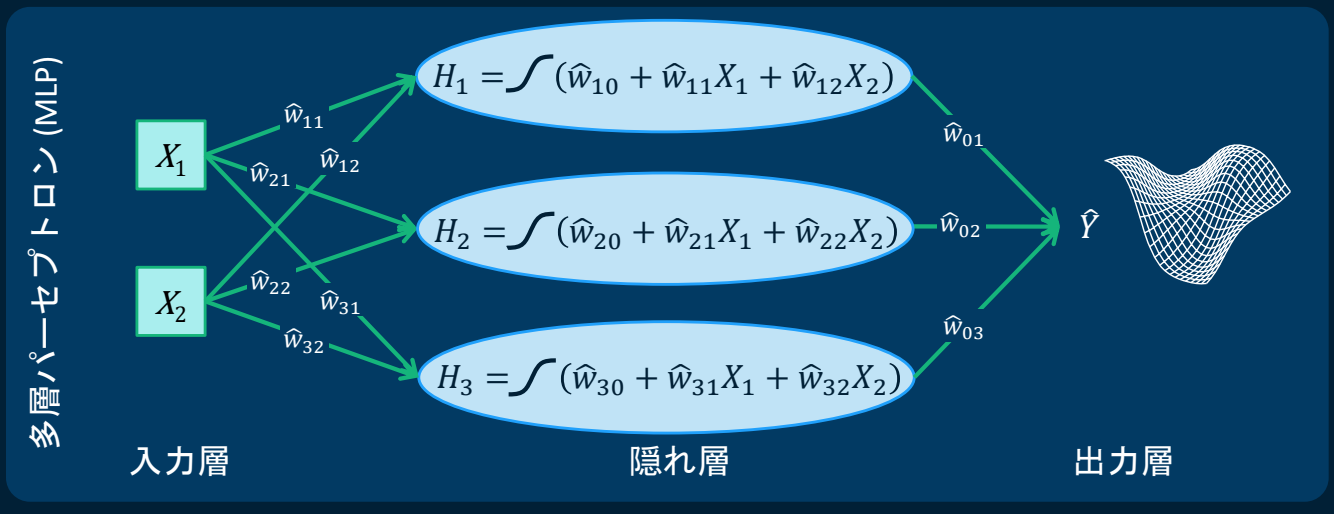
$$\hat{Y} = \hat{w}_{00} + \hat{w}_{01}H_1 + \hat{w}_{02}H_2 + \hat{w}_{03}H_3$$



ニューラルネットワーク

ニューラルネットワークモデル

$$\hat{Y} = \hat{w}_{00} + \hat{w}_{01}H_1 + \hat{w}_{02}H_2 + \hat{w}_{03}H_3$$



問題：ニューラルネットワーク

線形回帰モデルやロジスティック回帰モデルは、隠れ層を持たないニューラルネットワークと見なすことができます。

a. 正しい
b. 誤り

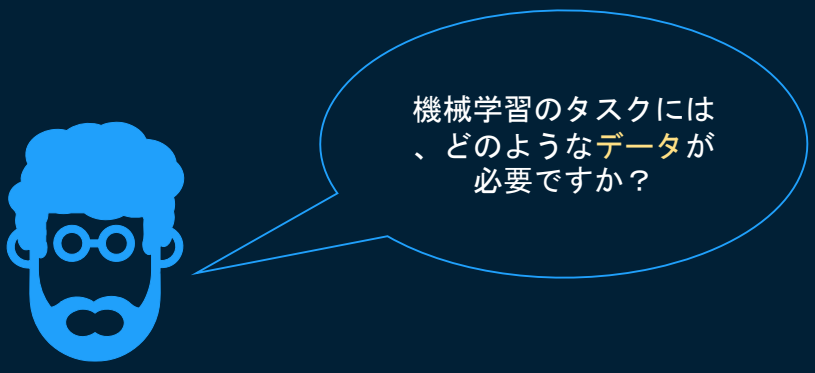
回答：ニューラルネットワーク

線形回帰モデルやロジスティック回帰モデルは、隠れ層を持たないニューラルネットワークと見なすことができます。

a. 正しい
b. 誤り

詳細情報：一般的なアルゴリズム

機械学習のためのデータ準備



機械学習のためのデータ準備

データの配置

長い・細い

Acct	Type
2133	MTG
2133	SVG
2133	CK
2653	CK
2653	SVG
3544	MTG
3544	CK
3544	MMF
3544	CD
3544	LOC

この銀行データはどのように整理すべきでしょうか？

機械学習の手法では、データはエンティティごとに1行ずつ構成されている必要があります。

短い・広い

Acct	CK	SVG	MMF	CD	LOC	MTG
2133	1	1	0	0	0	1
2653	1	1	0	0	0	0
3544	1	0	1	1	1	1

機械学習のためのデータ準備

横断データの集計

特定の時点での比較を行う

HH	Acct	Sales
4461	2133	160
4461	2244	42
4461	2773	212
4461	2653	250
4461	2801	122
4911	3544	786
5630	2496	458
5630	2635	328
6225	4244	27
6225	4165	759

HH	Acct	Sales
4461	2133	?
4911	3544	?
5630	2496	?
6225	4244	?

機械学習のためのデータ準備

縦断データの集計

経時的な比較を行う

フライトマイルは、どのように統合すべきですか？

会員ID	月	マイレージ	VIP会員
10621	1月	650	なし
10621	2月	0	いいえ
10621	3月	0	いいえ
10621	4月	250	いいえ
33855	1月	350	いいえ
33855	2月	300	いいえ
33855	3月	1200	はい
33855	4月	850	はい

機械学習のためのデータ準備

テーブルの結合

集計されたフライトマイルデータ

会員ID	マイレージ	VIP会員
10621	900	いいえ
33855	2700	はい

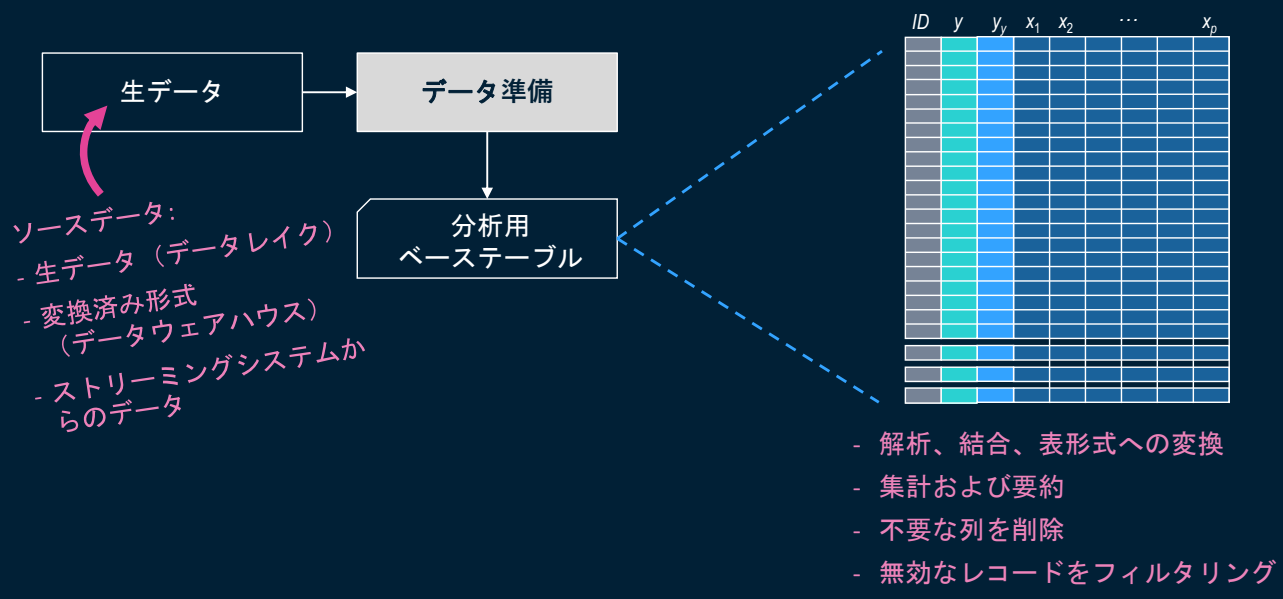
＋

人口統計データ

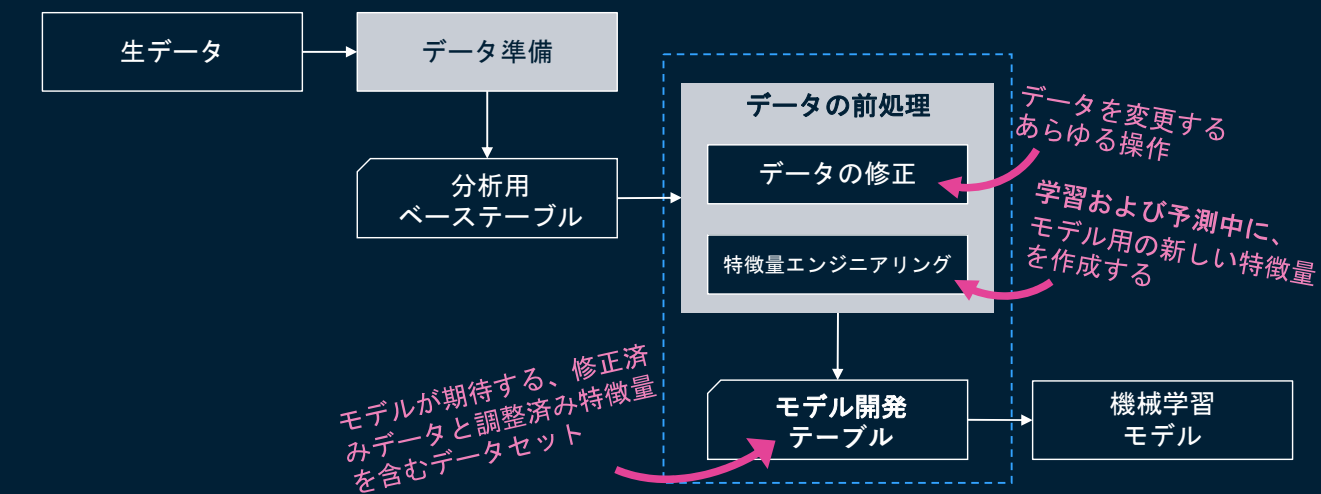
会員ID	年齢	所得水準	持ち家
10621	45	中	はい
33855	28	高	いいえ

会員ID	マイレージ	VIP会員	年齢	所得水準	持ち家
10621	900	いいえ	45	中	はい
33855	2700	はい	28	高	いいえ

機械学習のためのデータ準備



機械学習のためのデータ前処理



問題：データ準備とデータ前処理

機械学習のデータに関して、次の記述のうち正しいものはどれですか？
（該当するものをすべて選択してください。）

a. ソースデータは、事前の準備を一切行わないままの生のテーブルとして生成される。
b. データ準備を行うと、特徴量エンジニアリングなど、機械学習タスクにすぐに使える形式の分析用ベーステーブルが生成される。
c. データ前処理を行うと、機械学習モデルへの適用に適したモデル開発用テーブルが生成される。
d. 機械学習データは一般的にクリーンであり、準備や処理を必要としない。

回答：データ準備とデータ前処理

機械学習のデータに関して、次の記述のうち正しいものはどれですか？
（該当するものをすべて選択してください。）

a. ソースデータは、事前の準備を一切行わないままの生のテーブルとして生成される。
b. データ準備を行うと、特徴量エンジニアリングなど、機械学習タスクにすぐに使える形式の分析用ベーステーブルが生成される。
c. データ前処理を行うと、機械学習モデルへの適用に適したモデル開発用テーブルが生成される。
d. 機械学習データは一般的にクリーンであり、準備や処理を必要としない。

機械学習モデルのためのデータ前処理

機械学習モデルのためのデータ前処理の理解

Lesson 05, Section 02

データの課題とモデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力のばらつきへの対処
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

データに関する課題
（このセクション）

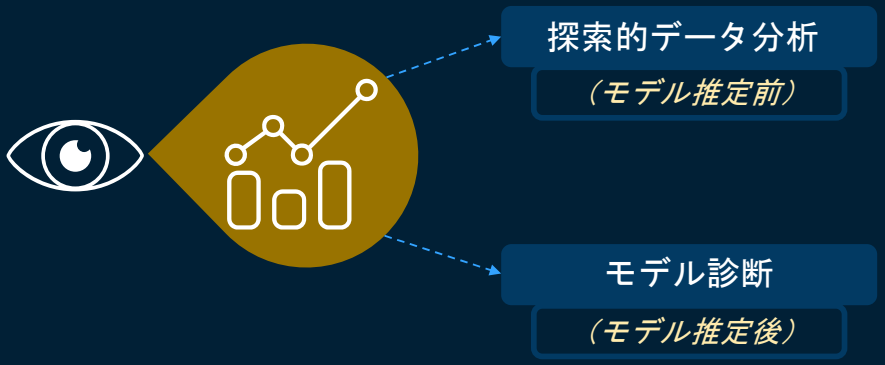
モデリング上の課題
（次のセクション）

データの可視化

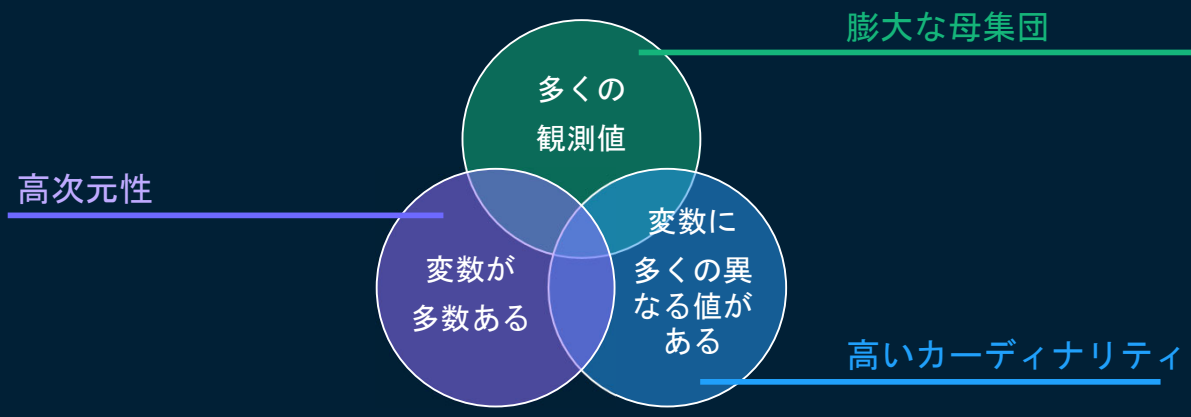
- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データのばらつきへの対応
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

