

Part 2

ビジネスデータ解析に必要なスキルを学ぶ

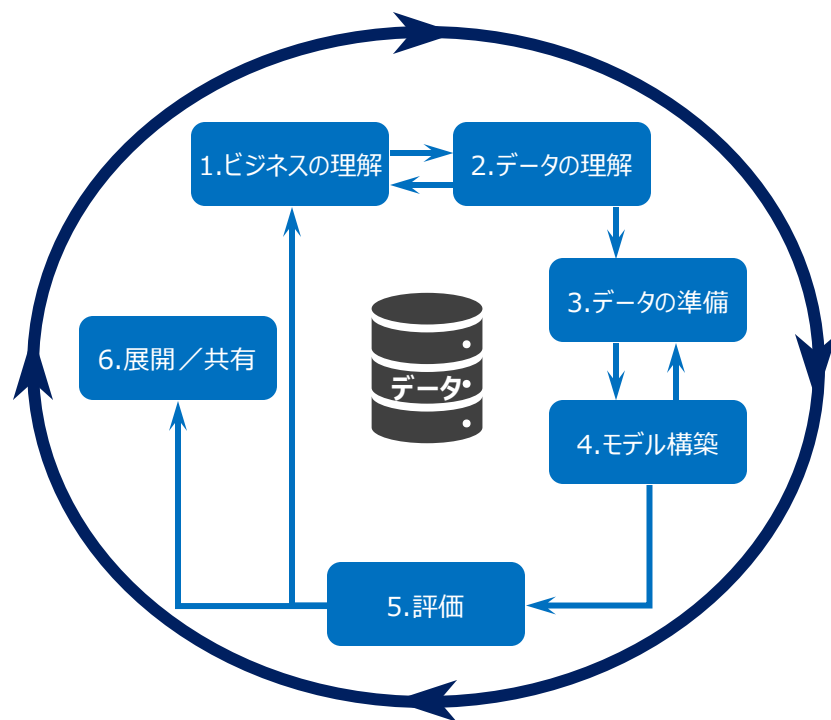
- ✓データ分析の進め方
- ✓分析に必要なビジネス力
- ✓分析に必要なデータサイエンス力とデータエンジニアリング力

データ分析の進め方

- データ分析の進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: 分析方法論

(CRoss Industry Standard Process for Data Mining)



1. ビジネスの理解

- ビジネス、分析目標の決定
- プロジェクトの立ち上げ

ビジネス

2. データの理解

- データの収集
- データの調査
- データ品質の検証

データサイエンス

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

データエンジニアリング

4. モデル構築

- モデリング手法の選択
- モデルの作成
- モデルの評価

データサイエンス

5. 評価

- 分析結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

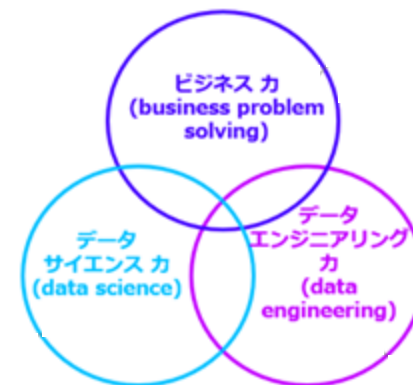
データエンジニアリング

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

ビジネス

一般社団法人
データサイエンティスト協会



分析に必要なビジネス力

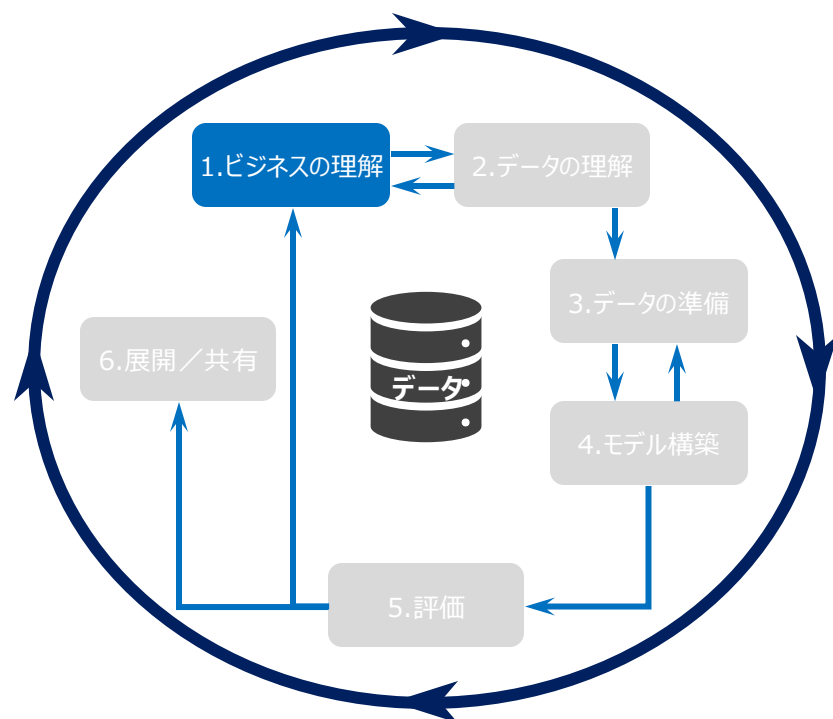
- ✓ 分析テーマの明確化
- ✓ 仮説思考力の重要性

データ分析の進め方

- データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(CRoss Industry Standard Process for Data Mining)



1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル構築

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

「なんかわかるでしょ？」の注意

- データ分析のきっかけは、大体ぼんやりとした依頼から始まることが多い・・・



最近ほら、AIとか流行ってるでしょ。
うちも売上データ使ってなんかできない？
データがあればなんかわかるよね？

売り上げを増やしたいの？
顧客を増やしたいの？

・分析対象（目的変数）
の明確化

今どのくらいの貸し倒れ？
どこまで改善目指す？

なんとか貸し倒れを減らしたいんだけど、
データ分析して原因を見つけられる？

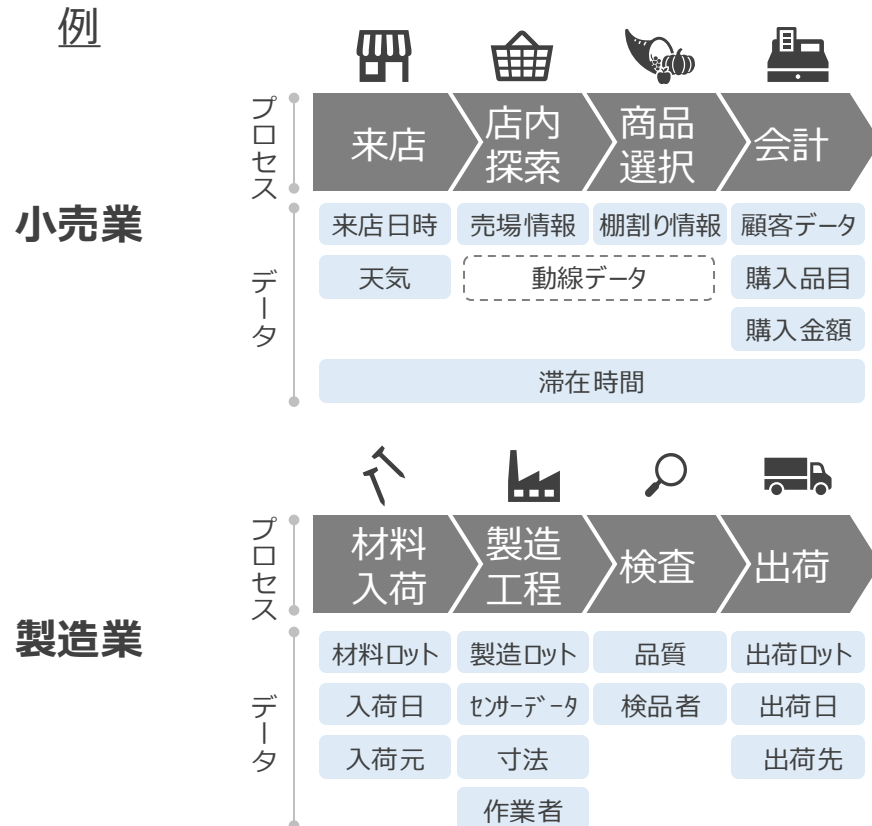


・現状の把握
・分析目標（ゴール）の設定

(参考) ビジネスの理解＝「プロセス」と「構造」の理解

- データに触る前に、分析対象ビジネスの「プロセス」と「構造」を理解しておくことが重要である
- これらと並行して、分析スコープと分析目的の検討、データの理解も行う