データの可視化

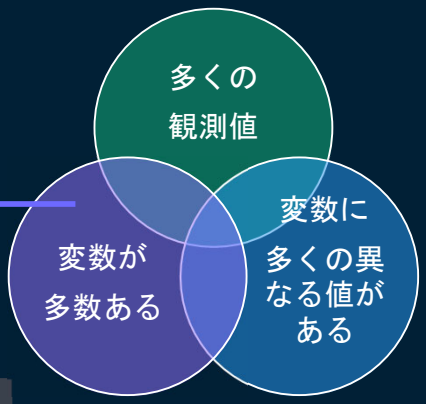
「百聞は一見に如かず。」



ビッグデータの可視化における課題



ビッグデータの可視化における課題

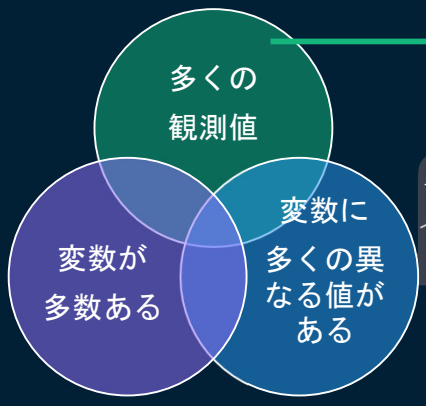


高次元性

- 多次元空間
- 変数間の関連性の提示
- 変数の相対的な重要性の提示

➢ 適切な視覚的表現を選択することが解決策となる。

ビッグデータの可視化における課題

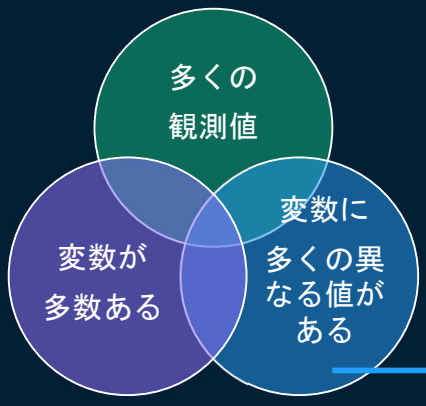


膨大な母集団

- 処理時間の長期化
- 視覚的なごちゃごちゃ

サンプリング、フィルタリング、およびビニングが選択肢となり得る。

ビッグデータの可視化における課題



高いカーディナリティ

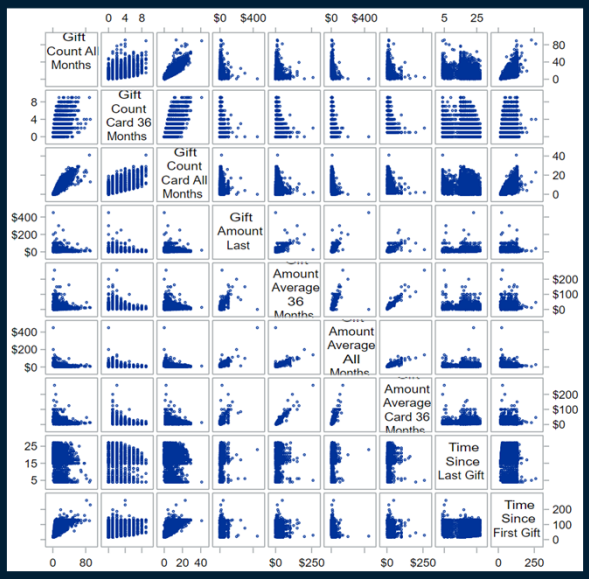
- 視覚的に非常に賑やかな
- ビジュアルのスペース不足

値を統合することも一つの解決策となる。

可視化の例

従来の統計プロット

散布図マトリックス



変数や観測値の数が増えるにつれて…

…プロットがごちゃごちゃした感じになることがある

従来のツールでは、かなりの時間がかかる可能性がある

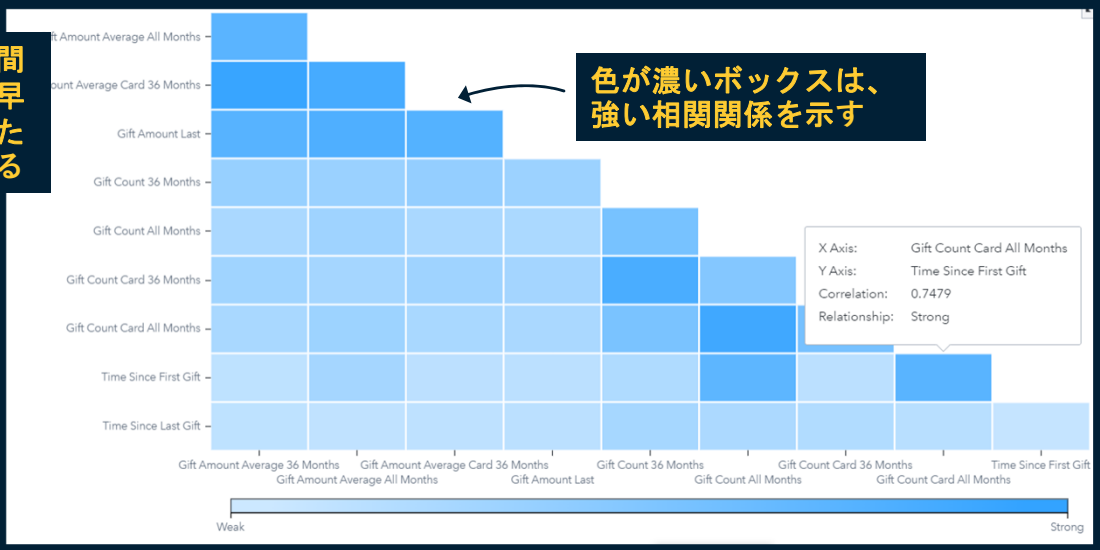
可視化の例

従来の統計プロットにとらわれない発想を！

相関行列

多くの関係ペアを表示するには不十分な場合がある

多数の変数間の関係を素早く把握するために使用する



色が濃いボックスは、強い相関関係を示す

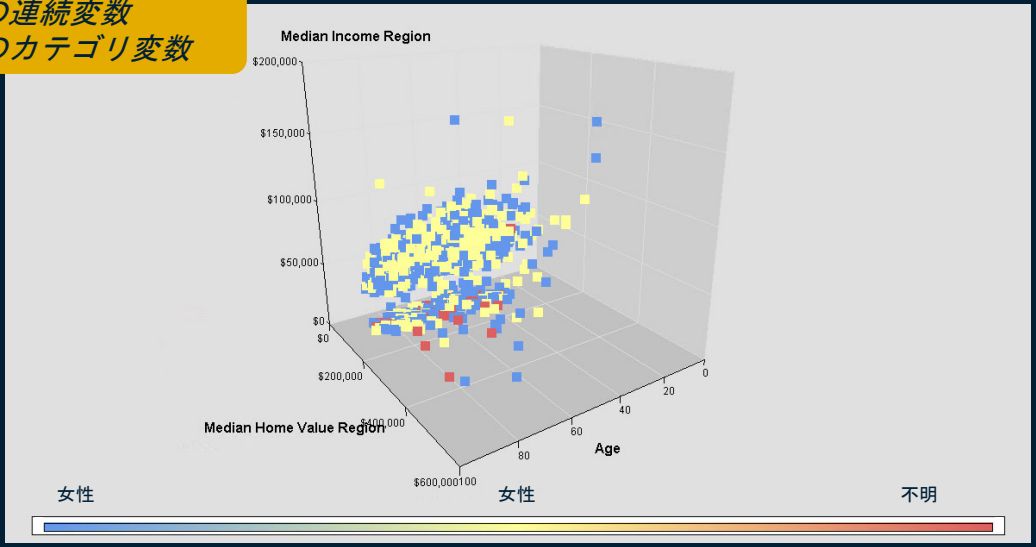
可視化の例

従来の統計プロット

3つの連続変数
1つのカテゴリ変数

色分けされた3D散布図

解釈が難しい場合があります！

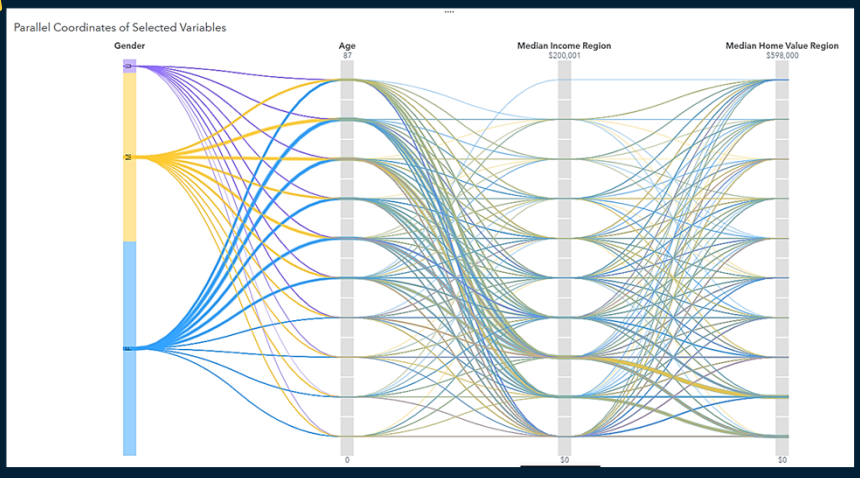


可視化の例

従来の統計プロットを超えた発想を！

多変量数値データの分析に有効

平行座標プロット



異なる規模の複数の数値変数を比較するために使用

- 異なる大きさや
- 異なる測定単位を持つ

可視化の例についてさらに詳しく

問題：データ可視化の課題

以下の用語のうち、*変数に多くの異なる値*が含まれている状況を表すものはどれですか？

a. 膨大な母集団
b. 高次元性
c. 高いカーディナリティ
d. 変動が小さい

回答：データ可視化の課題

以下の用語のうち、*変数に多くの異なる値*が含まれている状況を表すものはどれですか？

a. 膨大な母集団
b. 高次元性
c. 高いカーディナリティ
d. 変動が小さい

不適切なデータ：誤り、欠損値、外れ値

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データのばらつきへの対処
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

不適切なデータ：誤り、欠損値、外れ値

当座預金口座 / 当座預金口座数 / 平均日次残高 / 残高不足 / 確実な口座振込 / 普通預金口座 / 期末残高

cking	#cking	ADB	NSF	dirdep	SVG	bal
Y	1	468.11	1	1876	Y	1208
Y	1	68.75	0	0	Y	0
Y	1	212.04	0	6		0
	.	.	0	0	Y	4301
y	2	585.05	0	7218	Y	234
Y	1	47.69	2	1256		238
	1	4687.7	0	0		0
Y	.	.	1	0	Y	1208
	.	.	.	1598		0
	1	0.00	0	0		0
Y	3	89981.12	0	0	Y	45662
Y	2	585.05	0	7218	Y	234

問題： エラー

以下のデータの中で、間違っていると思われる点や、単に場違いに思える点はありますか？ここには決して「間違った答え」はありません。どんなに些細なことでも、調査すべきことは何でしょうか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

考えられる回答： エラー

1列目 [R5C1] に小文字の「y」があるのに対し、他はすべて大文字の「Y」です。

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

考えられる回答：エラー

6行目のC3～C7の値は、他の値と一致していません。

	C1	C1	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

考えられる回答：エラー

SVG列において、対応するbal列の数値が0以外の場合、その行は空白になります [R6C6]。

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

考えられる回答：エラー

5行目と12行目は、R5C1の小文字の「y」を除いて同じです。これらは、誰かが「y」を修正したものの、元の行を削除し忘れたという、同じ観測値なのでしょうか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

考えられる回答：エラー

NSF欄の[R6C4]にある「2」は何を意味しますか？NSF（資金不足）は二進法ですか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

エラー

エラーの種類

- 無効な値
- 不可能な値の組み合わせ
- 不整合なコーディング
- 不適切な値
- 重複レコード
- その他…

エラーの検出には、**粘り強さ**、**創造性**、そして**専門知識**が必要です！

デモ：データの変更と修正

このデモでは、欠損値のプレースホルダー値を、実際の欠損値に置き換える方法を示します。

問題：欠損値

欠損値はありますか？

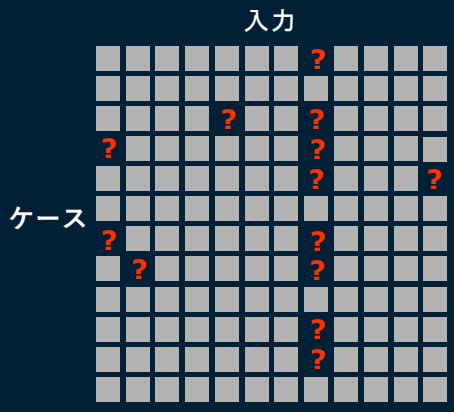
	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

回答：欠損値

欠損値はありますか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

欠損データ

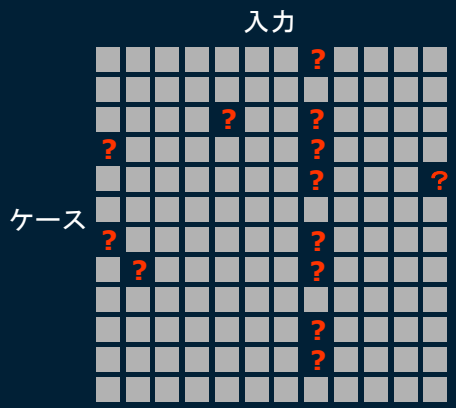


- 欠損データの問題
- 欠損データの分析戦略
- 欠損データのメカニズム
- 欠損データの原因

（自習）

データ欠落のメカニズム

欠損データの問題



欠損値によって引き起こされる主な問題：

- パラメータのバイアス
  - 過大評価または過小評価
- 推論誤差
  - 予測誤差または仮説検定の誤差

欠損データの問題（Newman 2014）

欠損データの問題についてさらに詳しく

データ欠落の原因

問題：欠損値の表現

SASでは、数値変数の欠損値はどのように表現されますか？

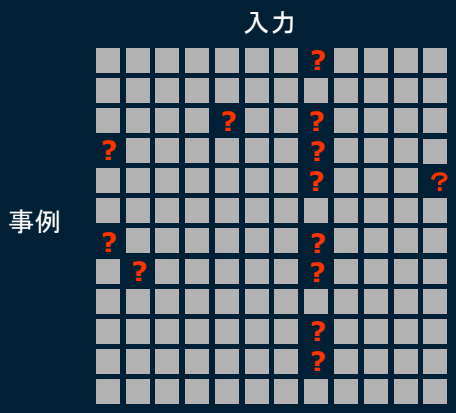
a. 空白（‘ ’）
b. ピリオド（ . ）
c. コンマ (,)
d. セミコロン (;)

回答：欠損値の表現

SASでは、数値変数の欠損値はどのように表現されますか？

a. 空白（‘ ’）
b. ピリオド（ . ）
c. コンマ (,)
d. セミコロン (;)

欠損データに対する分析戦略

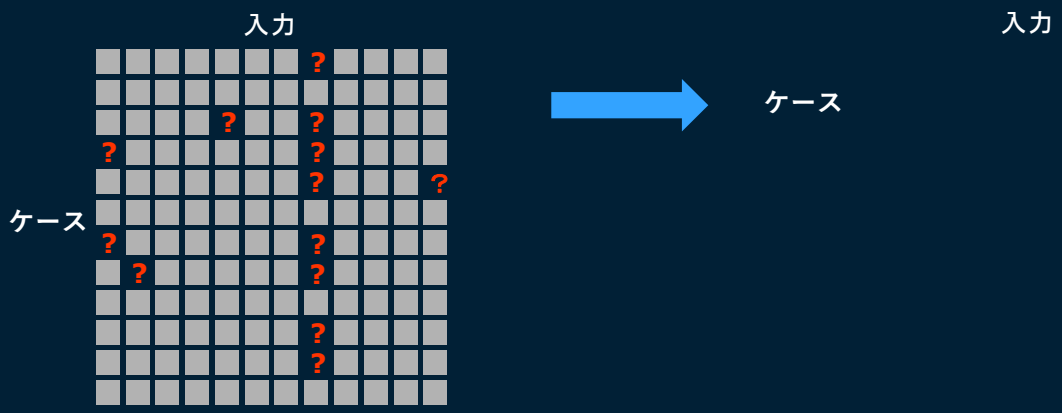


欠損値の取り扱いに関する分析戦略

- 完全ケース分析（CCA）
- 欠損値補完
- 欠損値の指標

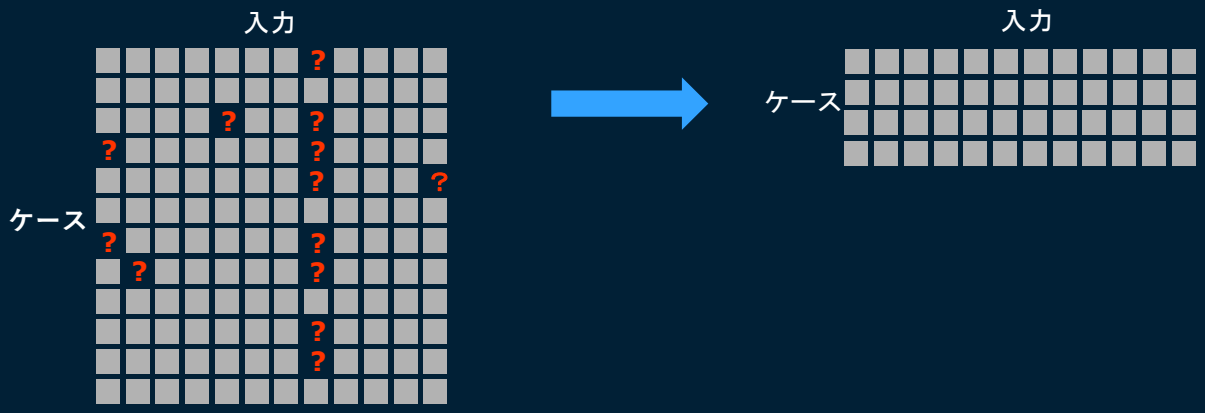
欠損データに対する分析戦略

完全ケース分析（CCA）



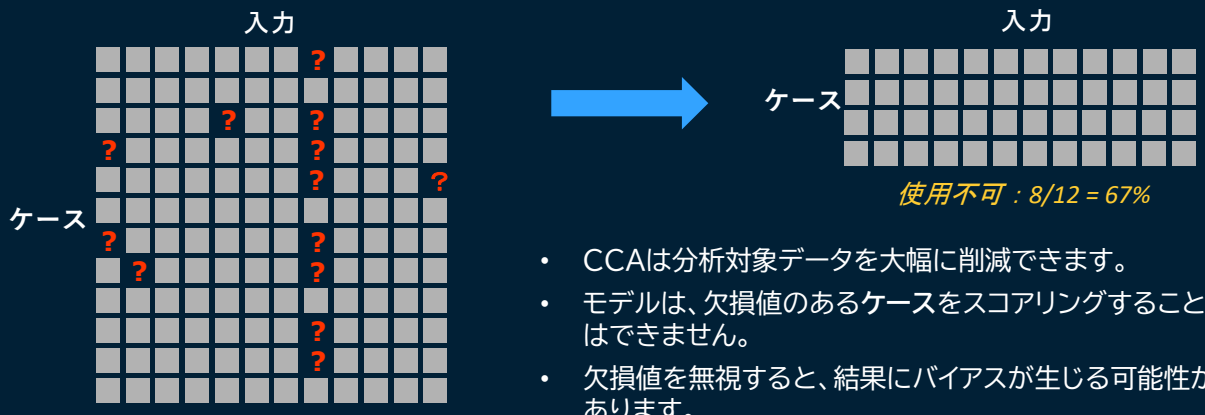
欠損データに対する分析戦略

完全ケース分析（CCA）



欠損データに対する分析戦略

完全ケース分析（CCA）



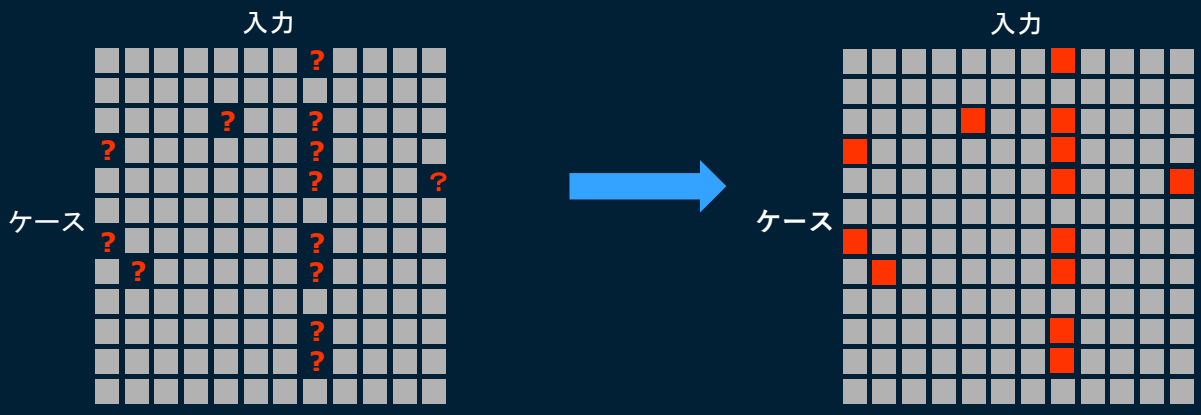
欠損：13/144 = 9%

使用不可：8/12 = 67%

- CCAは分析対象データを大幅に削減できます。
- モデルは、欠損値のある**ケース**をスコアリングすることはできません。
- 欠損値を無視すると、結果にバイアスが生じる可能性があります。
- ツリーベースのモデルは欠損値に対応しています。

欠損データに対する分析戦略

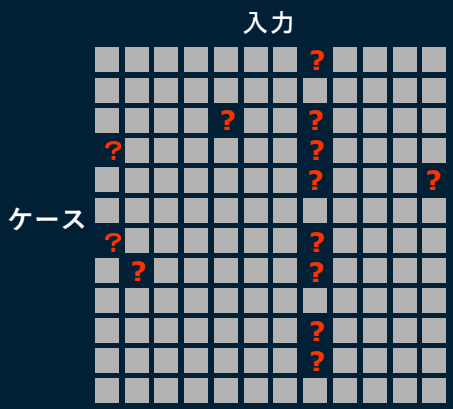
補完



- 固定値補完（合成分布）
- オーダーメイド値補完（推定）
- クラスター補完

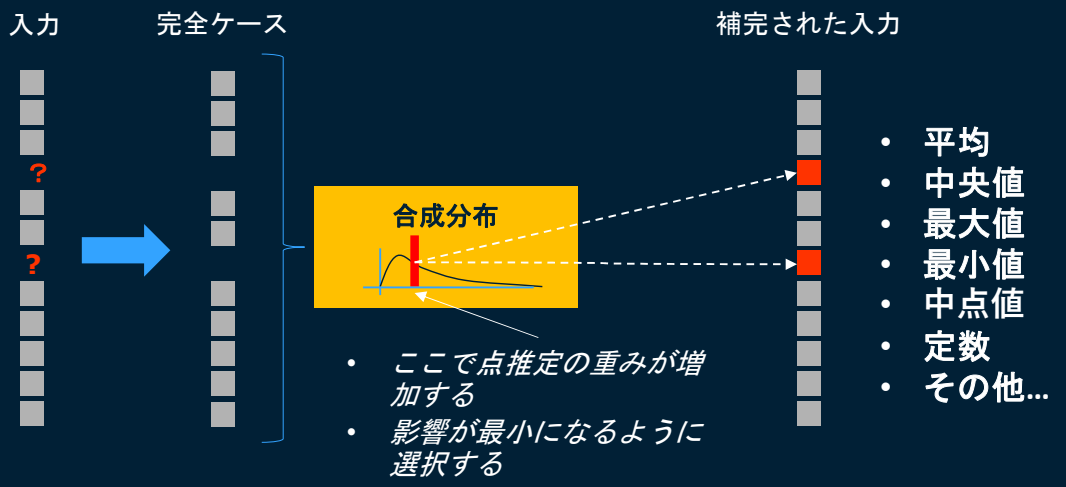
欠損データに対する分析戦略

固定値補完



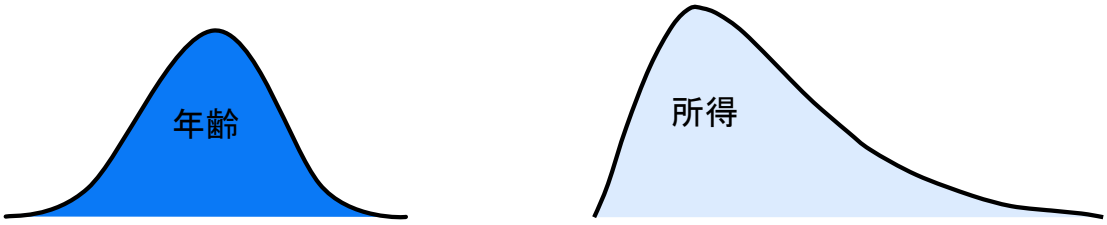
欠損データに対する分析戦略

固定値補完



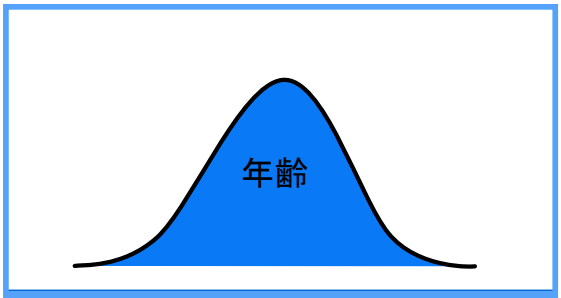
問題：平均値補完か中央値補完か

年齢と所得は、欠損値がある2つの変数です。平均値による補完または中央値による補完のいずれかを使用できますが、それぞれ1回のみ使用可能です。変数の分布は以下の通りです。どちらを使用し、その理由は何ですか？

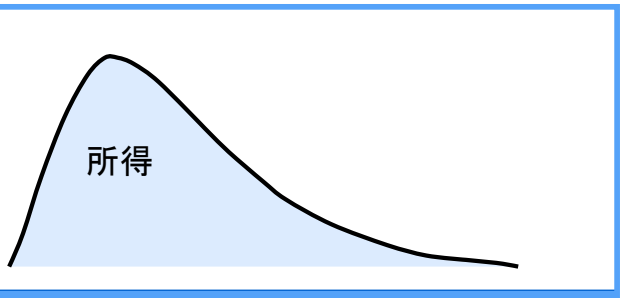


回答：平均値または中央値による補完

年齢と所得は、欠損値がある2つの変数です。平均値による補完または中央値による補完のいずれかを使用できますが、それぞれ1回のみ使用可能です。変数の分布は以下の通りです。どちらを使用し、その理由は何ですか？



年齢変数はほぼ正規分布しているため、**平均値補完が**有効です。とはいえ、正規分布データの場合、**平均値と中央値は一致します**。



データセットに**外れ値が多数含まれる**場合は、平均値ではなく**中央値を使用することが**推奨されます。したがって、所得変数は右偏りしているため、**中央値補完法**が用いられます。

問題：欠落した株価

証券取引所で最低価格に達し、取引が停止された場合、欠落している株価は次のうちどれで置き換えることができますか？

a. 平均
b. 中央値
c. 最小値
d. 最大値

回答：欠落した株価

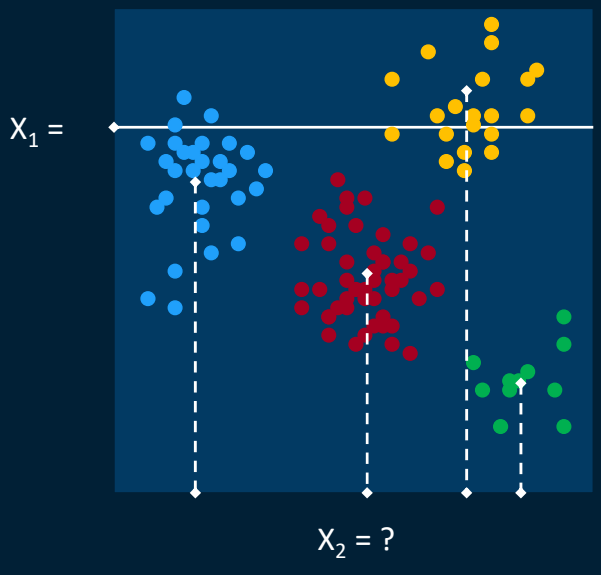
証券取引所で最低価格に達し、取引が停止された場合、欠落している株価は次のうちどれで置き換えることができますか？

a. 平均
b. 中央値
c. 最小値
d. 最大値

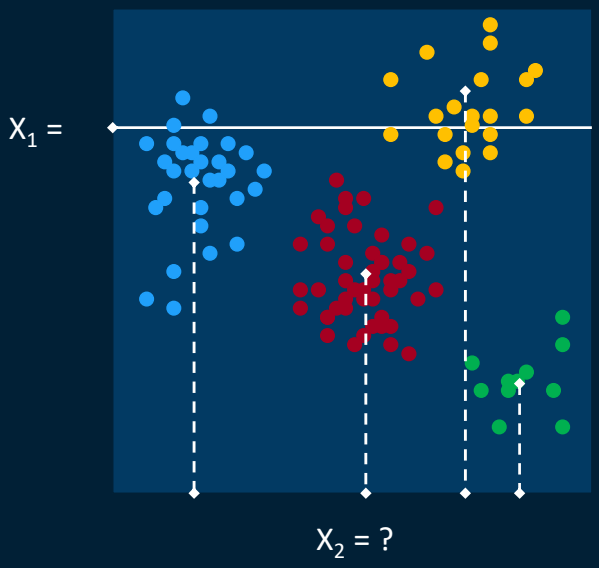
クラスター補完

クラスター平均法による補完：

1. ケースを比較的均質なグループにクラスタリングする。
2. 各グループ内で平均値補完を行う。
3. 複数の欠損値を持つ新規ケースについては、欠損のないすべての変数において最も近いクラスター平均を使用する。



クラスター補完

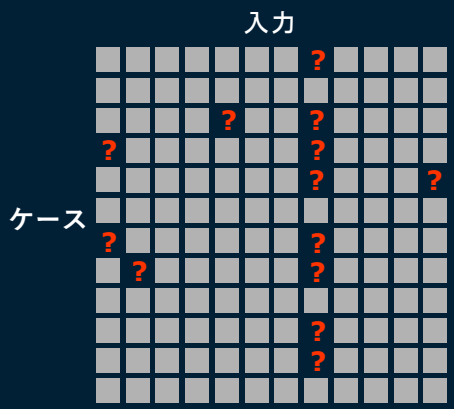


クラスター平均補完:

- 他の入力変数に依存する欠損に対応できる。
- クラスターごとに異なる値が補完される。
- 各クラスター内では、補完値は同一または固定される。

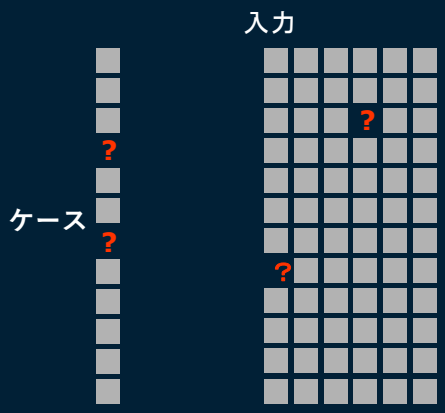
テーラードバリュー補完

欠損値がある各ケースに対して、カスタマイズされた補完を行う

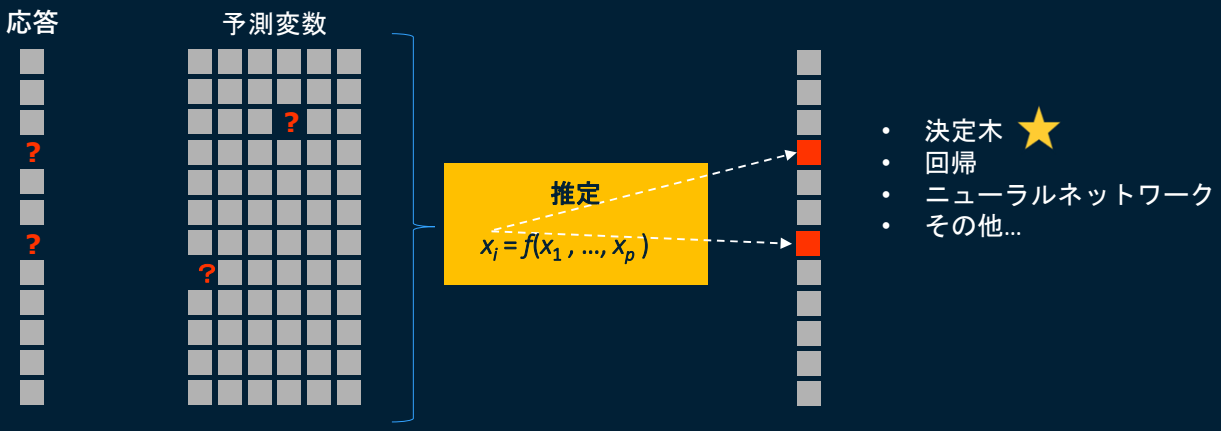


テーラードバリュー補完

欠損値がある各ケースに対して、
カスタマイズされた補完を行う



テーラードバリュー補完



- 決定木 ★
- 回帰
- ニューラルネットワーク
- その他...

問題：補完法

他の入力変数に依存する欠損値に対処する必要がある場合、以下のどの補完法を使用できますか？（該当するものをすべて選択してください）

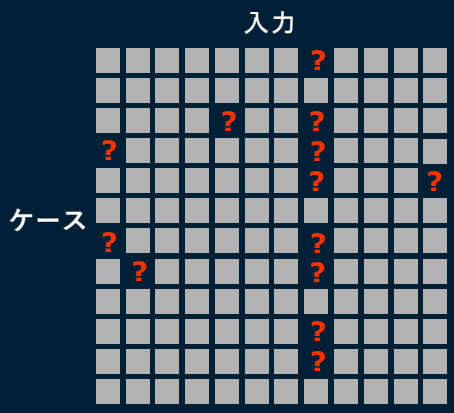
a. 固定値補完
b. テーラードバリュー補完
c. クラスター補完

回答：補完法

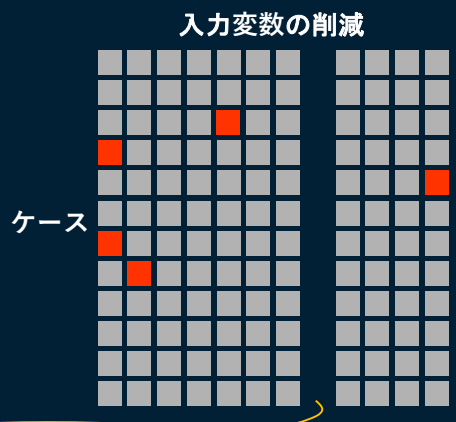
他の入力変数に依存する欠損値に対処する必要がある場合、以下のどの補完法を使用できますか？（該当するものをすべて選択してください）

a. 固定値補完
b. テーラードバリュー補完
c. クラスター補完

過度な欠損



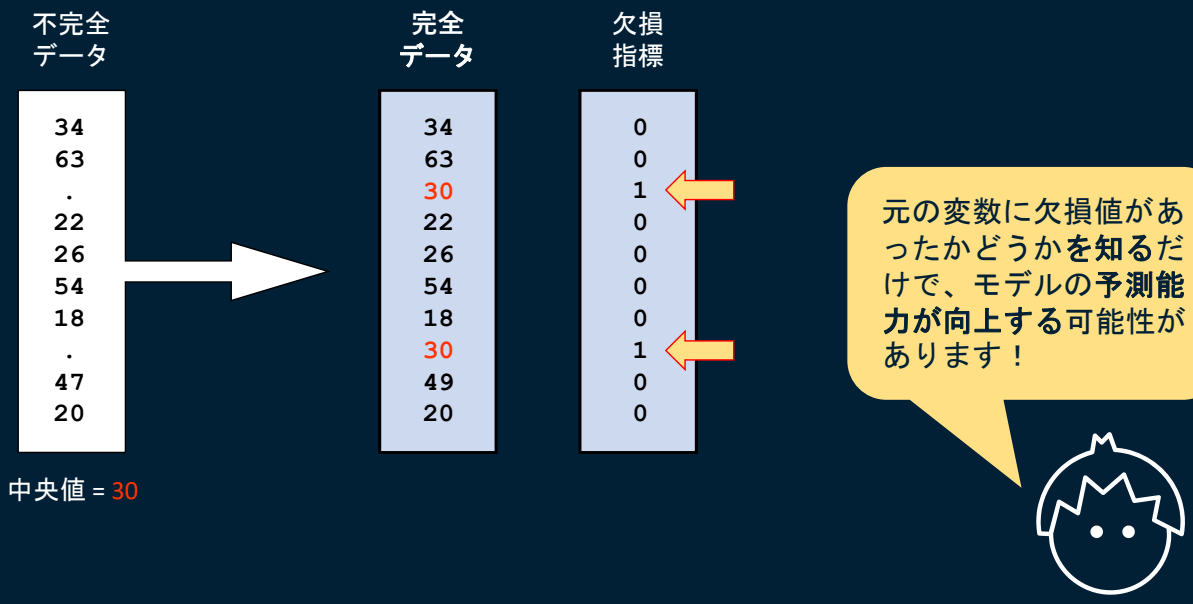
過度な欠損



変数の60～70%でデータが欠損している場合は、
その変数を削除することを検討してください。

重要な変数である場合は、
他の変数と入れ替えてください
。

欠損値の指標



デモ：欠損値の処理

このデモでは、合成データの値を補完する方法を説明します。

問題：欠損値の補完

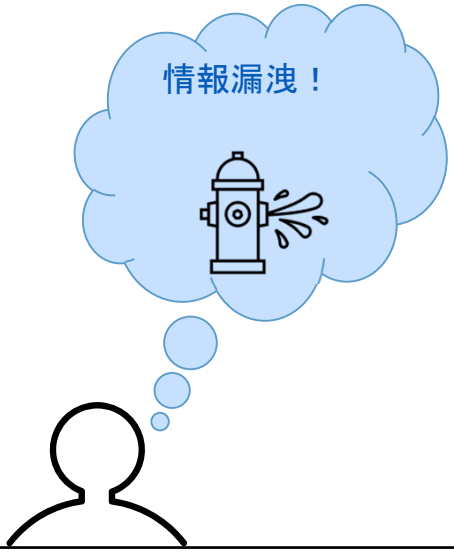
学習データにおける入力変数の平均が35で、検証データにおける同じ変数の平均が39の場合、欠損値の補完にはどちらを使用しますか？該当するものをすべて選択してください。

a. 学習データには35、検証データには39
b. 学習データと検証データの両方で35
c. 学習データと検証データの両方で39
d. 評価データには35

回答：欠損値の補完

学習データにおける入力変数の平均が35で、検証データにおける同じ変数の平均が39の場合、欠損値の補完にはどちらを使用しますか？該当するものをすべて選択してください。

a. 学習データには35、検証データには39
b. 学習データと検証データの両方で35
c. 学習データと検証データの両方で39
d. 評価データには35



問題：外れ値

外れ値とは、異常なデータ値のことです。ここには外れ値はありますか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

回答：外れ値

外れ値とは、異常なデータ値のことです。ここには外れ値はありますか？

	C1	C2	C3	C4	C5	C6	C7
	cking	#cking	ADB	NSF	dirdep	SVG	bal
R1	Y	1	468.11	1	1876	Y	1208
R2	Y	1	68.75	0	0	Y	0
R3	Y	1	212.04	0	6		0
R4		.	.	0	0	Y	4301
R5	y	2	585.05	0	7218	Y	234
R6	Y	1	47.69	2	1256		238
R7		1	4687.7	0	0		0
R8	Y	.	.	1	0	Y	1208
R9		.	.	.	1598		0
R10		1	0.00	0	0		0
R11	Y	3	89981.12	0	0	Y	45662
R12	Y	2	585.05	0	7218	Y	234

外れ値とは何か？

- 外れ値とは、他のデータとは明らかに異なるデータポイントのことです。

- **外れ値は、以下の方法で検出可能**
  - 視覚化手法：箱ひげ図、散布図など
  - 数学的関数：四分位範囲（IQR）、Zスコアなど
  - その分野の専門家への相談