1 「ビジネスプロセス」を理解する

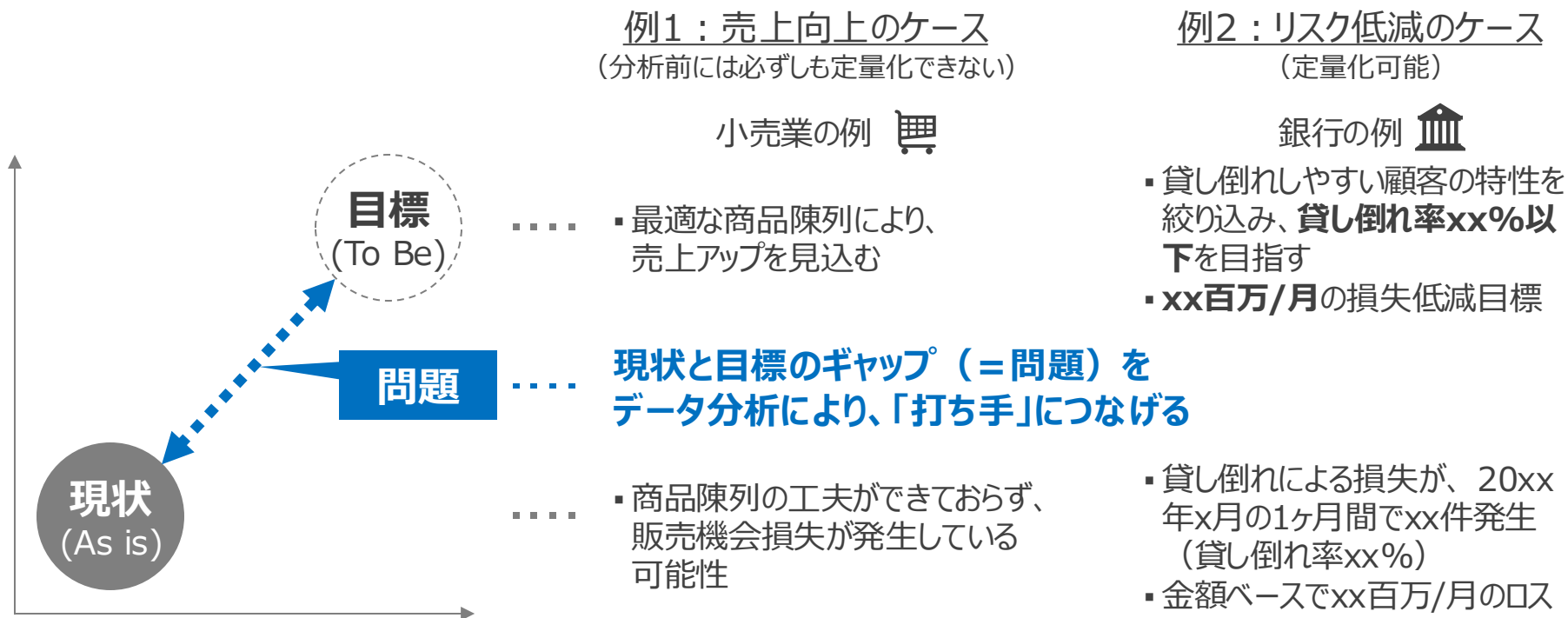


2 「ビジネス構造」を理解する



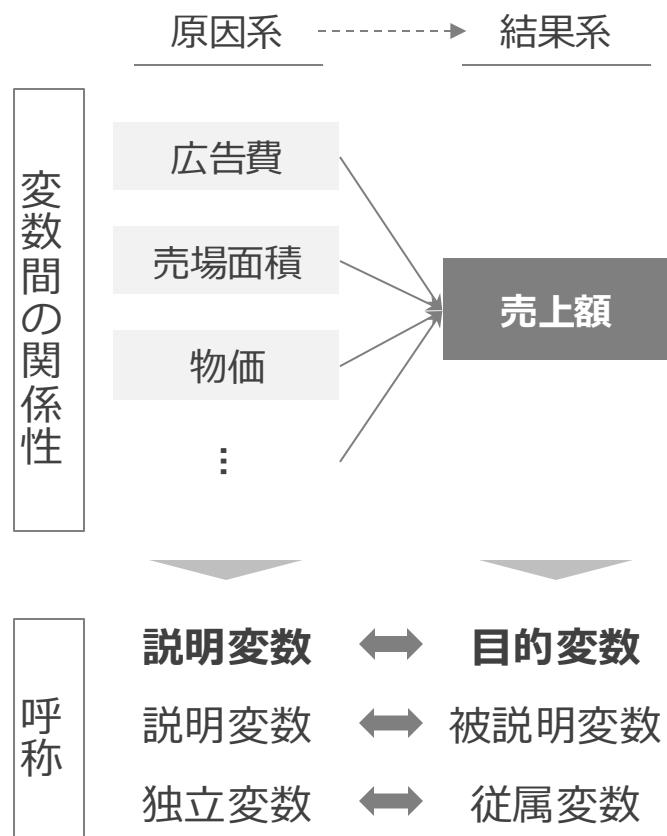
(参考) 現状把握と目標の設定

- 理解したビジネスのプロセス・構造に基づいて、現状の把握・分析目標の設定を行ない、**データ分析により解くべき問題を明確化する**
- (特にコスト削減のケースでは) **可能な限り定量的な現状把握、目標設定を行う**



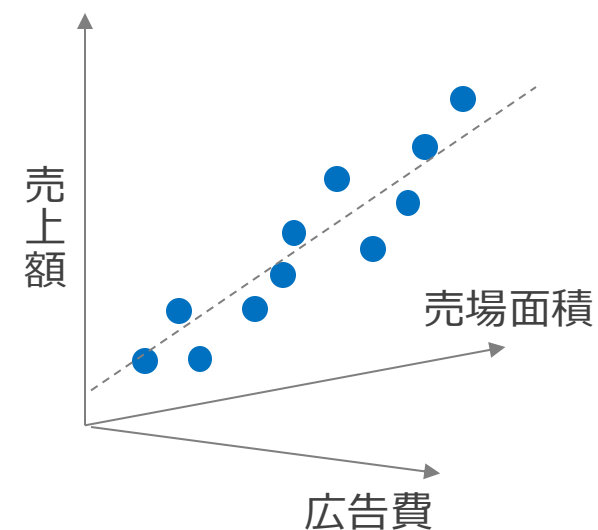
目的変数・説明変数とは

- 説明変数は四則演算などで必要に応じて新規変数を作る
- 目的変数は、必要に応じてフラグ値化／カテゴリ値化などの加工をする
(分布の形を見てbinの幅や閾値を判断する)



重回帰の場合：

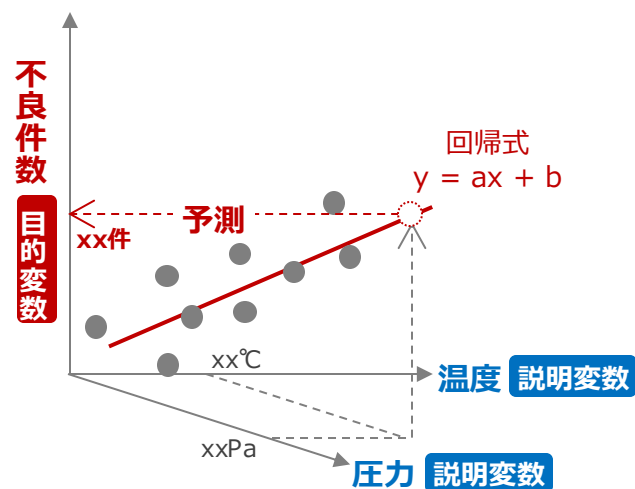
$$\text{売上額} = a_1 \cdot \text{広告費} + a_2 \cdot \text{売場面積} + a_3 \cdot \text{物価} + \dots$$



目的変数・説明変数の検討

- 分析目的やデータ形式に応じて、目的変数や説明変数の形式は異なる
「どんなデータを用いてどんな分析をしたいか」を明確にすることが重要

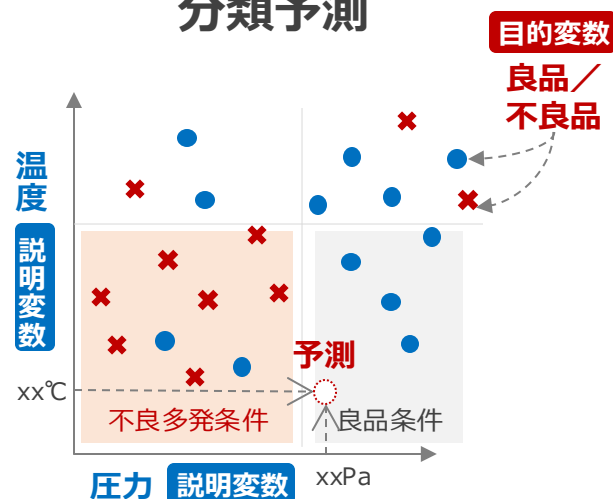
数値予測



分析例

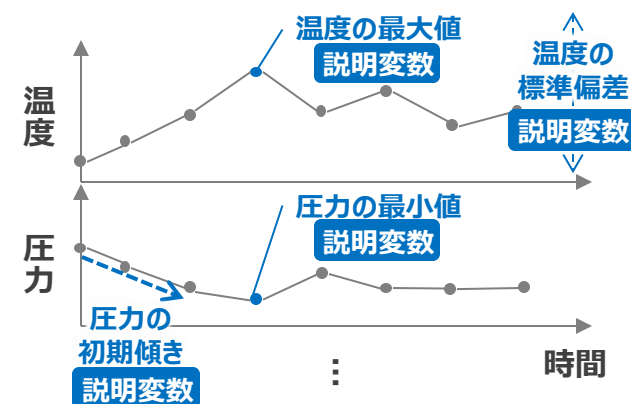
- 小売：売上金額の予測
- 小売：需要予測
- 製造：不良発生率の予測

分類予測



- 小売：購入／非購入顧客の分類
- 医療：生存条件の分析
- 製造：故障種類の分類

時系列分析



目的変数 良品／不良品

- 設備稼働の異常検知
- リアルタイム不良・不具合検出

統計学の基礎

- ✓ 記述統計学 と 推測統計学
- ✓ 要約統計量 (代表値 / ばらつき指標 / 順序統計量)

統計の落とし穴

- 以下は、日本国内でのある統計調査の結果です
- みなさん、こんな危険な食べ物、即刻禁止にすべきと思いませんか？
 - ・ **心筋梗塞による死亡者の95%以上**が生前ずっとこの食べ物を食べていた
 - ・ **がん患者の98%**がこの食べ物を摂取していた
 - ・ **強盗や殺人などの凶悪犯の70%以上**が犯行の24時間以内にこの食べ物を口にしている