外れ値と機械学習モデル

- 外れ値がすべての機械学習モデルにとって問題となるわけではありません。
- 線形回帰やロジスティック回帰のような従来のモデルは、外れ値の影響を受けやすく、その結果、モデルの精度が低下することがあります。
- また、外れ値はニューラルネットワークで使用される一部の活性化関数にとっても問題となります。
- 他のいくつかの機械学習モデルでは、外れ値によって学習時間が長引くことがあります。
- 外れ値は必ずしも厄介な存在とは限りません。不正の検出（異常検知）のように、外れ値の検出自体が分析の主な目的となる場合もあります。

外れ値の扱い

- 何もしない

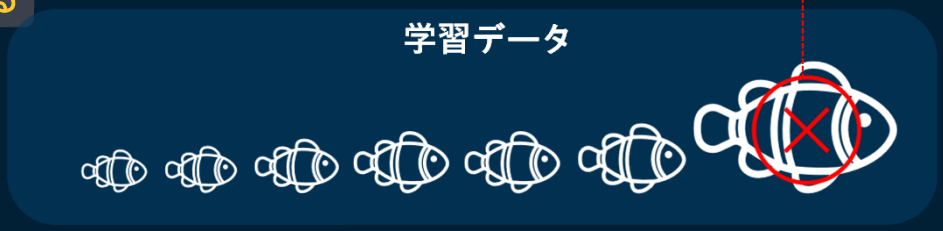
変数：魚の大きさ

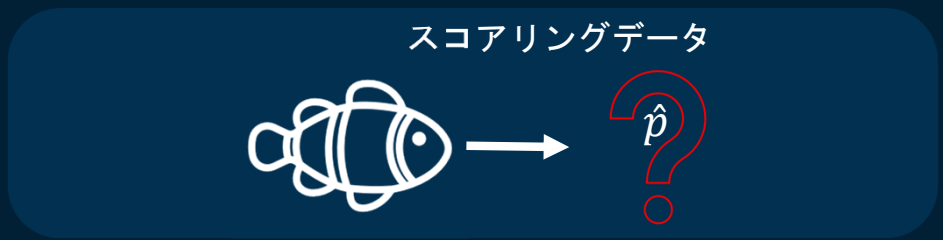
外れ値の処理

- 何もしない
- **外れ値を削除する**

誤ったデータポイントにのみ適している

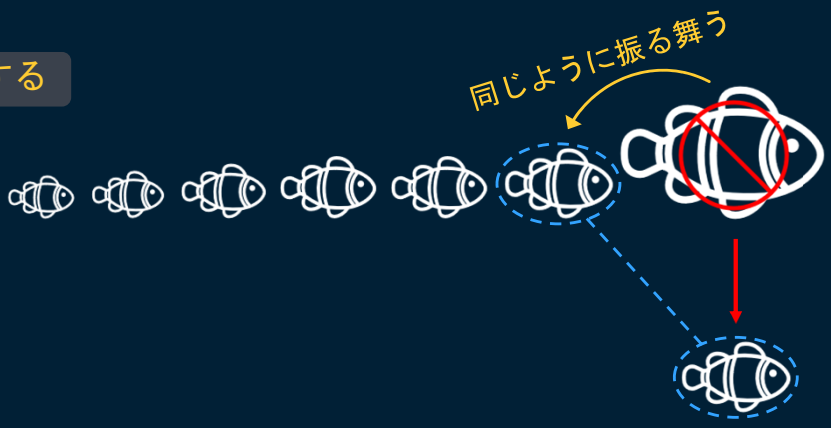
バイアスのあるモデルが生成される





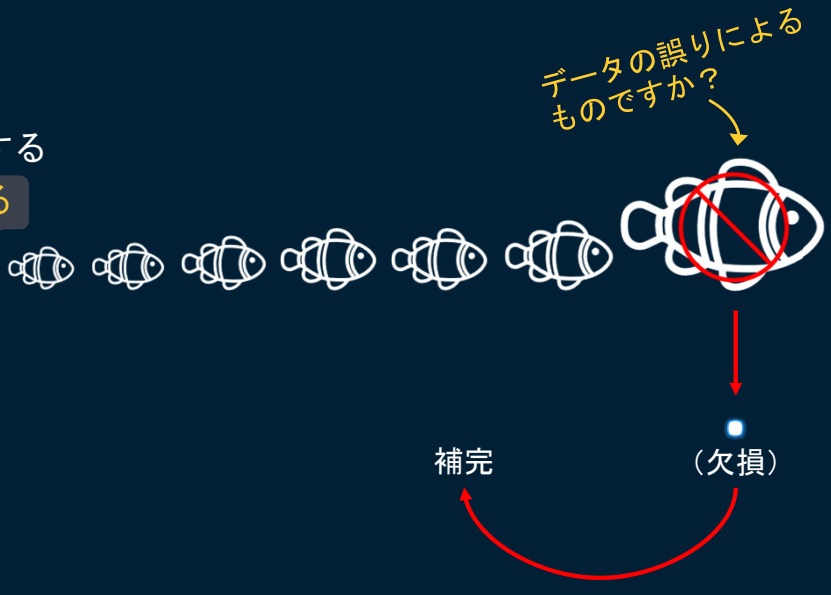
外れ値への対処

- 何もしない
- 外れ値を削除する
- 外れ値の上限を設定する



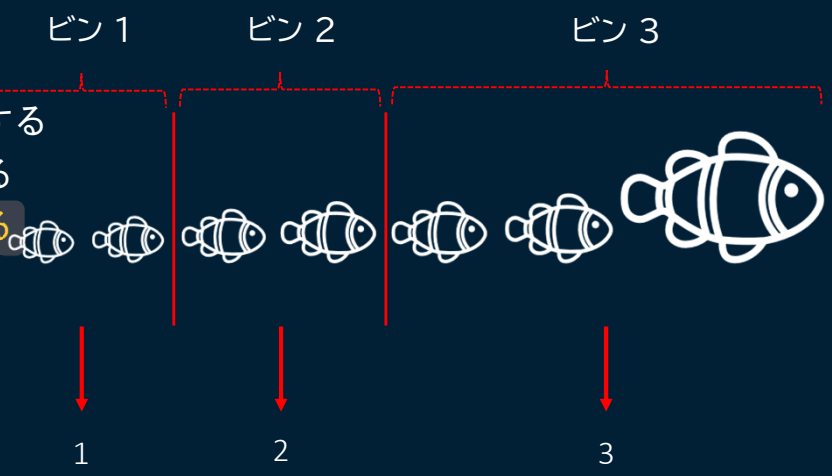
外れ値の処理

- 何もしない
- 外れ値を削除する
- 外れ値の上限を設定する
- 新しい値を割り当てる



外れ値の処理

- 何もしない
- 外れ値を削除する
- 外れ値の上限を設定する
- 新しい値を割り当てる
- 変数をビンニングする



バケット ビニング

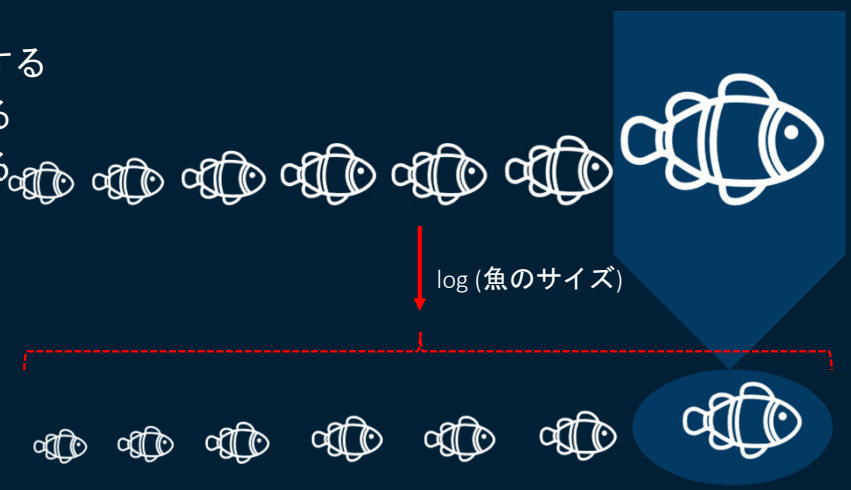
各ビンの等間隔幅

分位数 ビンニング

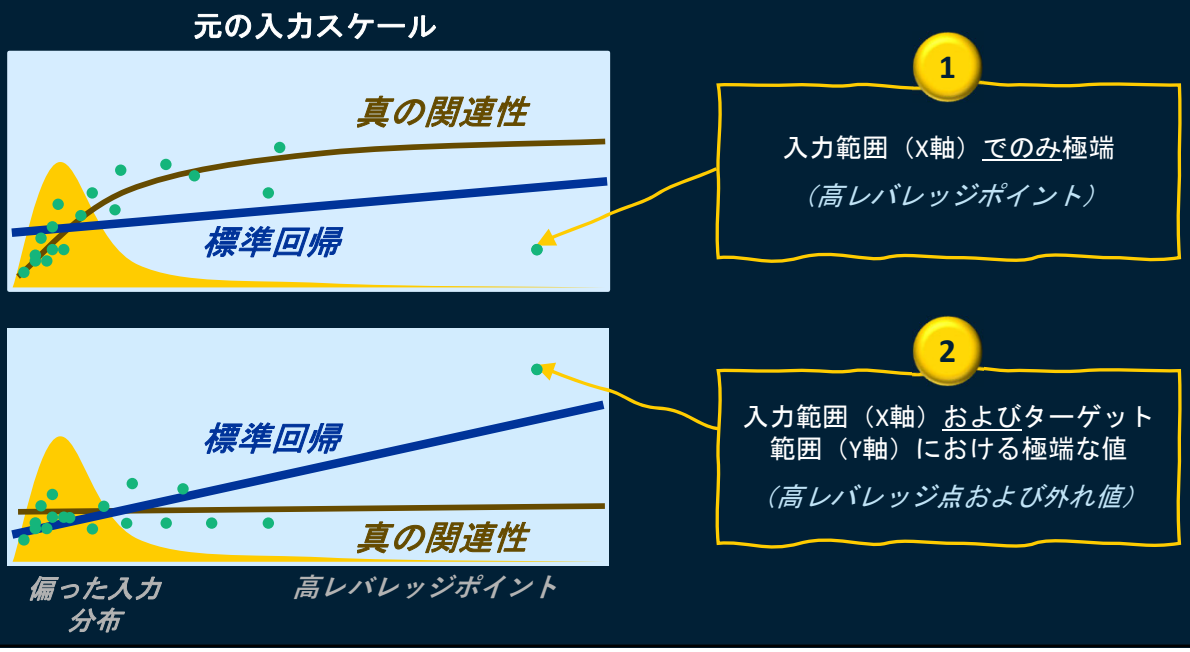
各ビンに等しい観測値数

外れ値の処理

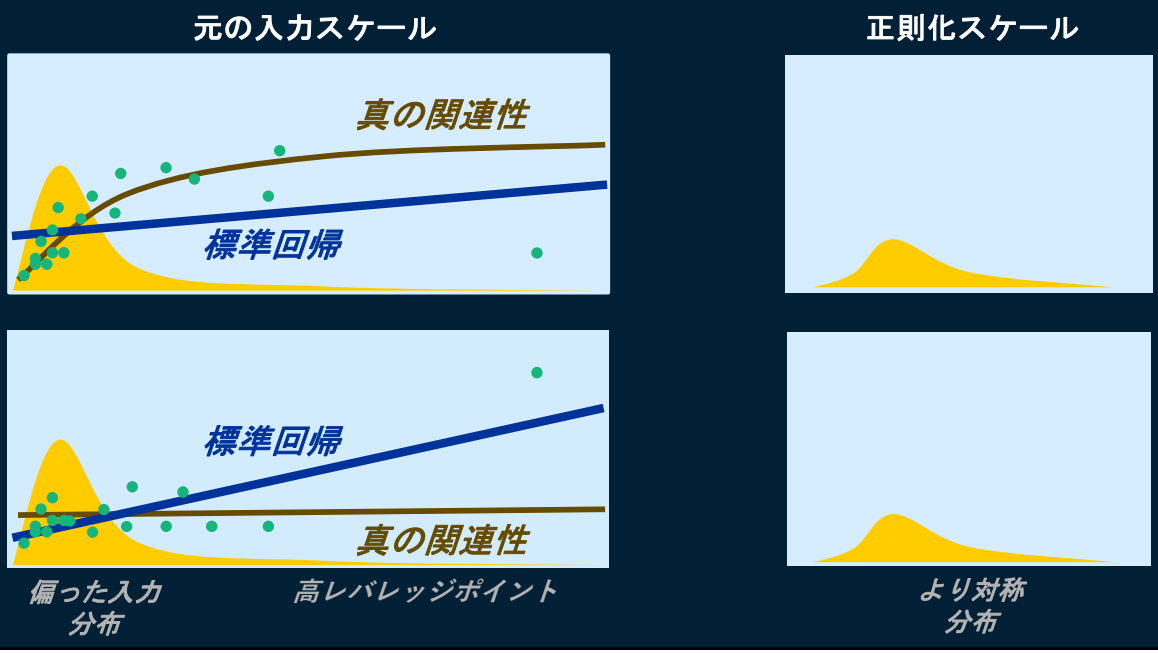
- 何もしない
- 外れ値を削除する
- 外れ値の上限を設定する
- 新しい値を割り当てる
- 変数をビンニングする
- 変数の変換



変換にはどのような効果があるのでしょうか？



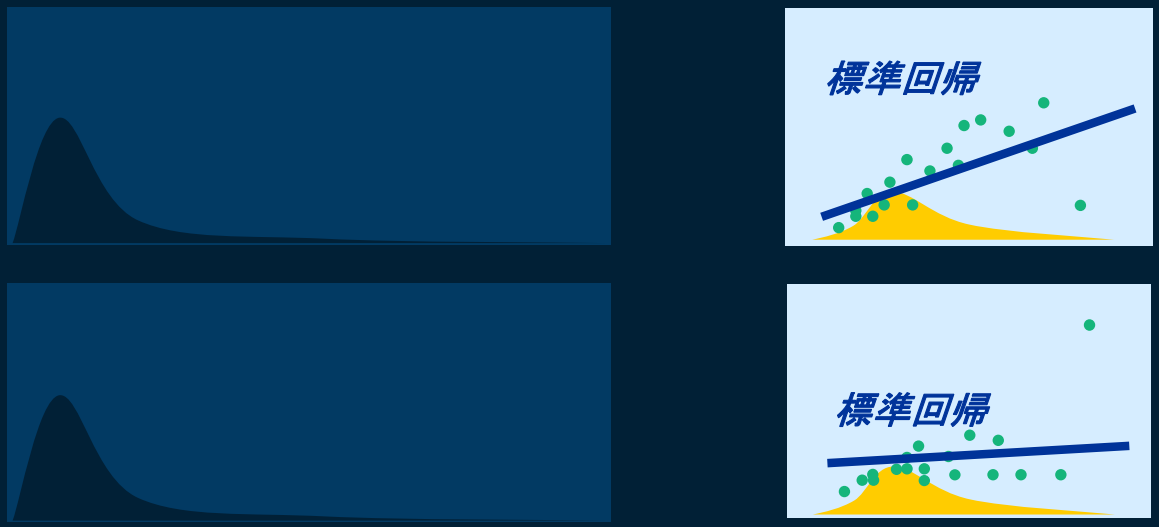
トランスフォーメーションにはどのような役割があるのでしょうか？



変換とは何をするものか？

元の入力スケール

正則化スケール



変換とは何をするものか？

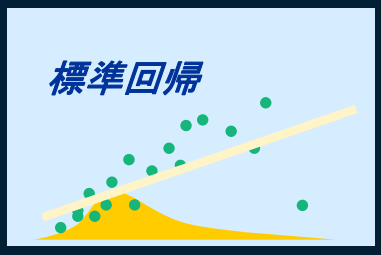
元の入力スケール

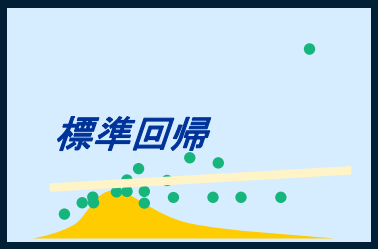
正則化スケール

変換とは何をするものか？

元の入力スケール

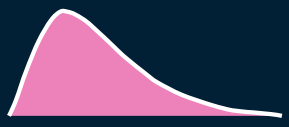
正則化スケール





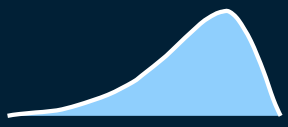
単純な変換

右に歪んだデータ



- 対数 $\log(x)$ 最も一般的
- 根 $\sqrt[n]{x}$ 最も弱い
- 逆数 $1/x$ 最も強い

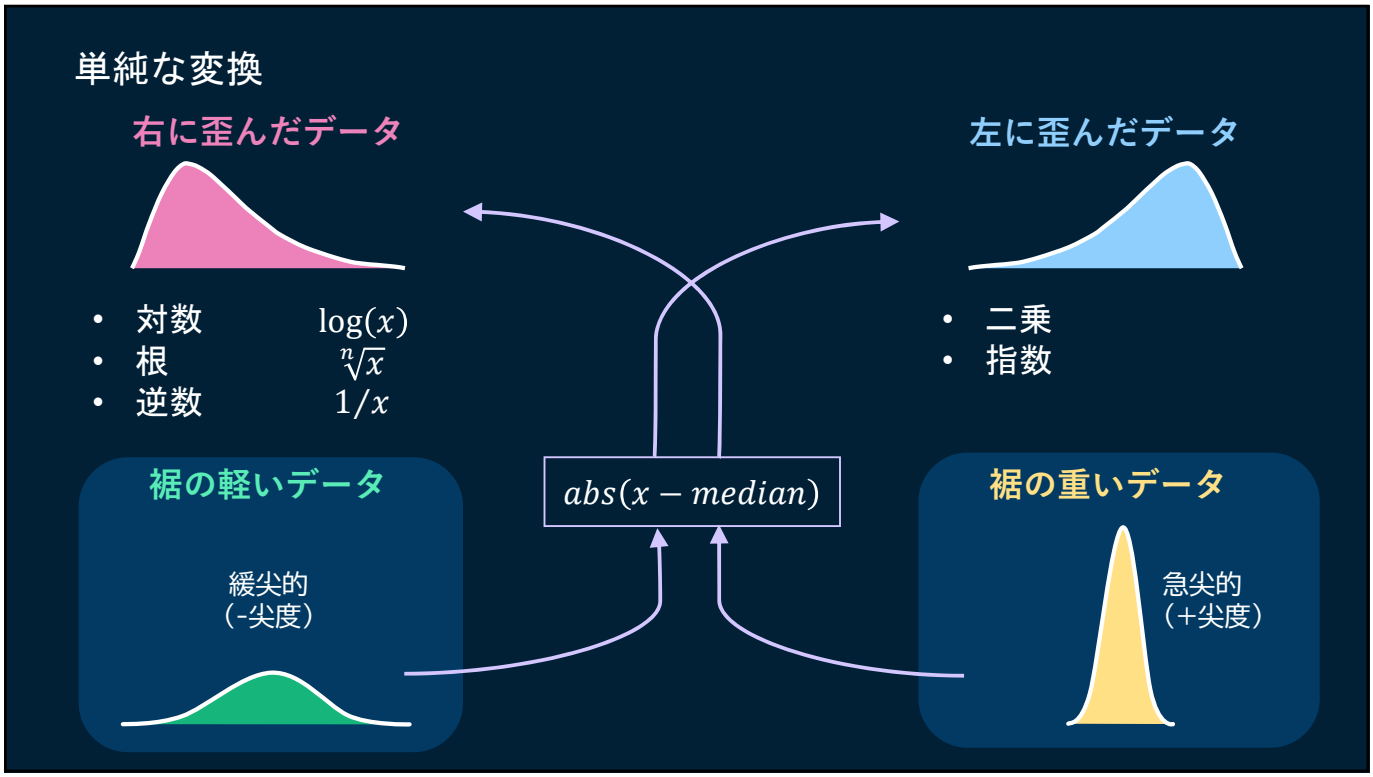
左に歪んだデータ



より高い指数で強くできる

- 二乗
- 指数

底数を大きくすることで強くなる



問題：変数の変換

回帰分析における入力変数の変換について、以下の記述のうち正しいものはどれか。

a. それらは決して良い考えではない。
b. それらは、モデルの仮定と最尤推定の仮定を一致させるのに役立つ。
c. モデル予測のバイアスを低減するために実施される。
d. 通常、名目（カテゴリ）変数に対して行われる。

回答：変数の変換

回帰分析における入力変数の変換について、以下の記述のうち正しいものはどれか。

a. それらは決して良い考えではない。
b. それらは、モデルの仮定と最尤推定の仮定を一致させるのに役立つ。
c. モデル予測のバイアスを低減するために実施される。
d. 通常、名目（カテゴリ）変数に対して行われる。

デモ：入力の変換

このデモでは、「列の変換」ステップを使用して、一連の入力データに標準的な変換を適用する方法を示します。

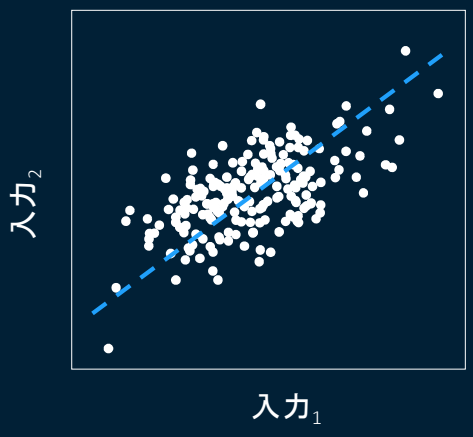
変数が多すぎる問題

データの課題と
モデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- **高次元データの取り扱い**
- 入力データのばらつきへの対処
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

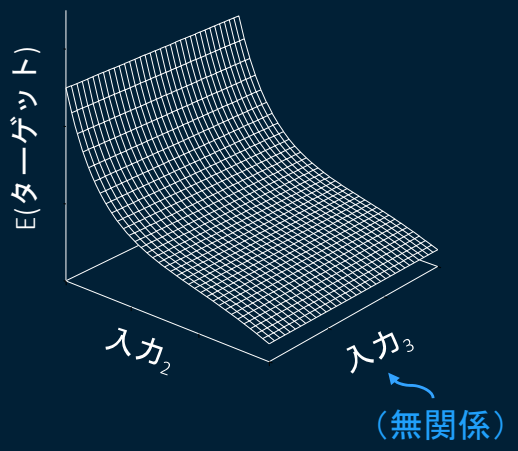
変数が多すぎる問題

- **冗長性**
  （相関のある入力）



変数が多すぎる問題

- 冗長性
- **無関係性**
  （ターゲットとの関連性がない）



変数が多すぎる問題

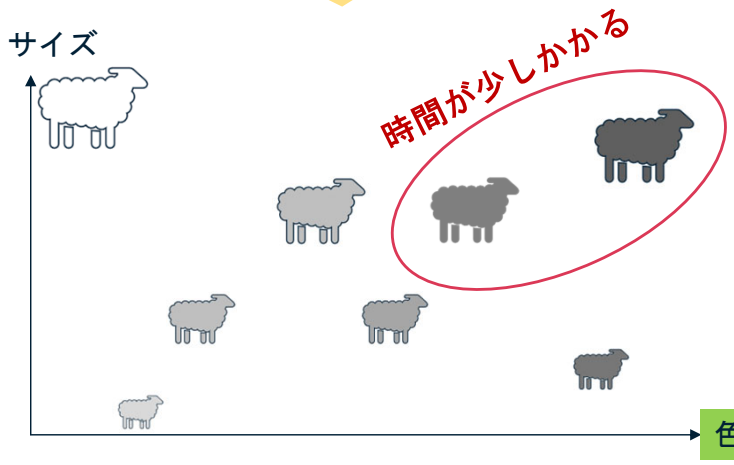
- 冗長性
- 無関係性
- **次元の呪い**
  （自由度）

◯ 特徴量の数がオブザベーションの数を上回る可能性がある
◯ 計算が極めて困難になる

変数が多すぎる問題

- 冗長性
- 無関係性
- 次元の呪い
  （自由度）

より暗くて大きな羊を見つける



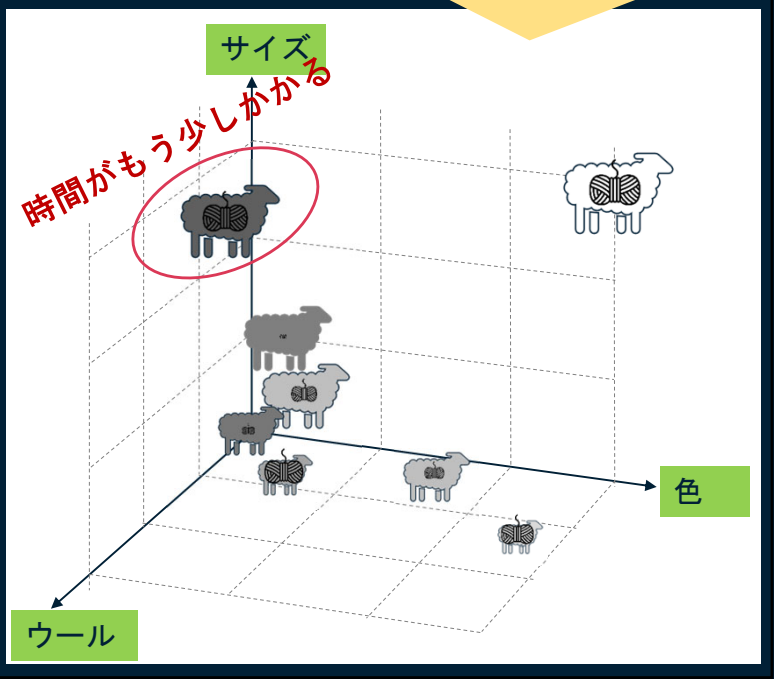
入力変数2つ = 2次元

変数が多すぎる問題

- 冗長性
- 無関係性
- 次元の呪い
  （自由度）

3つの入力変数 = 3次元

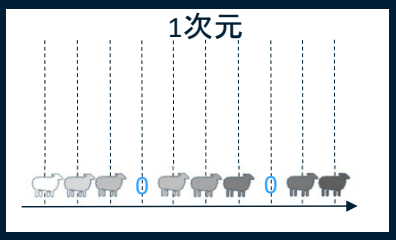
より色が濃く、体が大きく、羊毛密度が高い羊を見つける



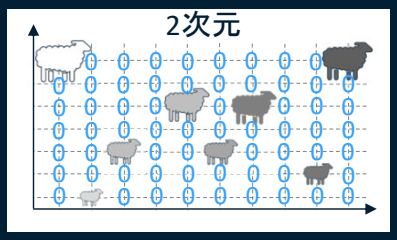
変数が多すぎる問題

- 冗長性
- 無関係性
- 次元の呪い
- スパース性
  （計算行列に多くのゼロが含まれること）

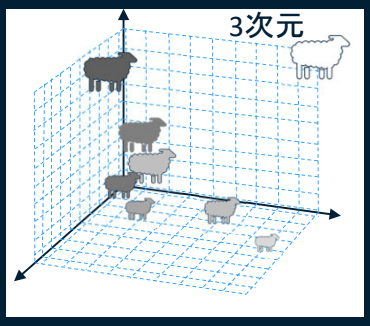
応用機械学習で一般的なもの：
→ カウントデータ
→ カテゴリをカウント値にマッピングするデータエンコーディング
→ 自然言語処理



8/10=80%



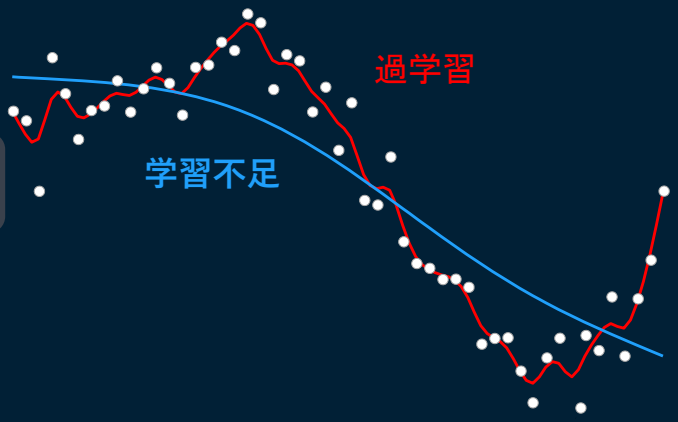
8/(10×10)=8%



8/(10×10×10) = 0.8%

変数が多すぎる問題

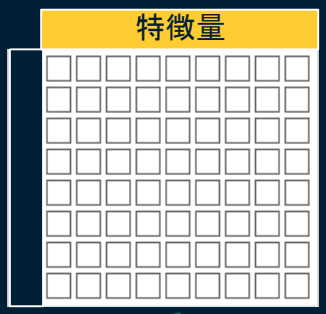
- 冗長性
- 無関係性
- 次元性の呪い
- スパース性
- 過学習
  （モデルの汎化性能が低い）



変数が多すぎる問題

- 冗長性
- 無関係性
- 次元性の呪い
- スパース性
- 過学習

特徴量エンジニアリング
が解決策です。



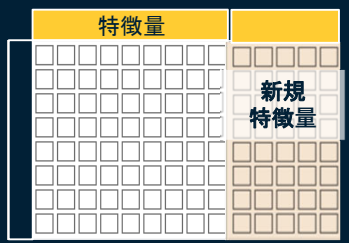
★ モデリングに使用する最適な入力変数を見つけ、作成する。
～ モデルの簡素化、学習時間の短縮、精度の向上 ～

特徴量エンジニアリングの種類

特徴量
構築

特徴量
選択

特徴
抽出



特徴量エンジニアリングの種類

特徴量の構築例

請求日	事故日時
2022年11月11日	2022年11月11日
2021年12月22日	2021年1月23日 01:42
2021年4月26日	2021年4月23日 03:05
2020年7月2日	2020年2月7日 06:25
2022年3月8日	2021年12月30日 18:33
2022年12月15日	2022年12月6日 18:12
2020年11月9日	2020年5月11日 22:14

派生特徴量

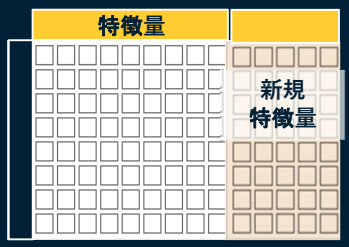
遅延	シーズン	ダーク
19	秋	0
333	冬	1
3	春	1
0	夏	0
69	冬	0
1816	夏	0
4	秋	1

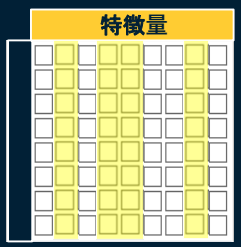
特徴量エンジニアリングの種類

特徴量構築

特徴量選択

特徴抽出





特徴量選択の例

冗長

ダイレクトメールへの反応	世帯収入	住宅価格
1	76,000	370,000
0	21,000	102,000
0	50,000	243,000
0	67,000	325,600
1	100,000	485,800
0	45,000	146,500
1	28,000	202,700

特徴量選択の例

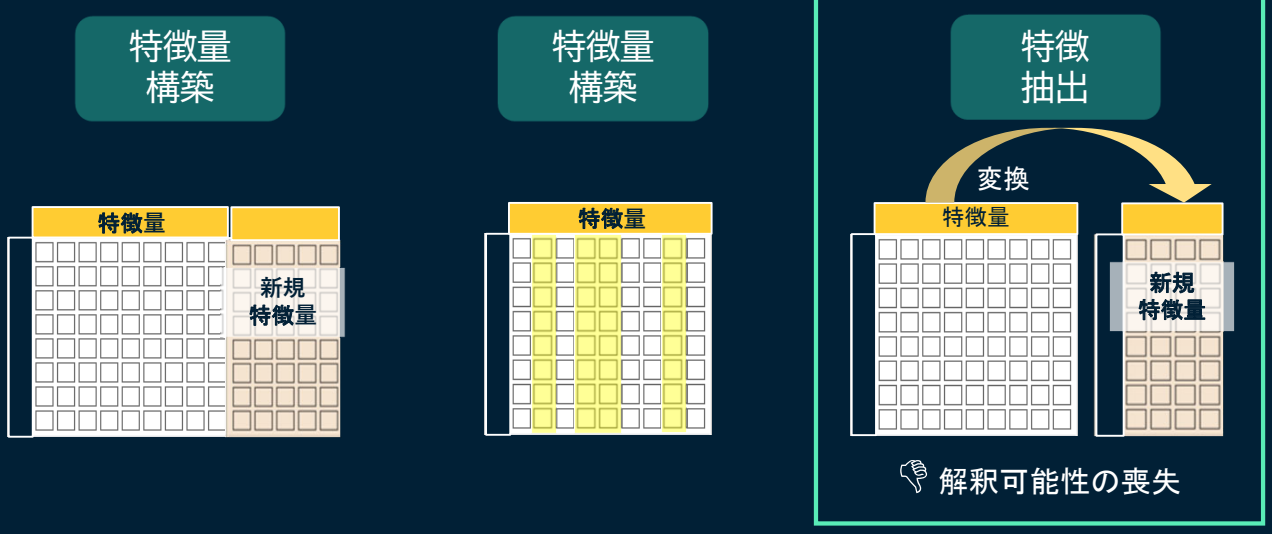
冗長

ダイレクトメールへの反応	世帯収入	住宅価格	靴のサイズ	過去の勧誘
1	76,000	370,000	10	3
0	21,000	102,000	11	1
0	50,000	243,000	9	0
0	67,000	325,600	7	1
1	100,000	485,800	6	4
0	45,000	146,500	13	0
1	28,000	202,700	12	2

無関係

関連あり

特徴量エンジニアリングの種類



特徴抽出の例

主成分分析（PCA）

世帯収入	住宅価格	過去の勧誘
76,000	370,300	3
21,000	102,000	1
50,000	243,000	0
67,000	325,600	1
100,000	485,800	4
45,000	146,500	0
28,000	202,700	2

低次元への数学的投影

抽出された特徴量

主成分
1.406
-1.678
-0.780
0.306
2.790
-1.324
-0.720

1つの主成分が3つの入力変数の変動の87%を説明している

問題：特徴量エンジニアリング

「選択」と「抽出」の主な違いは、特徴量選択では元の特徴量の一部を残すのに対し、特徴量抽出では全く新しい特徴量を作成する点にあります。

a. 正しい
b. 誤り

回答：特徴量エンジニアリング

「選択」と「抽出」の主な違いは、特徴量選択では元の特徴量の一部を残すのに対し、特徴量抽出では全く新しい特徴量を作成する点にあります。

a. 正しい
b. 誤り

デモ：特徴量選択の実行

このデモでは、「変数選択」ステップを使用した特徴量選択の手順を説明します。

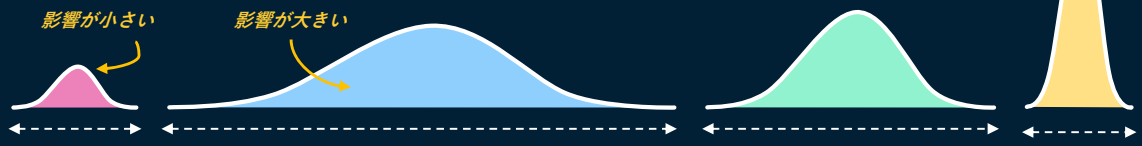
明確にスケーリングされた変数

データの課題とモデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- **入力データにおける明確な変動への対応**
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

スケールが著しく異なる変数

- 実世界のデータには、入力値に不均衡な変動性があります。
  （大きさ、単位、範囲が大きく異なる）



- **データセットの特徴量の間でスケールが大きく異なる場合、モデルは入力のスケールに敏感な影響を受けます**：
  - モデルの学習速度の低下
  - パラメータ推定の精度低下
  - 予測変数への重みの恣意性
  - モデルの不安定性