西内 啓「統計学が最強の学問である」より



「お米」 禁止しますか？

2つの「統計学」

■ 国勢調査と視聴率調査は何が違う？

国勢調査



Source: https://www.stat.go.jp/data/kokusei/2020/kekka/pdf/summary_01.pdf

視聴率調査

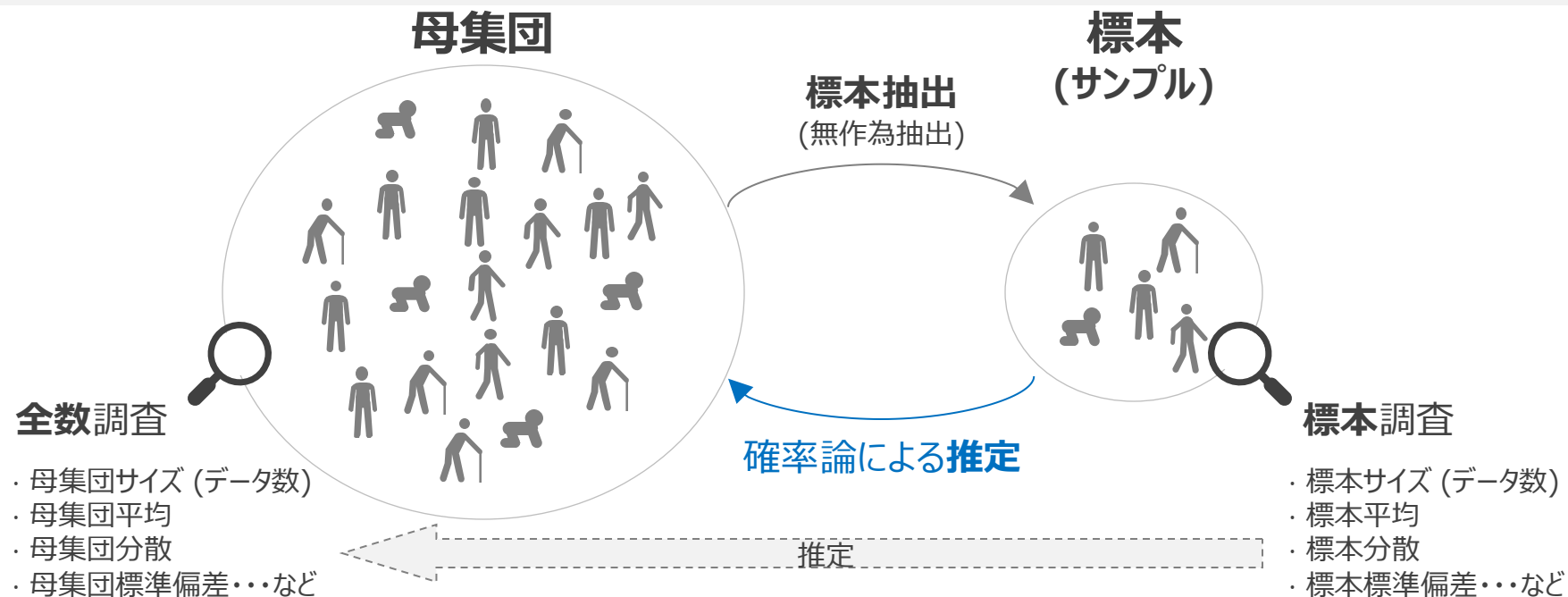
順位	番組タイトル	放送局	視聴率
1	連続テレビ小説・らんまん（7/14）	NHK総合	17.5 %
2	ザワつく！金曜日（7/14）	テレビ朝日	13.1 %
3	ボツンと一軒家3時間SP（7/16）	テレビ朝日	12.3 %
4	大相撲 2023年 名古屋場所・8日目（7/16 17:06～）	NHK総合	12.1 %
5	日曜劇場・VIVANT（7/16）	TBS	11.5 %
6	プラタモリ（7/15）	NHK総合	11.3 %
7	1億人の大質問!?笑ってコラえて！ 27周年2時間SP（7/12）	日本テレビ	11.1 %
8	世界まる見え！テレビ特捜部 2時間SP（7/10）	日本テレビ	10.6 %
9	木曜ドラマ・ハヤブサ消防団（7/13）	テレビ朝日	10.5 %
10	音楽の日2023（7/15 19:00～）	TBS	10.1 %

2023年7月10日(月)～7月16日(日) 視聴率ランキング

Source: <https://www.tvguide.or.jp/tvranking/tvrank-20230728/>

2つの「統計学」－ 記述統計学 と 推測統計学

- データ全体を調べてデータの特徴や傾向を把握する「記述統計学」と、一部のサンプルを抽出して調べ、全体の特徴や傾向を推測する「推測統計学」がある



記述統計学

概要

全数調査により観測されたデータ全体の特徴を調べたり、データを要約するアプローチのこと

例

国勢調査、(学校の)健康診断・定期テスト など

例えば
企業では...

(CRM内の) 顧客データの分析 など

推測統計学

データ全体 (母集団) を調べあげるのが困難な場合に、抽出したデータの一部 (標本) の特徴から、データ全体の特徴を確率的に推測するアプローチのこと

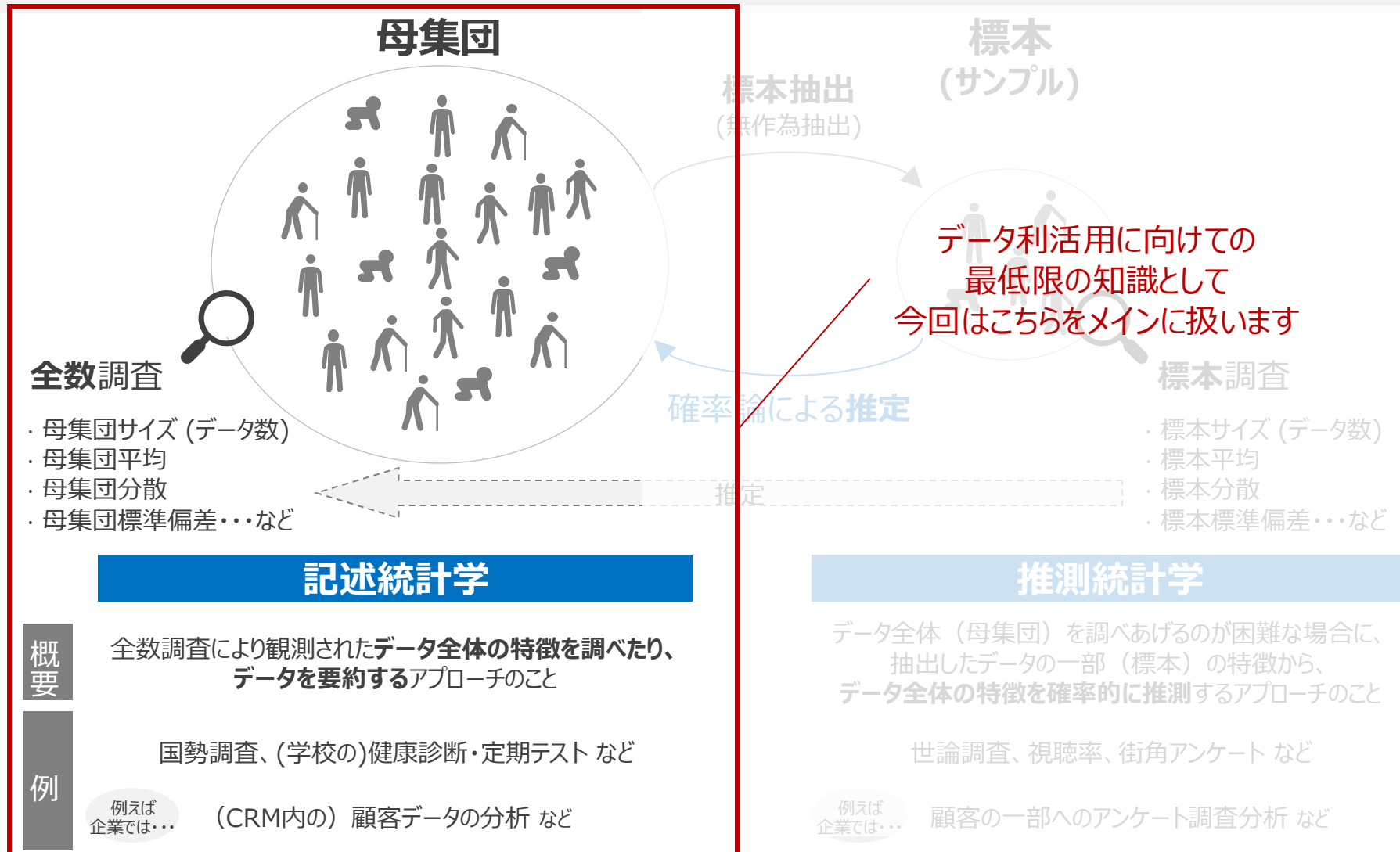
世論調査、視聴率、街角アンケート など

例えば
企業では...

顧客の一部へのアンケート調査分析 など

2つの「統計学」－ 記述統計学 と 推測統計学

- データ全体を調べてデータの特徴や傾向を把握する「記述統計学」と、一部のサンプルを抽出して調べ、全体の特徴や傾向を推測する「推測統計学」がある

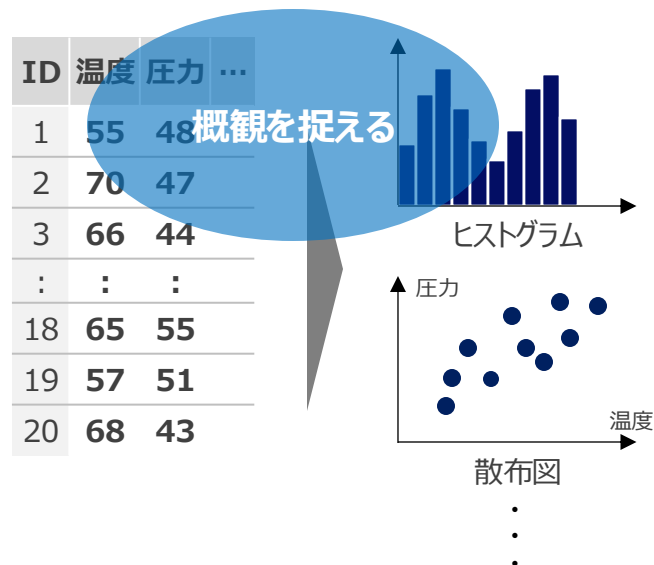


記述統計学の基本：データの特徴を捉える

- 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や要約統計量（平均値や標準偏差）を用いて全体傾向を把握する