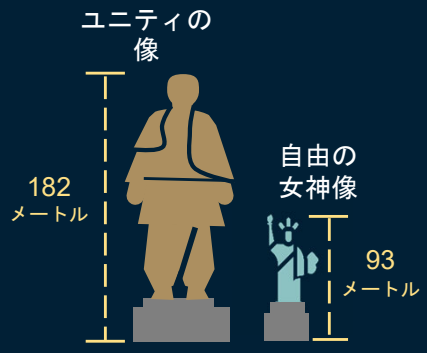
特徴量スケーリング

これらの像の建設の節目について、どのように分析できますか？

使用された鉄とコンクリートの量

建設期間

問題：標準化されていない変数を使用すると、分析において範囲の広い変数がより重要視される可能性があります。



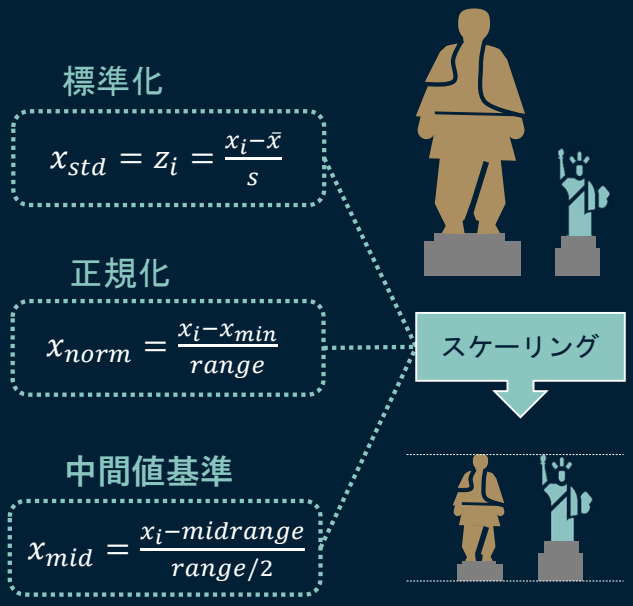
スケーリング

解決策：データを比較可能な尺度で再スケーリングする。

特徴量スケーリング

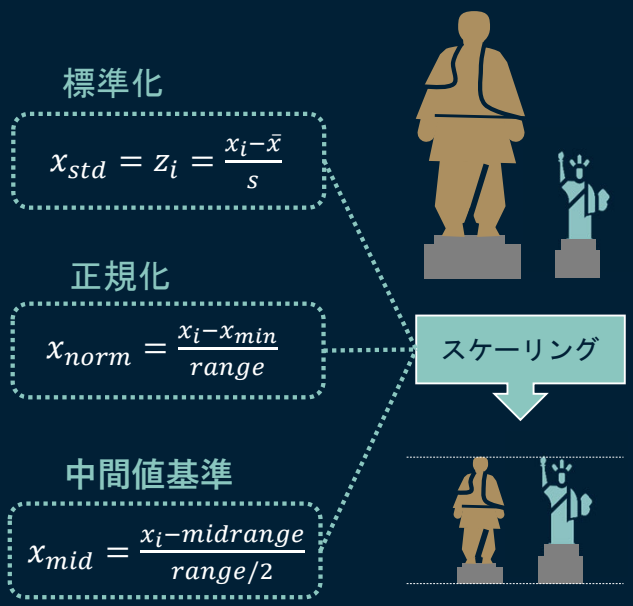
一般的な再スケーリング手法：

- 標準化
  (主に-3 から 3)
- 正規化
  (0～1)
- **中間値基準**
  (-1～1)



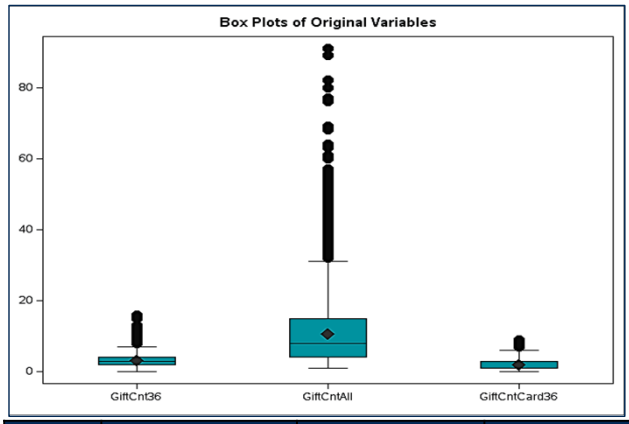
特徴量スケーリング

- 勾配降下法に基づくアルゴリズム
  ニューラルネットワーク、回帰*
- 距離ベースのアルゴリズム
  KNN、K-means、SVM
- **ツリーベースのアルゴリズム**
  決定木、フォレスト、グラディエントブースティング



考えてみよう：特徴量スケーリングの手法

以下の3つの変数を用いて、PVA_DONORSデータセット上でニューラルネットワークモデルを作成する任務を任されたと仮定します。

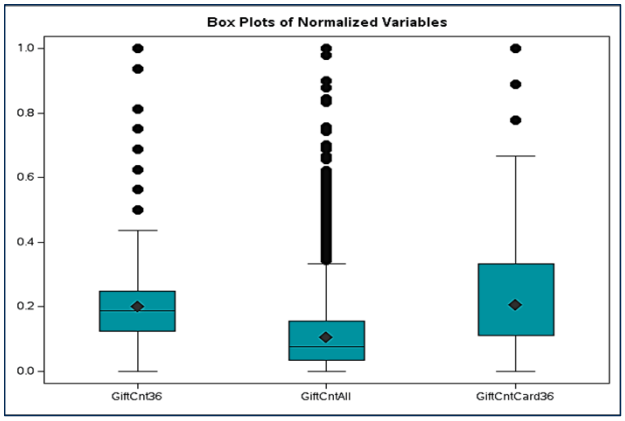


Std	16.515	10.057	8.993
Mean	59.151	14.876	10.508
Max	87	260	91
Min	0	0	1

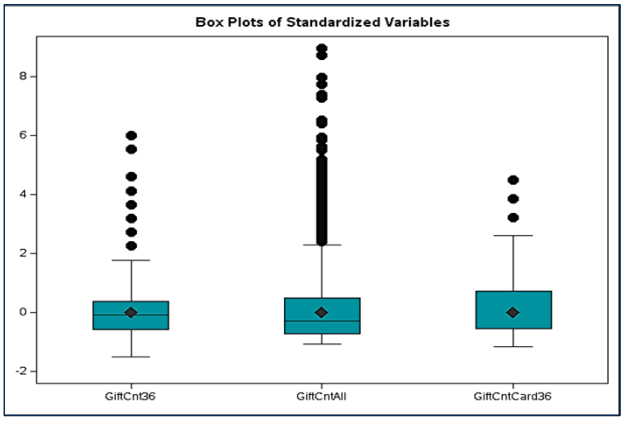
考えてみよう：特徴量スケーリング手法

不均衡な尺度を補正するために、どの特徴量スケーリング手法を使用しますか（もしあるなら）？

正規化

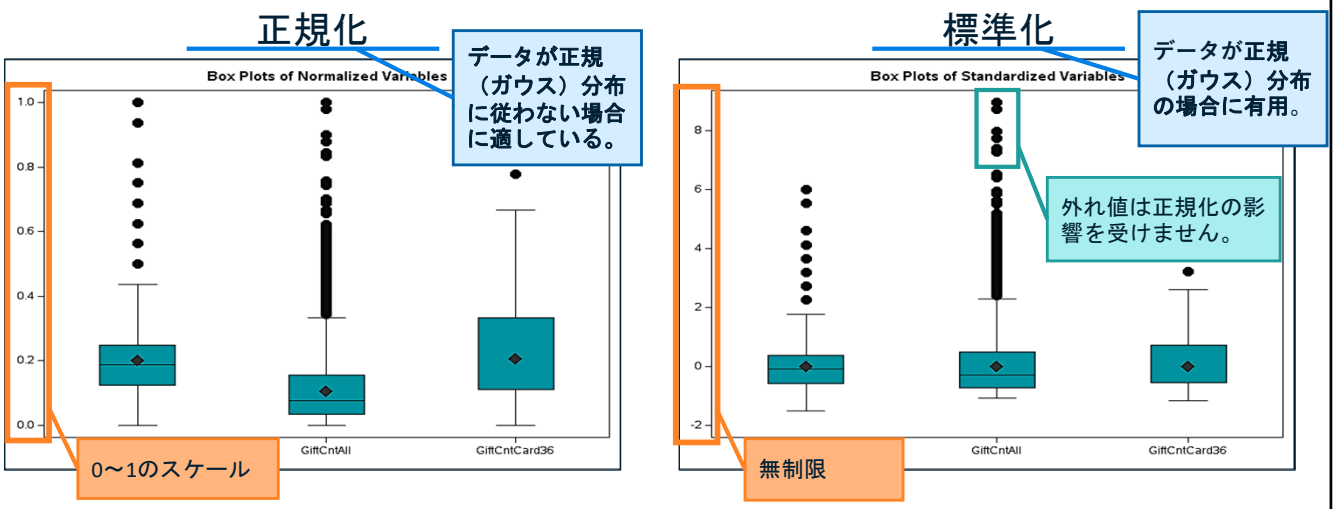


標準化



回答：特徴量スケーリング手法

不均衡なスケールを補正するために、どのような特徴量スケーリング手法を使用しますか（もしあるなら）？ ***データ次第です！***



問題：特徴量のスケーリング

なぜ、一部の機械学習アルゴリズムでは、データ内の特徴量セットをスケーリングする必要があるのでしょうか？

a. 特徴量をより解釈しやすくするため
b. データ値を正の値にするため
c. 単にそれが楽しいから
d. 多くのアルゴリズムが、スケーリングが不均一なデータに敏感だから

回答：特徴量のスケーリング

なぜ、一部の機械学習アルゴリズムでは、データ内の特徴量セットをスケーリングする必要があるのでしょうか？

a. 特徴量をより解釈しやすくするため
b. データ値を正の値にするため
c. 単に楽しいから
d. 多くのアルゴリズムが、スケーリングが不均一なデータに敏感だから

モデルの評価、推定、および学習後のタスク

機械学習におけるモデル評価、推定、および学習後のタスクについて

Lesson 05, Section 03

機械学習データにおけるモデリングの課題

データの課題と
モデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データのばらつきへの対応
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

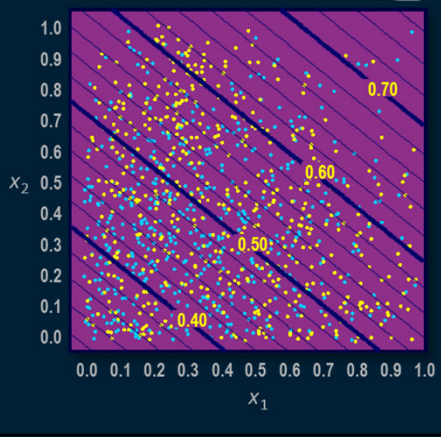
機械学習データにおけるモデリングの課題

- データセットが小さい

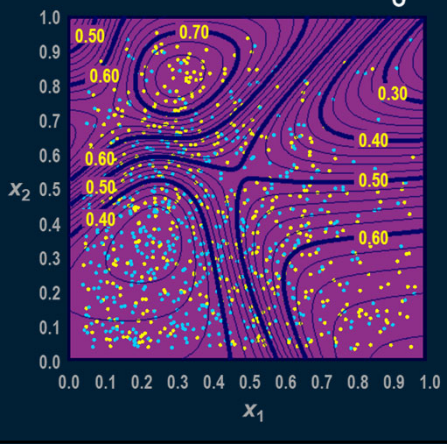
機械学習データにおけるモデリングの課題

- **データセットが小さい**

統計モデルまたはより単純な機械学習モデル

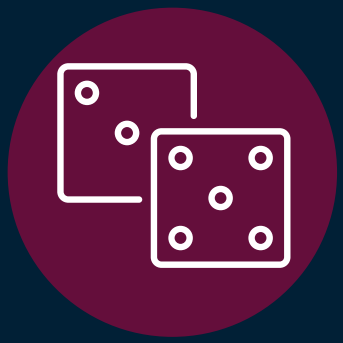


データ消費量の多い複雑な機械学習モデル



機械学習データにおけるモデリングの課題

- データセットが小さい
- 関連データの不足



データにはどの程度のシグナルが含まれていますか？

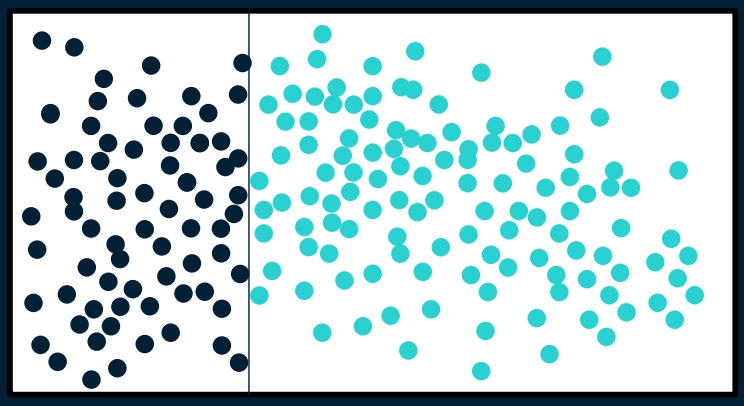
データから読み取れる真の根本的なパターン

ターゲット = シグナル + ノイズ

系統的変動	ランダムな変動
↓	↓
説明可能	説明不能
↓	↓
予測可能	予測不能

機械学習データにおけるモデリングの課題

純粋な分離 =
純粋な信号 =
ノイズなし

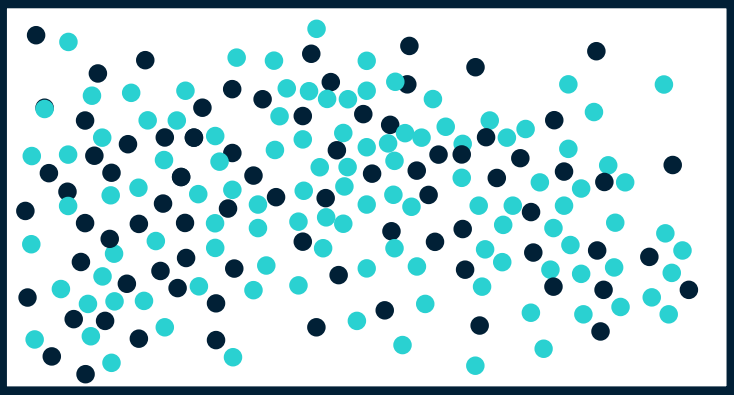


入力

ターゲット：主要アウトカム＝●　副次アウトカム＝●

機械学習データにおけるモデリングの課題

分離なし =
純粋なノイズ =
信号なし



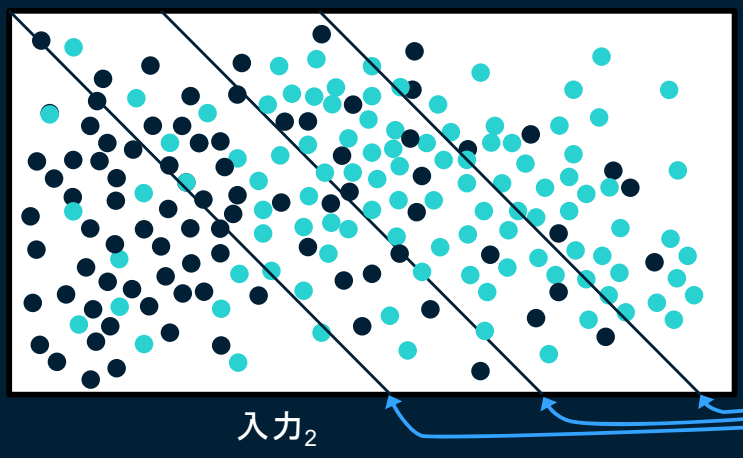
入力$_1$

入力$_2$

ターゲット：主要アウトカム＝●　副次アウトカム＝●

機械学習データにおけるモデリングの課題

ある程度の分離 =
信号 + ノイズ

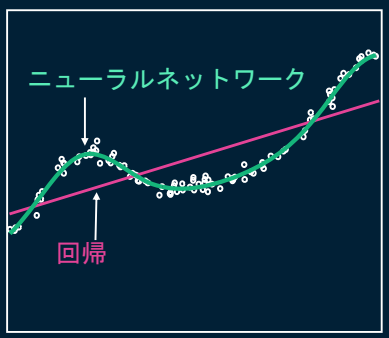


ターゲット：主要アウトカム＝●　副次アウトカム＝●

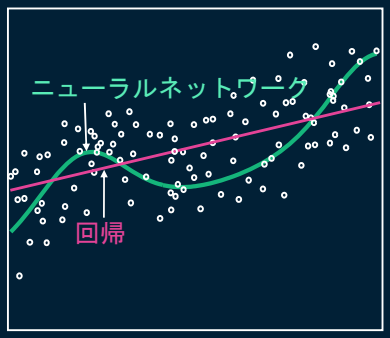
機械学習データにおけるモデリングの課題

機械学習モデルが常に優れているとは限らない！

ノイズの多いデータはモデルに影響を与える。



$\frac{\text{signal}}{\text{noise}} = \text{high}$



$\frac{\text{signal}}{\text{noise}} = \text{low}$

問題：シグナルとノイズ

データから学習したい真の根本的なパターンは「ノイズ」と呼ばれ、データセットに含まれる無関係な情報やランダム性は「シグナル」と呼ばれます。

a. 正しい
b. 誤り

回答：シグナルとノイズ

データから学習したい真の根本的なパターンは「ノイズ」と呼ばれ、データセットに含まれる無関係な情報やランダム性は「シグナル」と呼ばれます。

a. 正しい
b. 誤り

機械学習データにおけるモデリングの課題についてさらに詳しく

- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い

不正
病気の進行
勧誘への反応
ローン延滞
航空会社のノーショー
顧客離反
ネットワークへの侵入
通信回線の過負荷

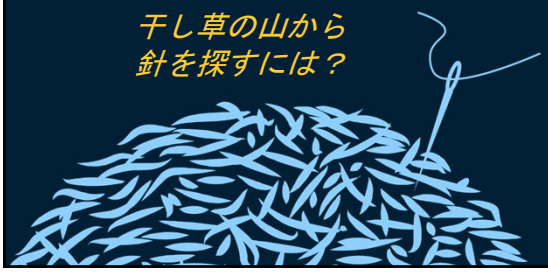
機械学習データにおけるモデリングの課題についてさらに詳しく

- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い

課題

- ターゲットの検索はしばしば困難である
- 多数派グループから多くを学習し、少数派の事象グループを誤分類しやすい
- 信頼性の高いモデルを構築するには、はるかに大規模なサンプルが必要となる（Harrell 1997）

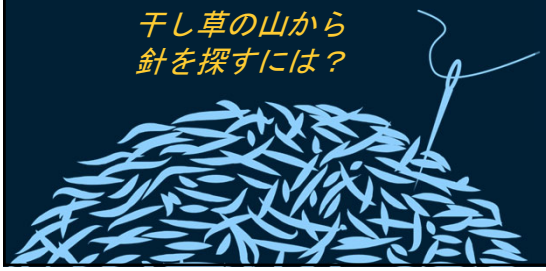
干し草の山から針を探すには？



機械学習データを用いたモデリングの課題についてさらに詳しく

- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い

干し草の山から針を探すには？



課題

- ターゲットの検索はしばしば困難である
- 多数派グループから多くを学習し、少数派の事象グループを誤分類しやすい
- 信頼性の高いモデルを構築するには、はるかに大規模なサンプルが必要となる（Harrell 1997）