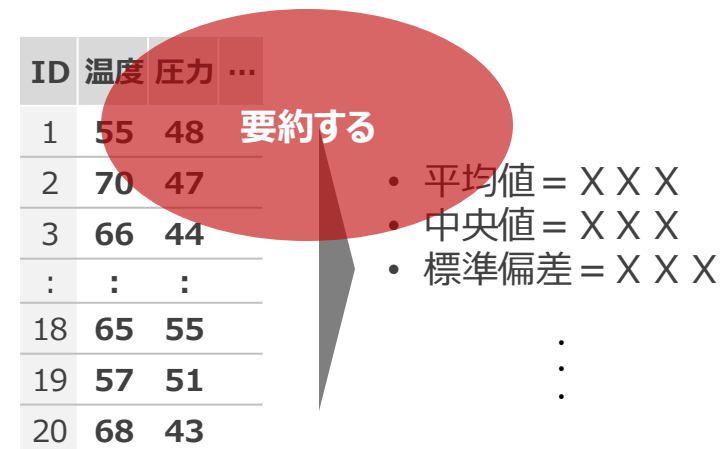
可視化（グラフ化）

視覚的にデータの特徴や傾向を把握



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
(要約統計量)

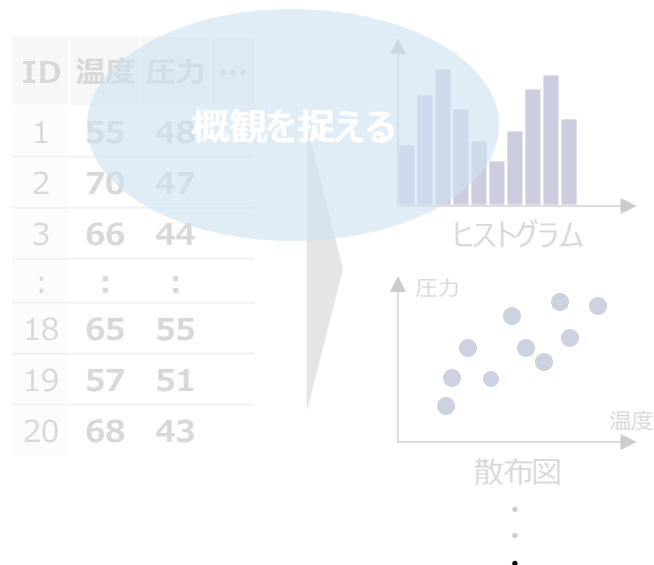


記述統計学の基本：データの特徴を捉える

- 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や要約統計量（平均値や標準偏差）を用いて全体傾向を把握する

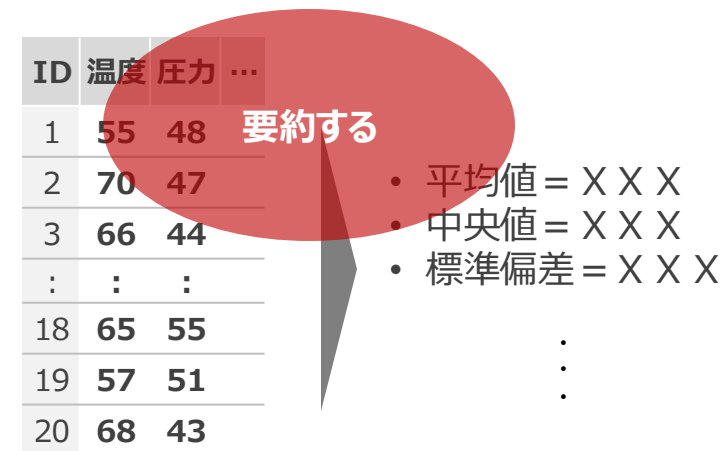
可視化（グラフ化）

視覚的にデータの特徴や傾向を把握



数値化（データ要約）

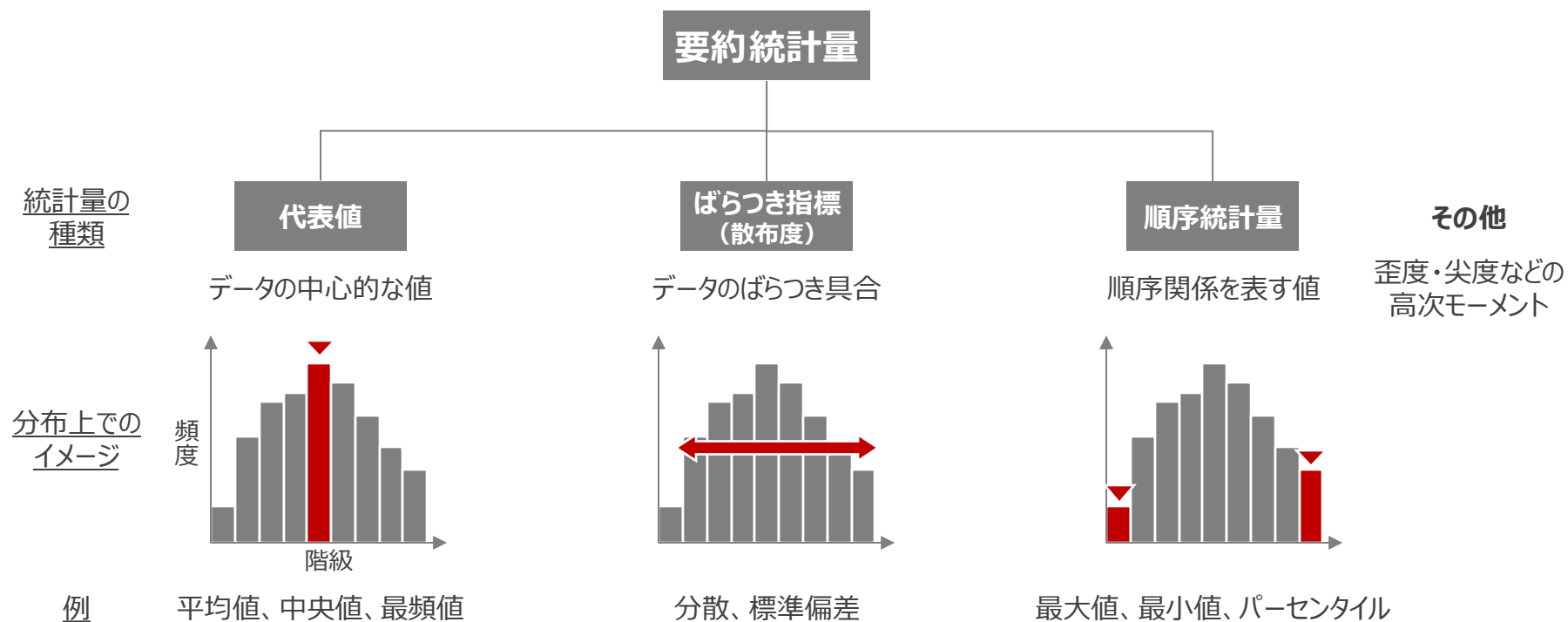
データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
(要約統計量)



※可視化については、別途「データの理解（可視化）」のプロセスで取り扱う

数値化（データ要約）

- 全データをくまなく見るのは難しいため、「要約した値（＝要約統計量）」を用いて全体傾向を把握するのが一般的



代表値（平均値）

- 平均値は最もよく用いられる代表値であるが、分布の形によっては思わぬ落とし穴がある

ある会社の社員9名の年収

社員	年収
A	200万円
B	200万円
C	200万円
D	400万円
E	400万円
F	500万円
G	700万円
H	900万円
I	1億円

平均値の算出

$$\begin{aligned}\text{平均値} &= \frac{200\text{万円} + 200\text{万円} + \cdots + 1\text{億円}}{9\text{名}} \\ &= \mathbf{1,500\text{万円}}\end{aligned}$$

この平均値、直感と合っている？



代表値（中央値・最頻値）

- 中央値：データを値の小さい順に並べたときに中央に位置する値
- 最頻値：データの中で最も多く出現する値。複数存在するケースもある

ある会社の社員9名の年収

中央値と最頻値

社員	年収
A	200万円
B	200万円
C	200万円
D	400万円
E	400万円
F	500万円
G	700万円
H	900万円
I	1億円

最頻値 = 200万円

社内で最も多い年収値

中央値 = 400万円

社内の年収を小さい順に並べ、
ちょうど真ん中に位置する年収値

外れ値

中央値・最頻値は外れ値の影響を受けにくい！

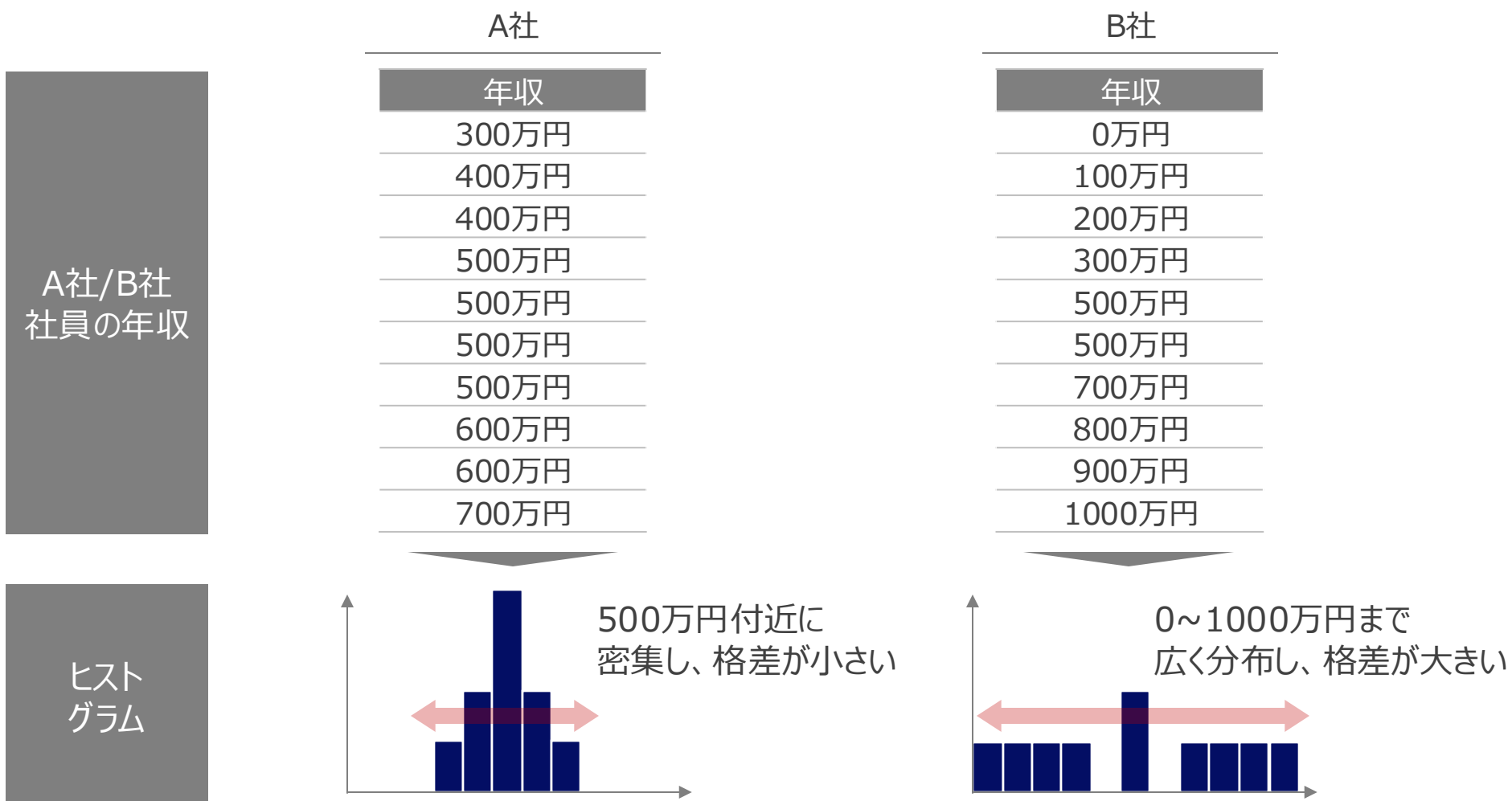
代表値だけではつかめない特徴

- 代表値はすべて同じ。どちらの会社に入っても待遇は同じ！と言ってよいのか？

A社/B社 社員の年収		A社	B社
代表値	平均値	500万円	500万円
	中央値	500万円	500万円
	最頻値	500万円	500万円
		年収	年収
		300万円	0万円
		400万円	100万円
		400万円	200万円
		500万円	300万円
		500万円	500万円
		500万円	500万円
		500万円	700万円
		600万円	800万円
		600万円	900万円
		700万円	1000万円

代表値だけではつかめない特徴

- ヒストグラムで見ると、社員によって年収が大きく異なっていることがわかる



データの「ばらつき」の表し方

- データのばらつきは、「平均からのズレ」= 偏差 として捉えるのが基本的な考え方

A社/B社 社員の年収	A社		B社	
	年収	平均値との差	年収	平均値との差
	300万円	-200万円	0万円	-500万円
	400万円	-100万円	100万円	-400万円
	400万円	-100万円	200万円	-300万円
	500万円	0万円	300万円	-200万円
	500万円	0万円	500万円	0万円
	500万円	0万円	500万円	0万円
	500万円	0万円	700万円	+200万円
	600万円	+100万円	800万円	+300万円
	600万円	+100万円	900万円	+400万円
	700万円	+200万円	1000万円	+500万円
	平均年収 500万円		平均年収 500万円	

データの「ばらつき」の計算

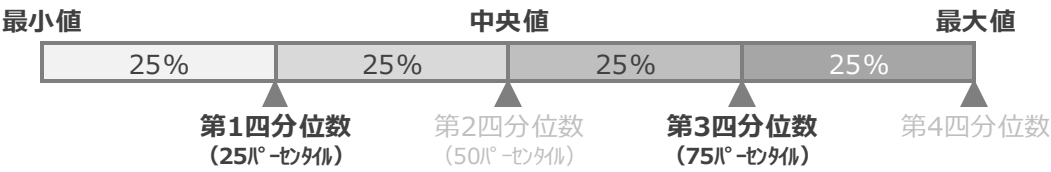
- 分散：平均からのズレを、（正負の符号を消すために）2乗して平均した値
- 標準偏差：元のデータと単位を揃えるために、分散の値を変換した値

A社				B社			
A社/B社 社員名 社員番号	年収	平均値との差		年収	平均値との差		
	300万円	-200万円		0万円	-500万円		
	400万円	-100万円		100万円	-400万円		
	400万円	-100万円		200万円	-300万円		
	500万円	0万円		300万円	-200万円		
	500万円	0万円		500万円	0万円		
	500万円	0万円		500万円	0万円		
$\text{A社年収の分散} = \frac{(-200\text{万円})^2 + (-100\text{万円})^2 + \dots + (100\text{万円})^2 + (200\text{万円})^2}{10\text{名}}$							
$= \boxed{\text{A社の年収標準偏差}} \times \text{A社の年収標準偏差}$							
実際に標準偏差を計算すると、A社は109.5万円、B社は328.6万円							

代表的な要約統計量まとめ

- どれか一つを算出すれば良いわけではなく、様々な観点でデータの概観を把握することが重要
- また、時系列データの分析ではこれら統計量自体を「特徴量」として扱うことも多い

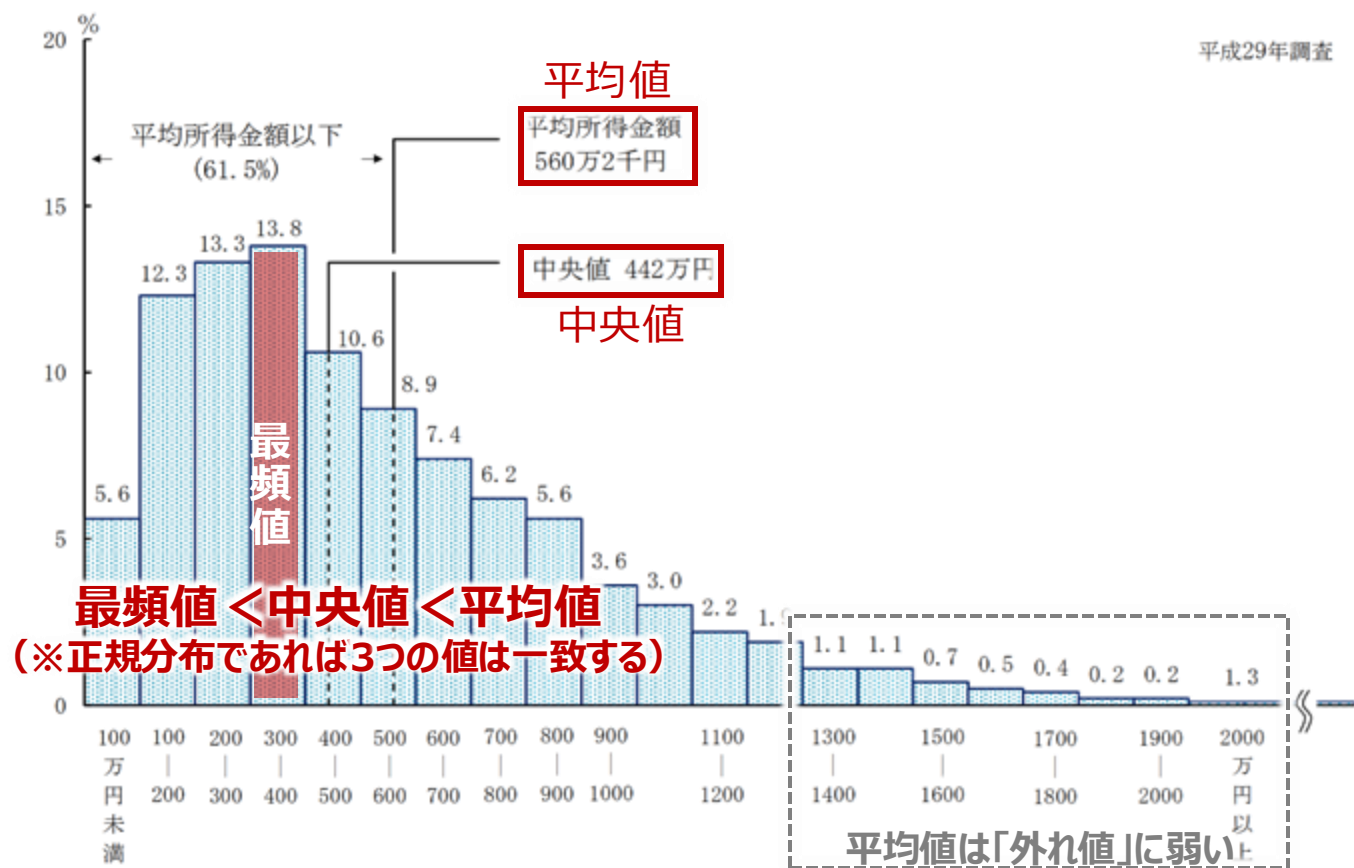
分類	統計量	説明
代表する値 (代表値)	平均	データの中心位置を表す値
	中央値 (メディアン)	データを小さい順に並べたとき、ちょうど中央の位置にくる値
	最頻値 (モード)	データの中で最頻出する値
ばらつき具合 を示す値 (散布度)	分散	データのばらつき具合を表す値
	標準偏差	データのばらつき具合を表す値 元データと 同じ単位 のため直感的なばらつき把握が可能
順序を表す値 (順序統計量)	最大値／最小値	データの中で最も大きい／小さい値
	第1／第3四分位数	データを小さい順に並べたとき、 それぞれ25%／75% の位置にくる値