解決策

- イベントベースのサンプリング
  - アンダサンプリング
  - オーバーサンプリング
- 人工ケースの利用
  - SMOTE
- 教師なし学習
- ルール誘導

機械学習データにおけるモデリングの課題についてさらに詳しく

- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い
- 利用可能なデータの効率的な活用

リサンプリング法

- 交差検証法
- ブートストラップ法

機械学習データにおけるモデリングの課題についてさらに詳しく

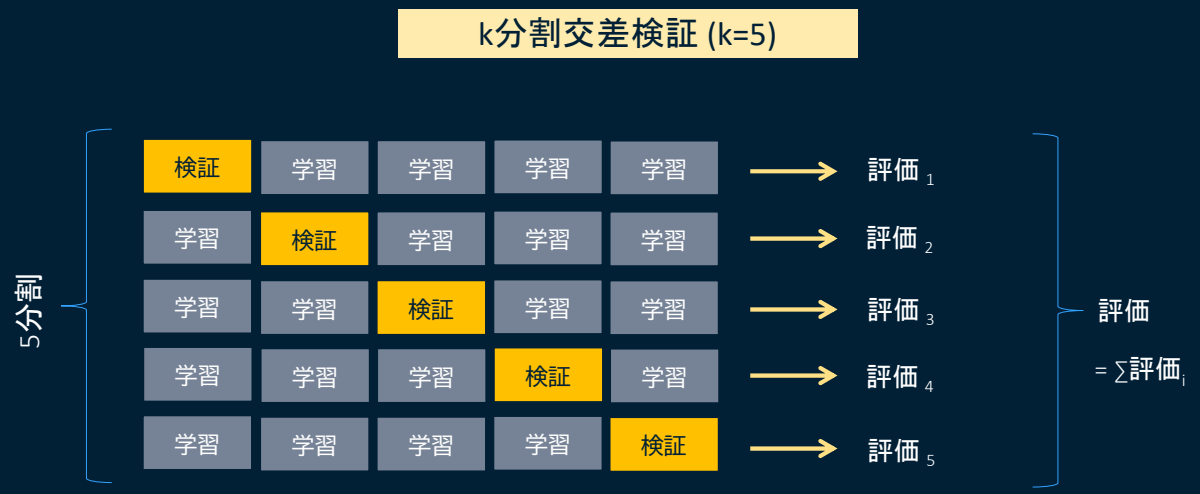
- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い
- 利用可能なデータの効率的な活用

交差検証

- データサンプルが小さい場合に用いられる
- データを複数の部分に分割する
- 学習データの利用可能性を最大化する
- すべてのデータが学習と検証の両方に使用される
- 評価結果は平均化される

小規模および中規模のデータセットについては、k分割交差検証（Breiman et al. 1984; Ripley 1996; Hand 1997）がより優れた戦略である。

交差検証の例



ブートストラップ集計

- データセットが小さい
- 関連データの不足
- ターゲットイベントの発生頻度が低い
- 利用可能なデータの効率的な活用

ブートストラップ法

- 同じサイズの小さなサンプルを多数、繰り返し抽出する
- 機械学習モデルに伴う不確実性を定量化する

ブートストラップ集計

学習データ

ケース	入力		ターゲット
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

k = 1

ケース	度数
1	1
2	0
3	2
4	0
5	2
6	1

k = 2

ケース	度数
1	0
2	1
3	0
4	2
5	2
6	1

k = 3

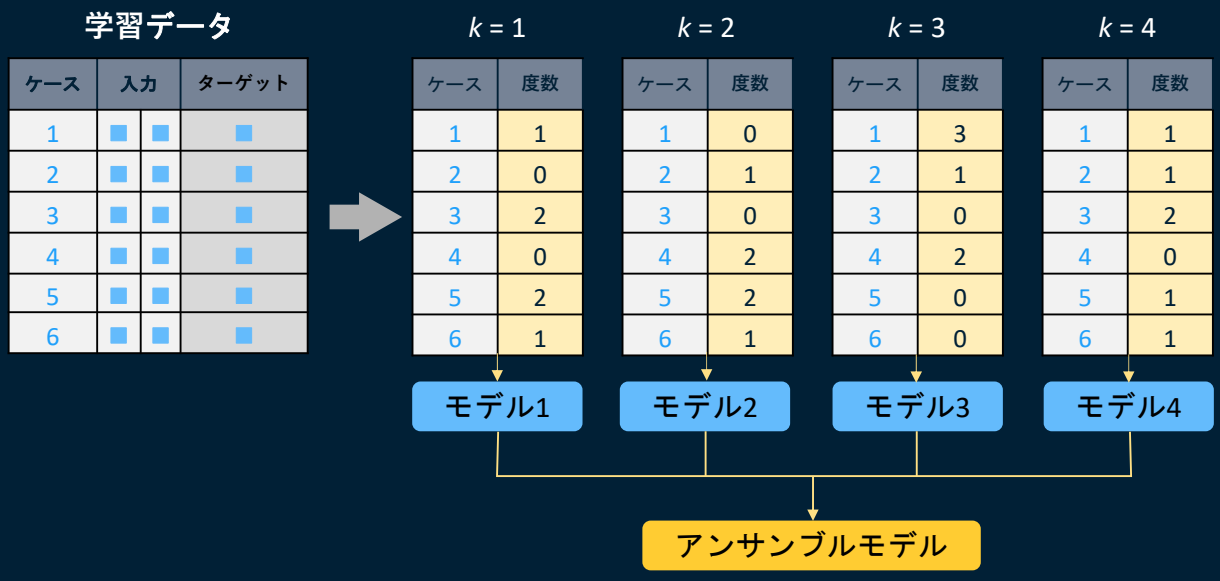
ケース	度数
1	3
2	1
3	0
4	2
5	0
6	0

k = 4

ケース	度数
1	1
2	1
3	2
4	0
5	1
6	1

反復ありのランダム抽出

ブートストラップ集約



問題：交差検証

交差検証は、データサイズが小さく、検証やテスト、あるいはその両方のためにデータを別途確保する余裕がない場合に利用できる、もう一つの客観的な評価手法です。

a. 真
b. 誤り

回答：交差検証

交差検証は、データサイズが小さく、検証やテスト、あるいはその両方のためにデータを別途確保する余裕がない場合に利用できる、もう一つの公正な評価手法です。

a. 真
b. 誤り

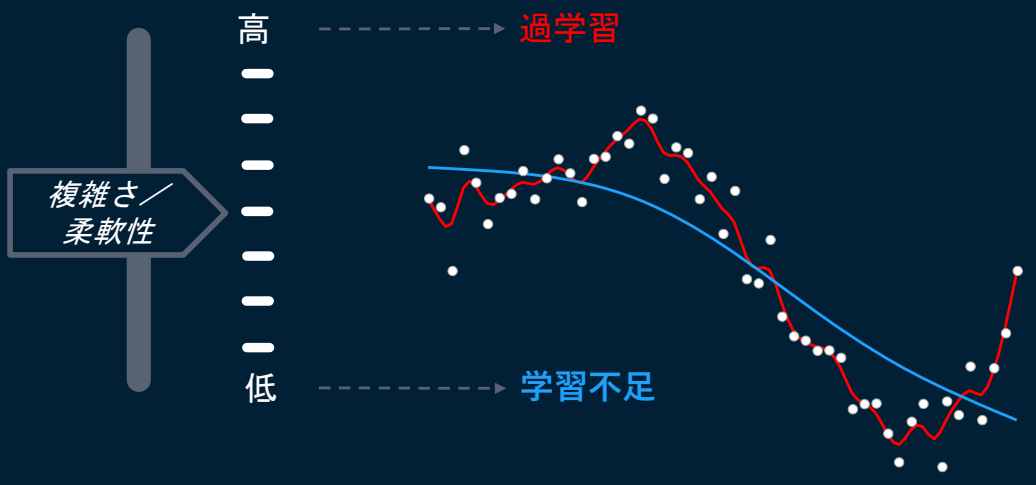
モデルの適合

データの課題とモデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データのばらつきへの対応
- 利用可能なデータの適切な活用
- **モデルの柔軟性と複雑さの調整**
- モデルの学習とチューニングの理解
- モデルの説明可能性の確保

モデル適合

モデル適合



モデル適合

異なるモデルにおける複雑さと柔軟性

モデル	各モデルの複雑さに影響を与える要因
回帰分析	**パラメータ／効果の数**
決定木	**葉の数**
勾配ブースティング	**ツリーの数**
SVM	**サポートベクターの数**
ニューラルネットワーク	**反復の回数**

問題：モデルの複雑さ

次の記述のうち、正しいものはどれですか。該当するものすべてを選択してください。

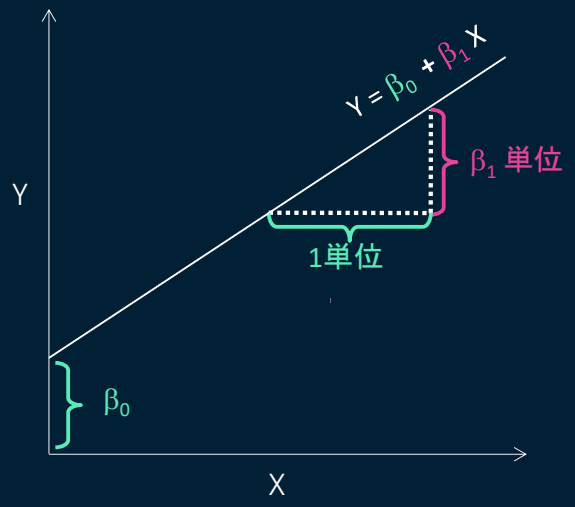
a. より複雑なモデルは、一般的に過学習を招く。
b. より複雑なモデルは、一般的に学習不足を引き起こす。
c. モデルの複雑さは、過学習や学習不足とは無関係である。
d. モデルの複雑さが低いほど、その柔軟性は低くなる。

回答：モデルの複雑さ

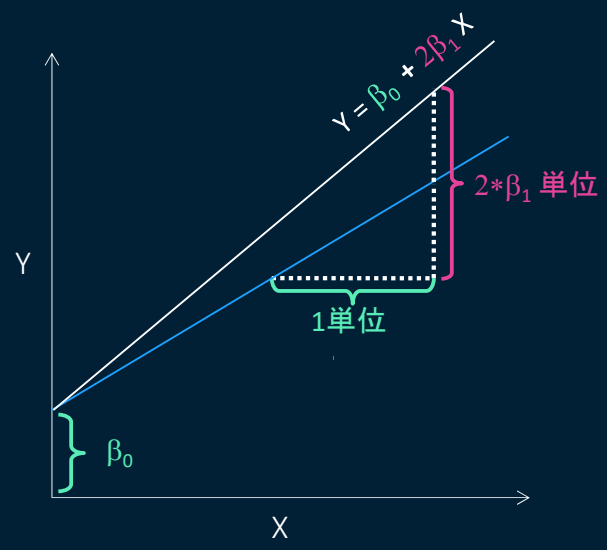
次の記述のうち、正しいものはどれですか。該当するものすべてを選択してください。

a. より複雑なモデルは、一般的に過学習を招く。
b. より複雑なモデルは、一般的に学習不足を引き起こす。
c. モデルの複雑さは、過学習や学習不足とは無関係である。
d. モデルの複雑さが低いほど、その柔軟性は低くなる。

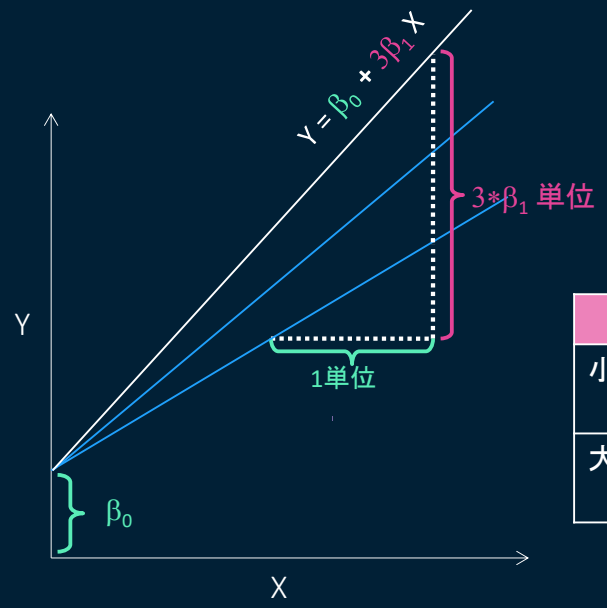
係数の大きさの影響



係数の大きさの影響



係数の大きさの影響



係数	結果
小さい	Yの予測値は、Xの変化に対して感度が低い
大きい	Yの予測値は、Xの変化に対してより敏感である

係数を縮小する

罰則
縮小
鈍感化
正則化

係数の大きい入力パラメータに適用される

複雑さに対する罰則

適切な量のバイアスを加える

モデルをテストデータに適切に適合させることの総称

係数を縮小する

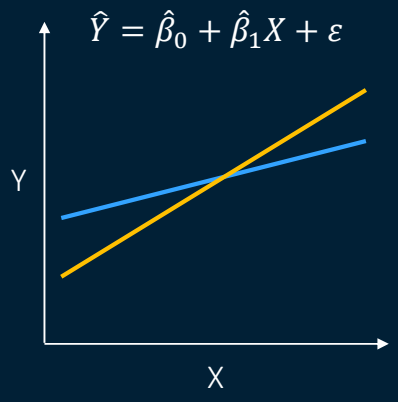
正則化

係数を縮小する

正則化ペナルティ

OLS回帰は

$$RSS = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$



収縮／正則化は以下を最小化する

$$RSS + Regularization\ Penalty$$

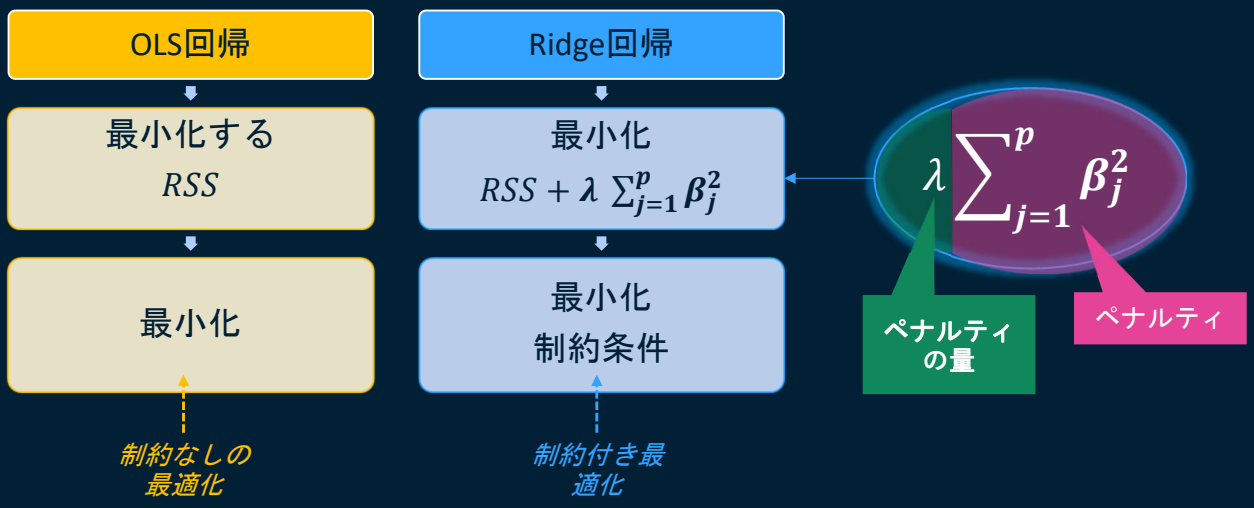
- L1 (Lasso) ..... 特徴量のサブサンプルを選択する
- L2 (Ridge) ..... より単純なモデルを使用する
- L12 (Elastic Net) ..... 両方

係数を縮小する

L1ペナルティの種類 – ラッソ正則化

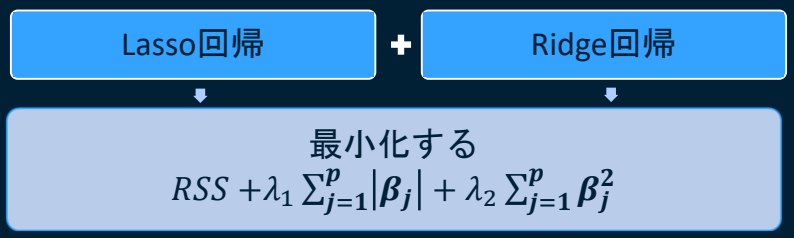
係数を縮小する

L2ペナルティの種類 – リッジ正則化



係数を縮小する

L12ペナルティの種類 – エラスティックネット正則化

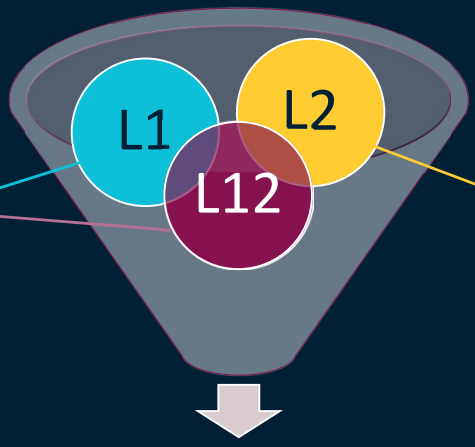


特徴量を過度に切り捨てずに選択する

係数を縮小する

どちらを使うべきか？

- ノイズの多いデータ
- 説明変数が多すぎる
- 特徴量選択が焦点



- ノイズの少ないデータ
- 説明変数の数が減少
- 予測が焦点

L1 対 L2

問題：正則化法

正則化手法に関して、次の記述のうち正しいものはどれか。

a. Lasso（L1）では、重みの絶対値の和がペナルティとして課される。
b. Ridge（L2）では、重みの二乗和がペナルティとして課される。
c. Elastic Net（L12）は、Lasso（L1）とRidge（L2）のハイブリッド。
d. すべて正しい。