(参考) 年収分布における代表値の差異

- 正規分布ではない分布では、平均値だけを採用するとデータの特徴を見誤る恐れがある

図9 所得金額階級別世帯数の相対度数分布



出典：厚生労働省 平成 29 年 国民生活基礎調査の概況

データ観察の基本

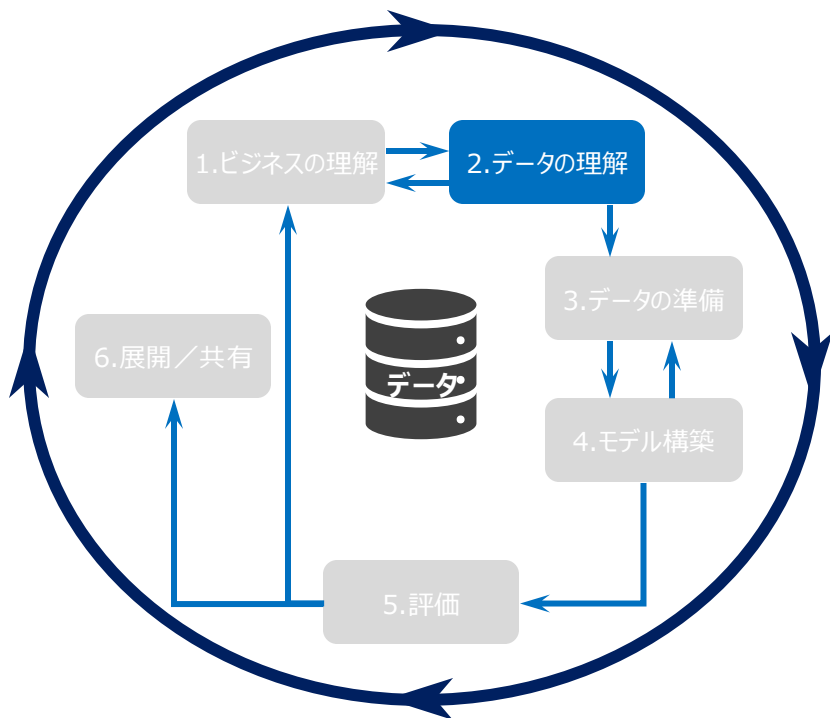
- ✓ グラフと要約統計量によるデータの傾向把握
- ✓ 相関分析によるデータ間の関係性把握
- ✓ 相関と因果の違い

データ分析の進め方

- データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(**C**Ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)



1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3.データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4.モデル構築

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6.展開／共有

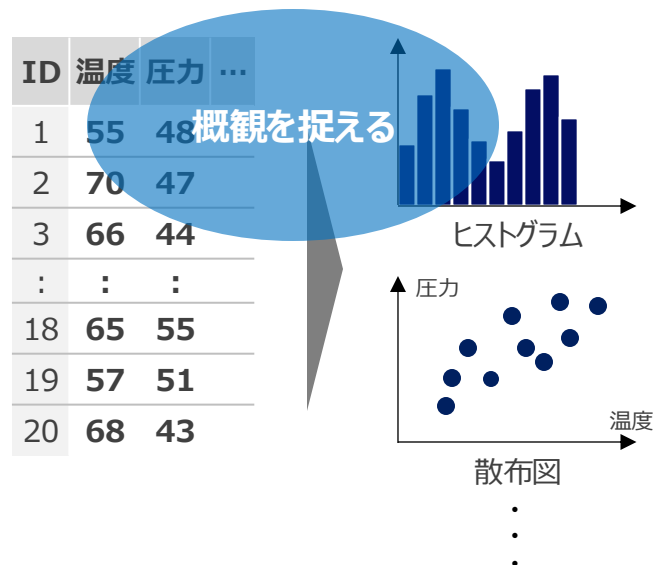
- 業務への導入計画
- モニタリング、メンテナンスの計画

記述統計学の基本：データの特徴を捉える

- 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や要約統計量（平均値や標準偏差）を用いて全体傾向を把握する

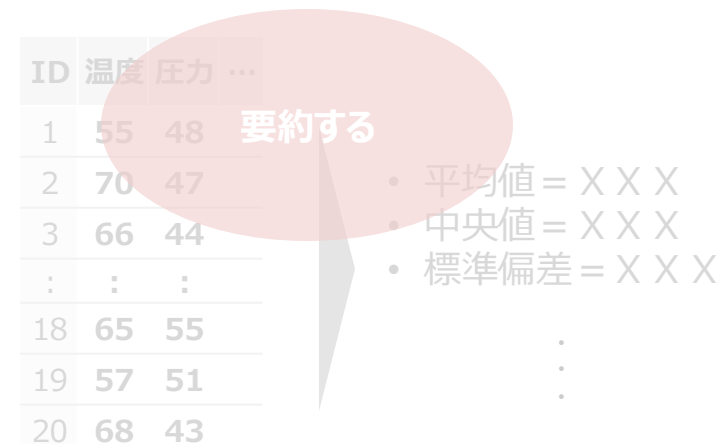
可視化（グラフ化）

視覚的にデータの特徴や傾向を把握



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
(要約統計量)



集団の比較

- あなたは、「飲むログ」でデートで使うお店を選んでいきます。候補となったお店は2つ。でもレビュー点数が全く同じです。
- 特徴を捉えるにはどうすれば良い？

レストランA



3.45

#	評点
1	3.5
2	3.5
3	4.5
4	3.5
5	3.5
6	2.5
7	3.0
8	3.5
9	4.0
10	3.0

#	評点
11	3.0
12	3.5
13	3.0
14	4.0
15	4.0
16	4.0
17	2.5
18	4.5
19	1.0
20	5.0



どちらも同じ
評価だから
どちらでも
いいのかな...

レストランB



3.45

#	評点
1	3.0
2	4.5
3	5.0
4	4.5
5	2.5
6	2.5
7	4.0
8	2.0
9	5.0
10	2.0

#	評点
11	1.0
12	2.0
13	4.5
14	2.0
15	5.0
16	4.5
17	4.5
18	4.0
19	2.5
20	4.0

ヒストグラム（度数分布図）

- ヒストグラムは、データを一定間隔（**階級**）ごとに頻度集計（**度数**）をとり、横軸に階級、縦軸に度数をとって柱状（**ビン**）の集合で表したグラフ
- データの分布状況を可視化して、直感的にデータの特徴を捉えやすくする目的がある

元データ

レストランAの
レビューアごとの評点データ

#	評点	#	評点
1	3.5	11	3.0
2	3.5	12	3.5
3	4.5	13	3.0
4	3.5	14	4.0
5	3.5	15	4.0
6	2.5	16	4.0
7	3.0	17	2.5
8	3.5	18	4.5
9	4.0	19	1.0
10	3.0	20	5.0

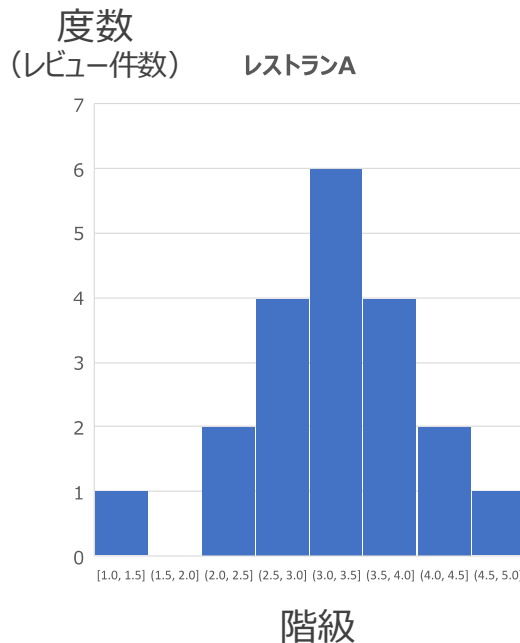
度数分布表

一定間隔（階級）ごとに
レビュー件数（度数）を集計

階級	度数
0.0~0.5	0
0.5~1.0	0
1.0~1.5	1
1.5~2.0	0
2.0~2.5	2
2.5~3.0	4
3.0~3.5	6
3.5~4.0	4
4.0~4.5	2
4.5~5.0	1

ヒストグラム（度数分布図）

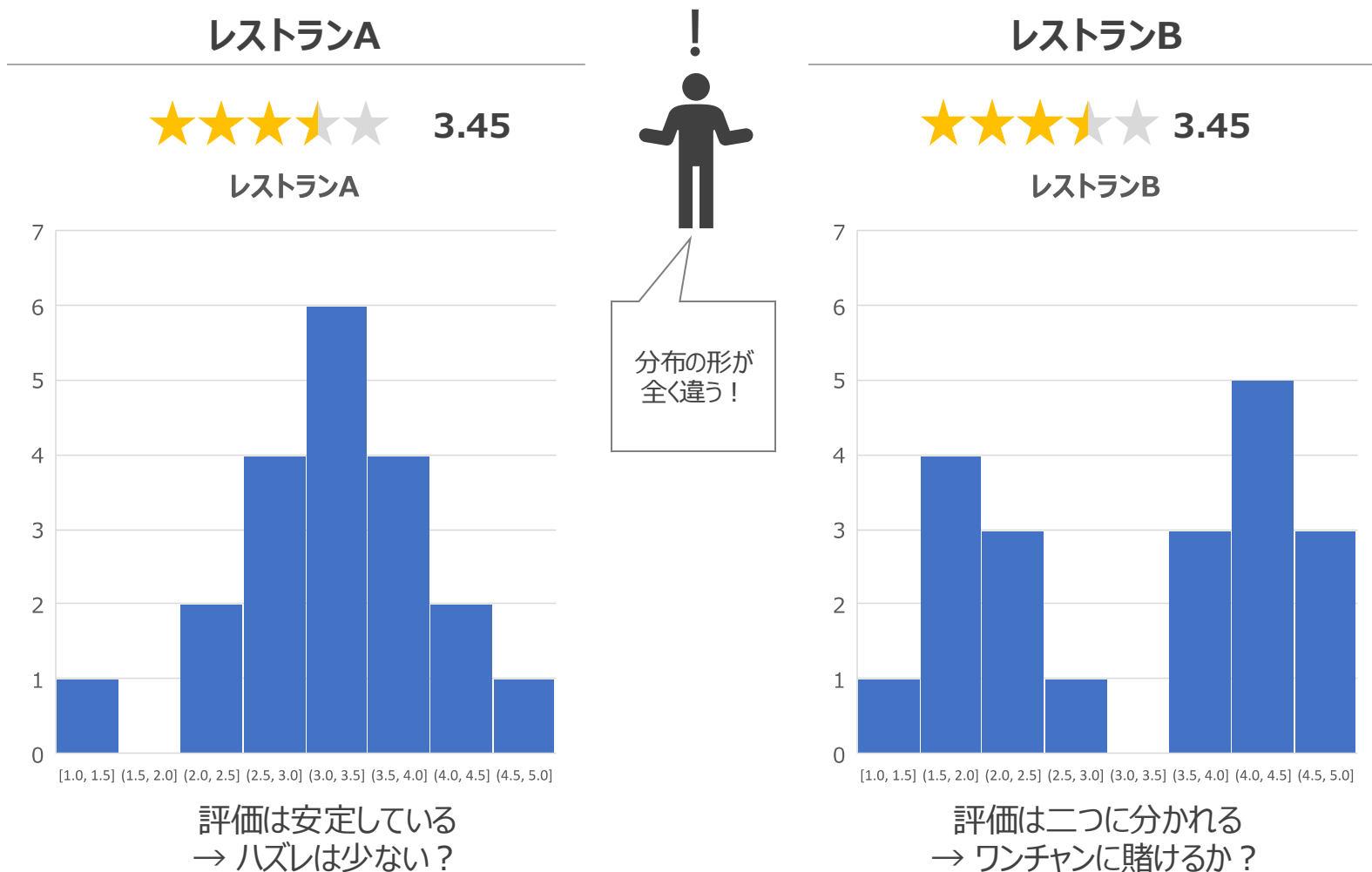
横軸に階級、縦軸に度数を取り、
柱（ビン）の集合で可視化



※境界値は小さい側の階級に含める
(ただし、1.0は便宜上1.0~1.5に含める)

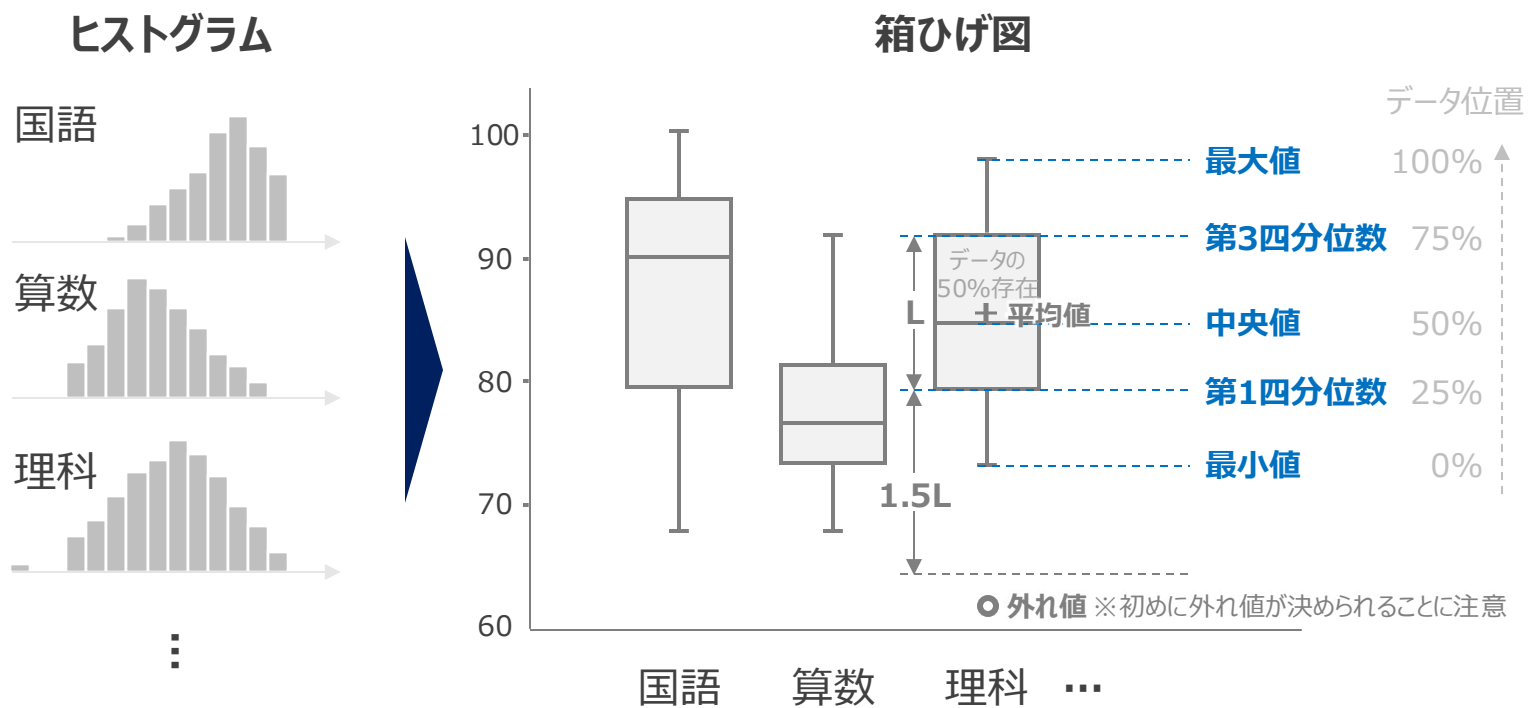
ヒストグラムの描画と比較

■ ヒストグラムを描くと、「分布の形」（≡データの散らばり具合）が可視化される



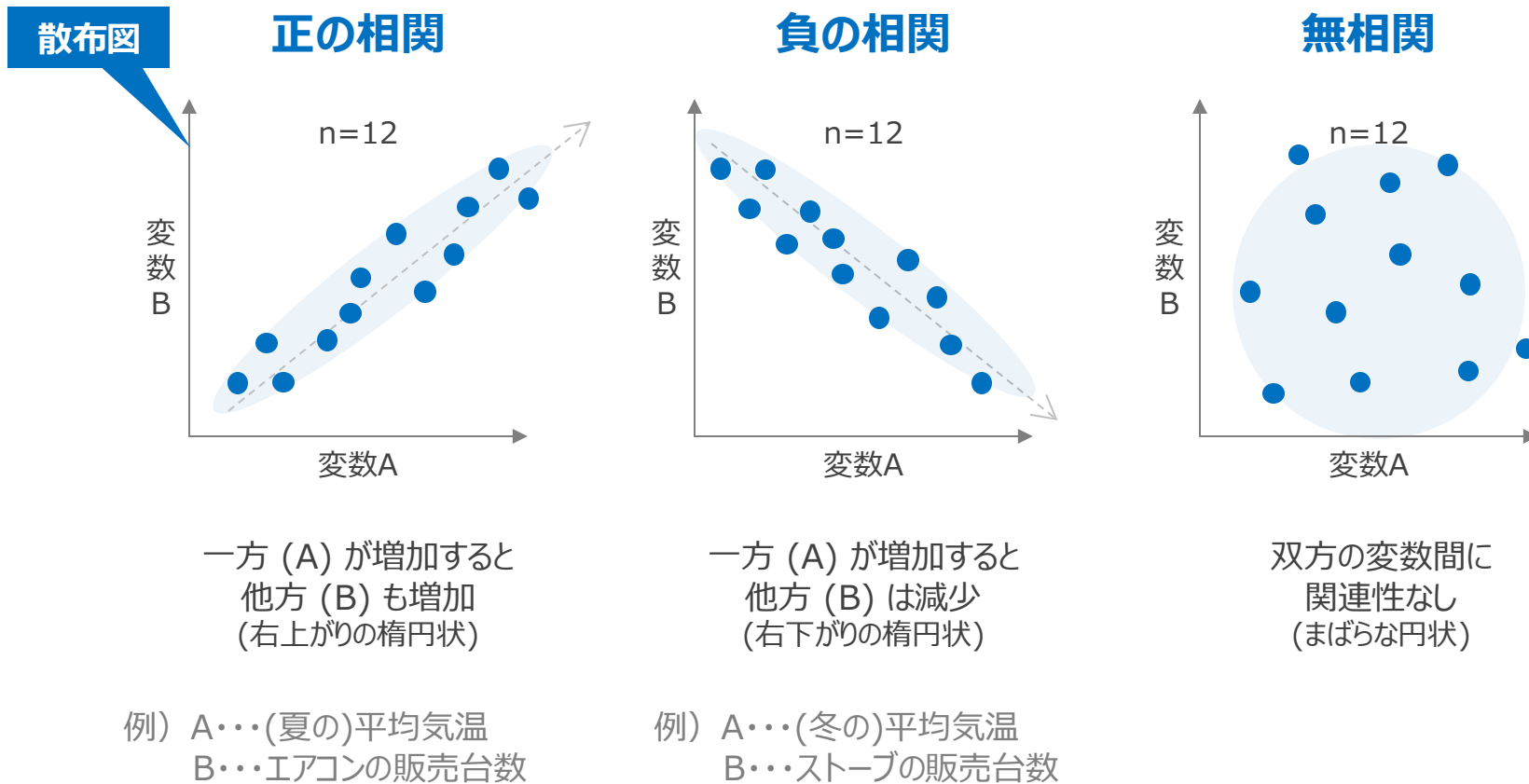
箱ひげ図

- 分布形状の把握は極めて重要であるが、主にヒストグラムは1変数、散布図は2変数での分布可視化に適している。一方、箱ひげ図は、分布の概形を**多変数間で比較**するのに適している