回答：正則化法

正則化手法に関して、次の記述のうち正しいものはどれか。

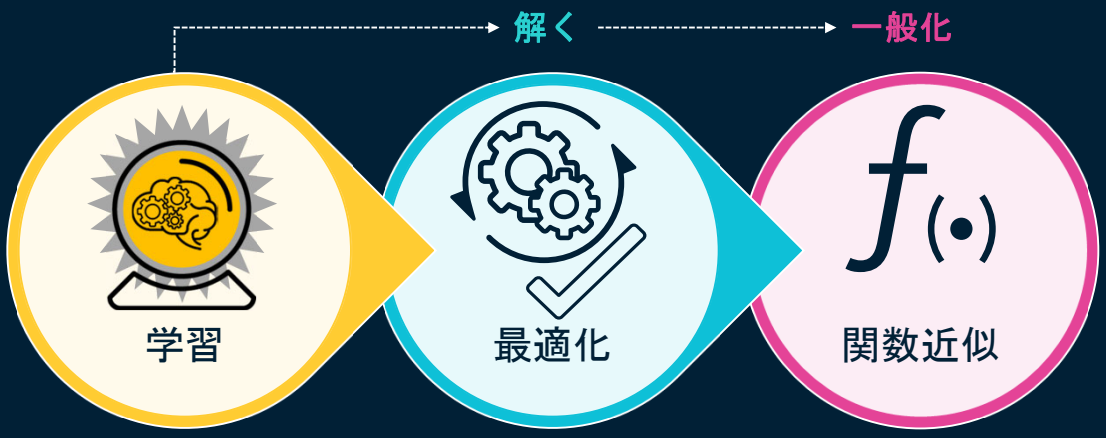
a. Lasso（L1）では、重みの絶対値の和がペナルティとして課される。
b. Ridge（L2）では、重みの二乗和がペナルティとして課される。
c. Elastic Net（L12）は、Lasso（L1）とRidge（L2）のハイブリッド。
d. すべて正しい。

学習プロセス

データの課題と
モデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データの顕著な変動への対応
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- **モデルの学習とチューニングの理解**
- モデルの説明可能性の確保

学習プロセス



学習プロセス

目的関数

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2$$

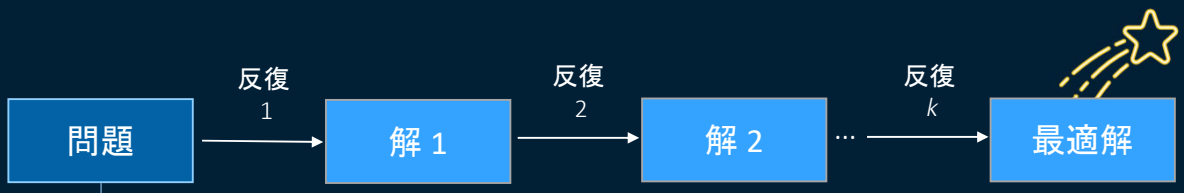
推定基準を最適化するために
パラメータの推定値を選択する。

- 誤差関数
- 偏差関数
- 損失関数
- コスト関数

二乗誤差関数の最小化: $\sum (Y_i - \hat{Y}_i)^2$

学習プロセス

反復プロセス



- 目的関数
- 決定変数
- 制約条件

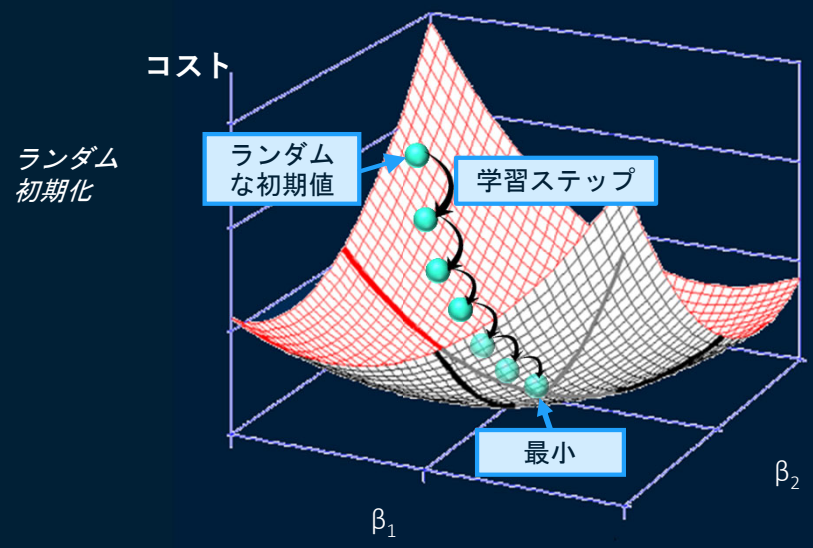
学習プロセス

勾配降下法

勾配 :
勾配の急峻さ

勾配降下法 :
コスト関数を最小化するためにパラメータを反復的に調整すること

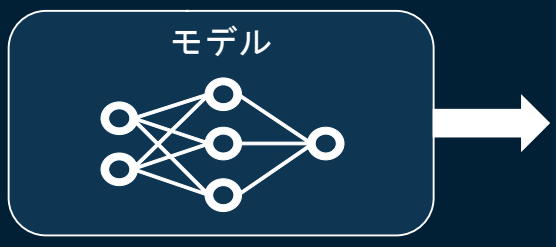
- 誤差関数の局所勾配を測定する
- 勾配が下がる方向に進む



学習プロセス

パラメータの推定

最適化は、モデルのパラメータを推定するために使用されます。

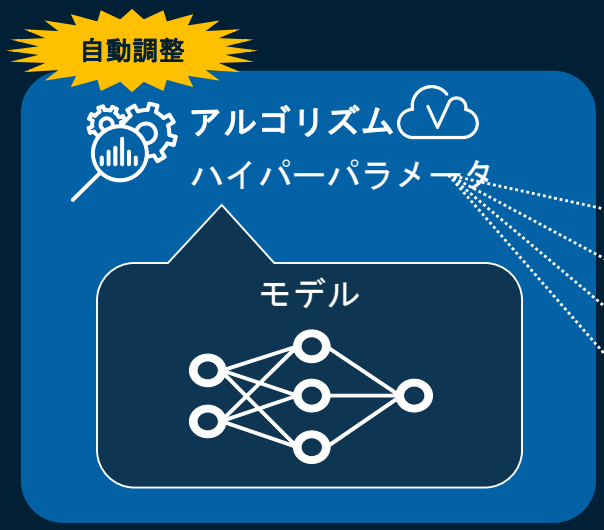


パラメータとは、学習プロセスにおいてデータを通じて学習される値のことです。

$$\text{logit}(\hat{p}) = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$
$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$
$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$
$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

学習プロセス

ハイパーパラメータのチューニング



ハイパーパラメータ ：学習プロセス前に指定される値

- 隠れ層の数
- ニューロンの数
- 学習率
- L1およびL2正則化

パラメータとハイパーパラメータ

推定基準の詳細

問題：学習用語

以下の説明に対して、次の用語のうちどれを使用しますか？

データセットで観測された値と、
モデルによって予測された値との差。

a. ランダム初期化
b. 反復
c. 誤差関数
d. 誤り

回答：学習用語

次の説明に対して、以下の用語のうちどれを使いますか？

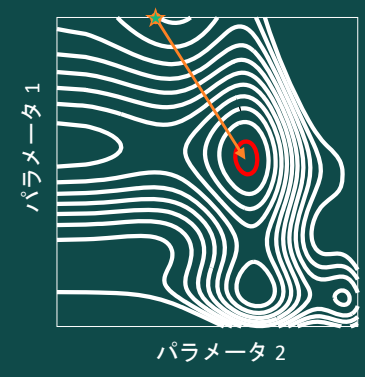
> *データセットで観測された値と、*
> *モデルによって予測された値との差。*

a. ランダム初期化
b. 反復
c. 誤差関数
d. 誤り

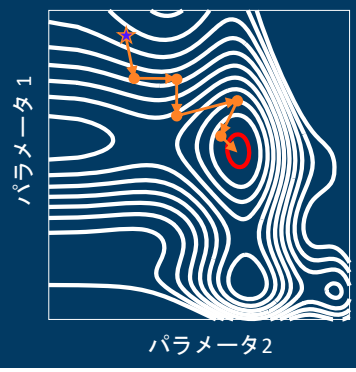
問題：パラメータ値の決定

ニューラルネットワークのパラメータ値を決定する際、以下のうち正しいものはどれか。

a. 最適推定値の明示的な式を求めることができる。



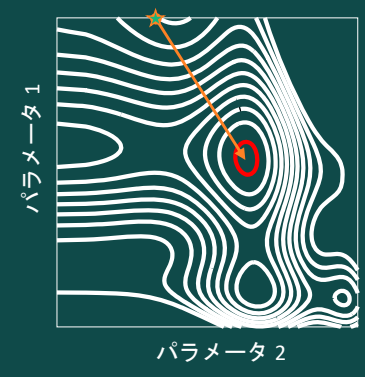
b. 反復的な数値最適化法が用いられる。



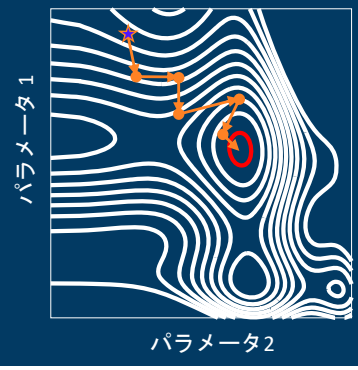
回答： パラメータ値の決定

ニューラルネットワークのパラメータ値を決定する際、以下のうち正しいものはどれか。

a. 最適推定値の明示的な式を求めることができる。



b. 反復的な数値最適化法が用いられる。



デモ：ニューラルネットワークモデルの実行とハイパーパラメータの調整

このデモでは、ニューラルネットワークモデルの学習を行い、そのハイパーパラメータを調整する方法を示します。

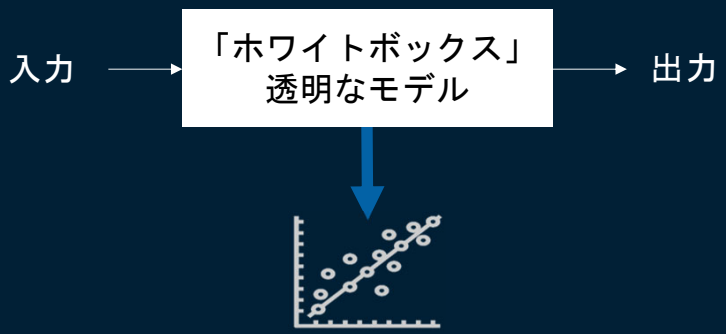
モデルの解釈可能性

データの課題とモデリング上の問題

- ビッグデータの可視化
- ノイズの多いデータの処理
- 高次元データの取り扱い
- 入力データにおける顕著な変動への対応
- 利用可能なデータの適切な活用
- モデルの柔軟性と複雑さの調整
- モデルの学習とチューニングの理解
- **モデルの説明可能性の確保**

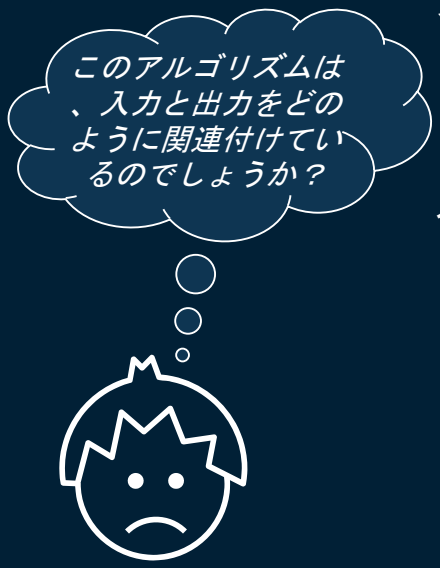
モデルの解釈可能性

ホワイトボックスモデル



モデルの解釈可能性

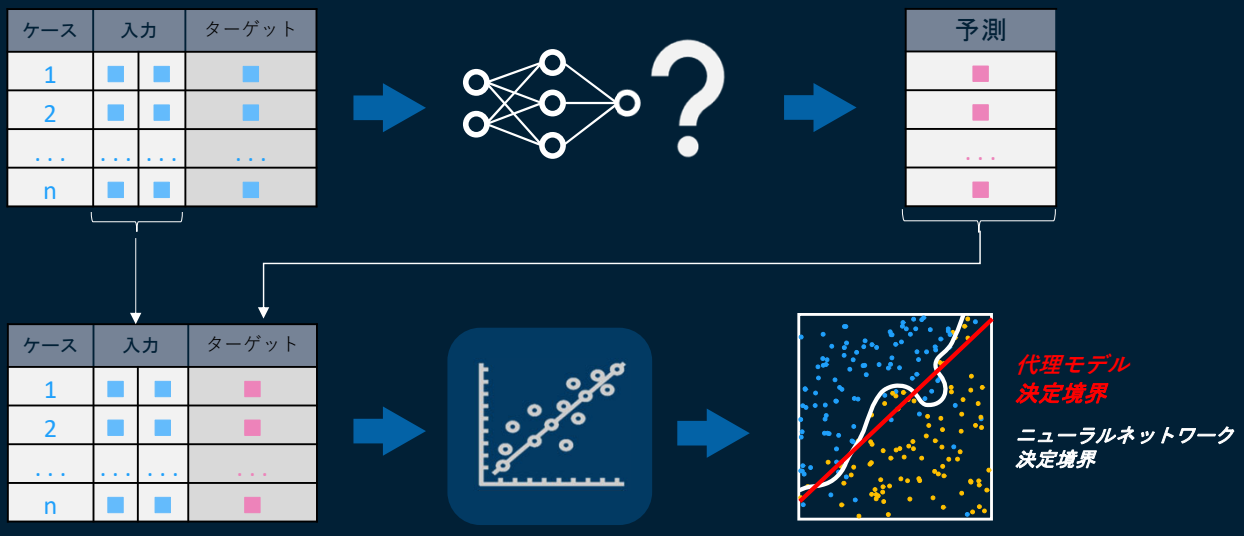
ブラックボックスモデル



入力 → 「ブラックボックス」不透明なモデル → 出力

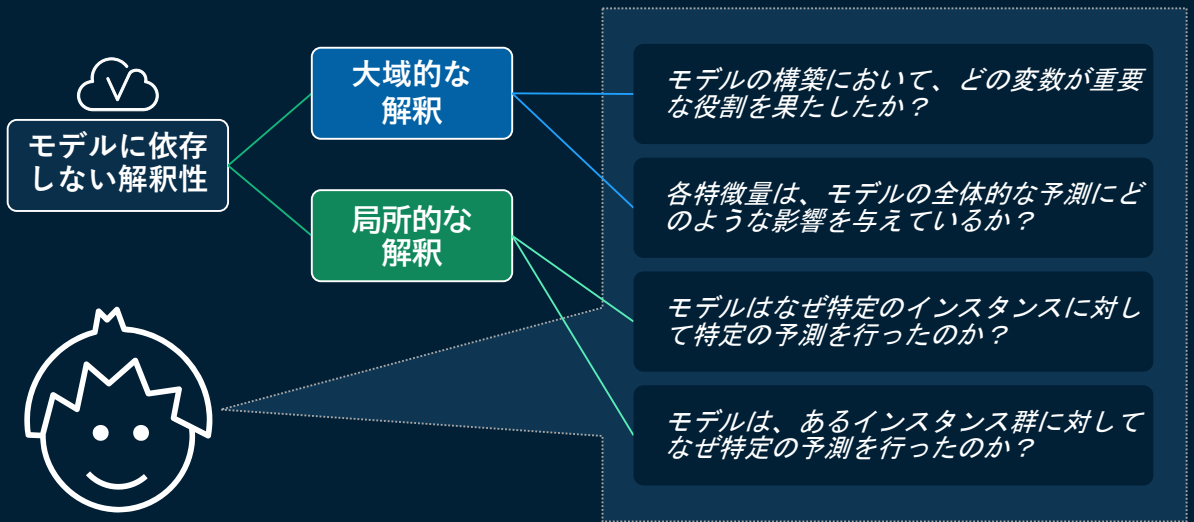
モデルの解釈可能性

代用モデル



モデルの解釈可能性

ブラックボックスモデルの理解



解釈可能性が重要な理由

問題：ブラックボックスモデルの解釈

機械学習モデルの多くはブラックボックスです。しかし、SAS Viyaには、機械学習モデルの結果を解釈するのに役立つ、いくつかのモデル解釈プロットが用意されています。

a. 正しい
b. 誤り

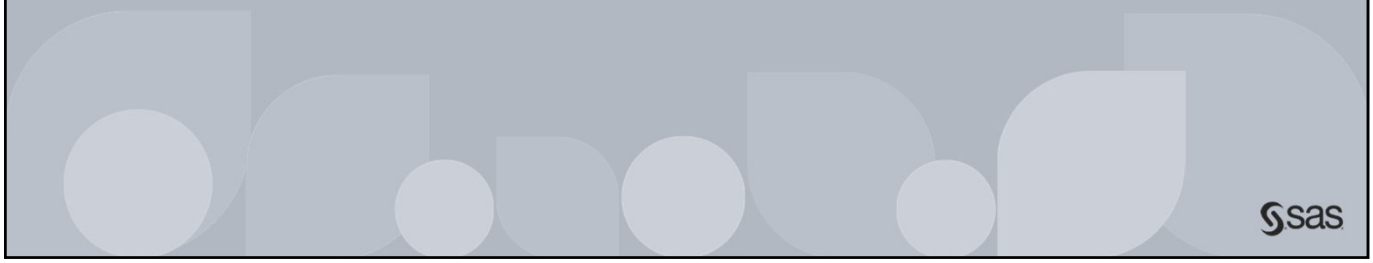
回答：ブラックボックスモデルの解釈

機械学習モデルの多くはブラックボックスです。しかし、SAS Viyaには、機械学習モデルの結果を解釈するのに役立つ、いくつかのモデル解釈プロットが用意されています。

a. 正しい
b. 誤り

ご質問はありますか？

レッスンのまとめ



次に何をすべきでしょうか？

以下のSASコースのいずれかを受講する準備が整っています：

SAS Viya を使用した機械学習

SAS Viya での対話型機械学習

SAS Studio における SAS Viya を使用した教師あり機械学習の手順

SAS Viya における SAS Visual Statistics：対話型モデル構築

機械学習で知っておくべき統計学

2026年8月28日　初版 第1刷

発行元　SAS Institute Japan株式会社

〒106-6661　東京都港区六本木6-10-1 六本木ヒルズ森タワー11F
℡ 03(6434)3690