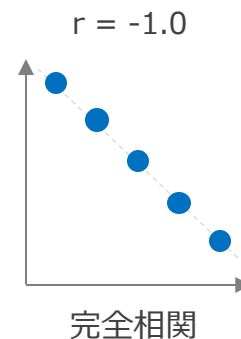
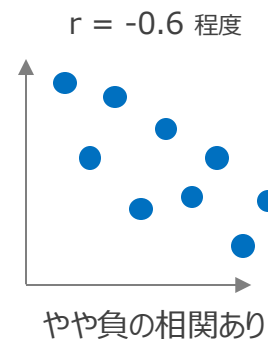
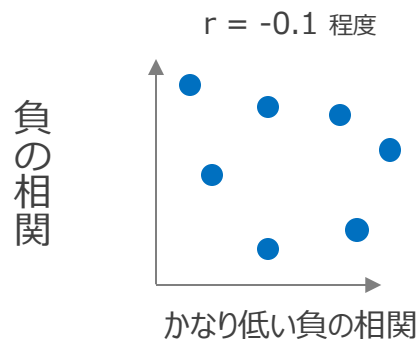
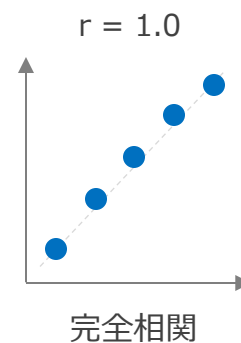
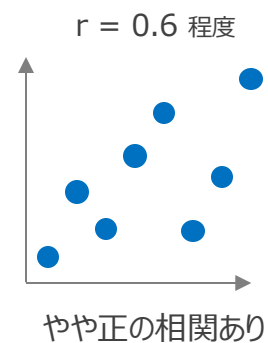
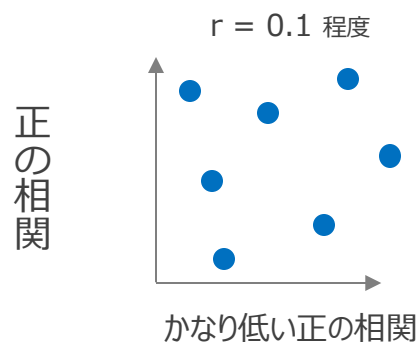
散布図 と 相関

- 二つの値（**変数**）間において、一方が上がれば他方も上がる（or 下がる）ような関係性のことを「**相関**」と呼ぶ
- 各変数を各軸にとってグラフ化した「**散布図**」を描くことで、相関関係の視覚的な把握が可能



相関係数

- 相関係数 r (correlation coefficient) とは、2つの変数間の相関の度合いを表す指標
- $-1 \leq r \leq 1$ の値を取り、正の場合は正相関、負の場合は負相関、0の場合は無相関



(参考) 相関係数の目安

相関係数 (絶対値)	相関の強弱
1.0	強い相関
0.7	やや相関あり
0.4	弱い相関
0.2	ほぼ相関なし
0	

相関行列による網羅的な変数間の関係性把握

- 事前に各変数同士の相関係数を総当たりで調べておくと、後々の結果解釈に役立つ（**相関行列**）
- また、**共線性が高い変数**（相関の高い）が複数混ざっていると、その変数の影響を強く受け、偏った分析結果になることがある。この場合、共線性が高い変数は除外することが有効

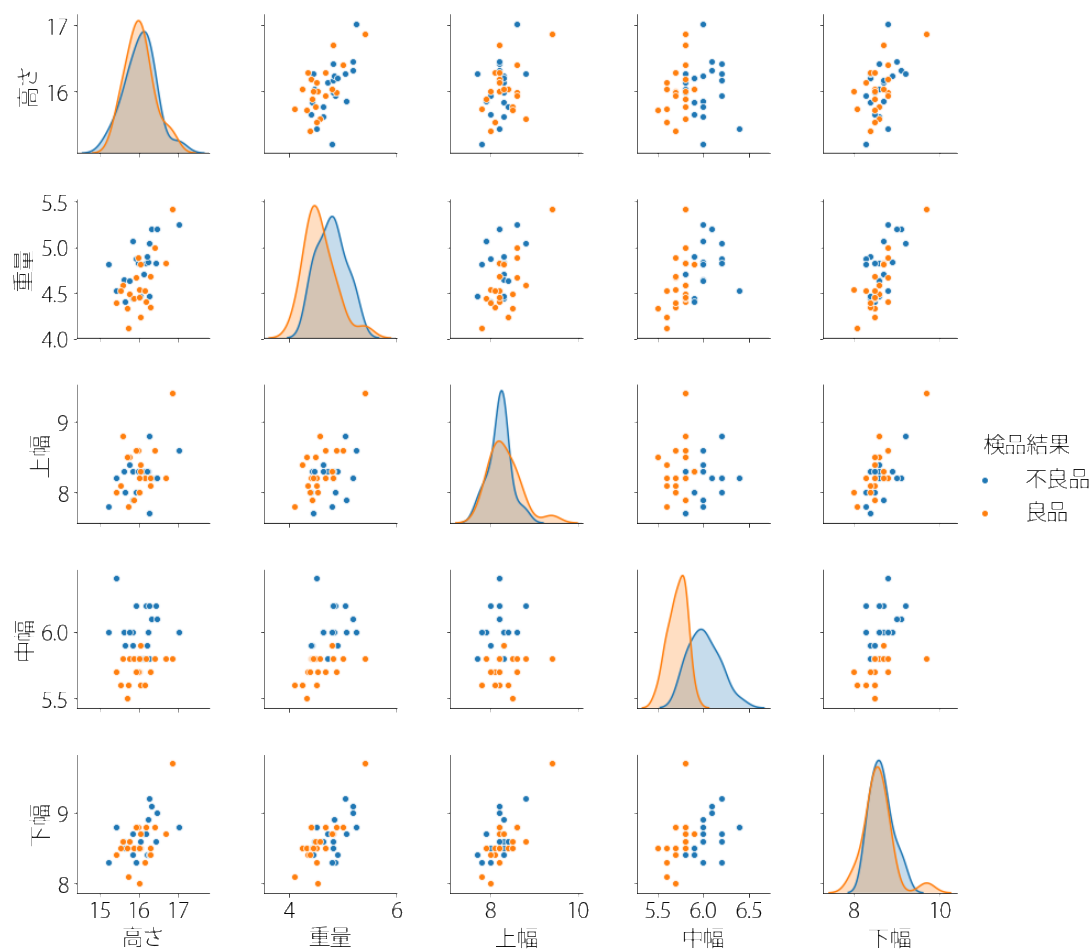
	打率	試合	打席数	打数	安打	本塁打	打点	盗
打率	1.00							
試合	-0.22	1.00						
打席数	0.02	0.85	1.00					
打数	-0.03	0.83	0.95	1.00				
安打	0.56	0.58	0.80	0.81	1.00			
本塁打	0.00	0.34	0.48	0.38	0.30	1.00		
打点	0.07	0.40	0.48	0.39	0.36	0.92	1.00	
盗塁	0.13	0.28	0.44	0.38	0.42	0.13	-0.17	1.00
四球							0.60	
死球	0.20	0.14			0.11	-0.16	-0.02	
三振	-0.34	0.44	0.47	0.41	0.13	0.63	0.57	
犠打	-0.33	0.12	0.05	0.09	-0.11	-0.49	-0.54	
併殺打	-0.06	0.08	0.10	0.12	0.05	0.32	0.46	
出塁率	0.73	-0.10	0.10	-0.14	0.32	0.31	0.34	
長打率	0.38	0.11	0.32	0.19	0.38	0.90	0.86	
OPS	0.55	0.04	0.27	0.09	0.40	0.77	0.76	
RC27	0.63	0.04	0.28	0.08	0.43	0.65	0.64	
XR27	0.59	0.07	0.30	0.09	0.42	0.70	0.67	

各変数同士の総当たり形式で相関係数を算出
(相関行列)

散布図行列と色分け可視化

- 各変数を個別に見るだけでなく、「**散布図行列**」として網羅的に各変数の関係性を俯瞰したり、**目的変数に応じて色分け**することで、影響因子を仮説立てできることもある

目的変数で色分けした
散布図行列の例





Google Colaboratory上での レクチャー&演習

(参考) 変数の尺度 (名義尺度・順序尺度・間隔尺度・比例尺度)

- 変数の種類は大きく「**質的データ** (カテゴリーデータ)」と「**量的データ** (数量データ)」に分けられ、それぞれの特性に合わせて扱う必要がある
- 数量データは先述の記述統計量、グラフが有効であるが、カテゴリーデータはクロス集計が有効

種類	変数の尺度	概要	データの例	扱い方
				大小 (A<B) 差分 (A-B) 比率 (A/B)
質的データ (カテゴリーデータ)	名義尺度	単にデータを区別するための分類ラベル。 演算不可で、順序も意味をなさない	・性別、血液型、顧客ID ・作業者、個品ID、 良品/不良品	- - - ※集計によるカウントのみ可能
	順序尺度	順序 (大小関係) にのみ意味がある尺度。 したがって、平均値は意味を持たないが、順序統計量 (最大・最小など) は算出可能	・顧客満足度、震度 ・書道の「段」や検定の「級」 ・不良レベル、工程順序	● - -
量的データ (数量データ)	間隔尺度	数値演算可能だが、 値の差 のみに意味がある尺度。 0はあくまで相対的な位置関係でしかない	・年齢、西暦、偏差値 ・温度 (℃)、製造日時	● ● -
	比例尺度	数値演算可能で、値の差に加え、 値の比 にも意味がある尺度。 0が「何もない」という絶対的な意味を持つ	・身長、売上金額 ・寸法、圧力、作業時間、 絶対温度	● ● ●

平均値に意味あり

(参考) クロス集計とは

- カテゴリーデータは、平均値や標準偏差などの記述統計量を算出してもあまり意味がないため、当該データ項目に登場する**カテゴリごとの頻度集計**を算出することが多い
- 特に、複数のデータ項目を掛け合せて算出（例えば年代×性別）した頻度集計表を**クロス集計**と呼ぶ

▼ある血液検査の結果

年代	性別	血清ナトリウム	...
70	男性	136	...
50	男性	136	...
60	男性	129	...
50	男性	137	...
60	女性	116	...
90	男性	132	...
70	男性	137	...
60	男性	131	...
60	女性	138	...
80	男性	133	...
70	男性	131	...
60	男性	140	...
40	男性	137	...
50	男性	137	...
40	女性	138	...
80	男性	136	...
80	男性	140	...
...	...	...	...

カテゴリーデータのため
平均値などの統計量
に意味を持たない

▼性別ごとの頻度集計

性別	人数
男性	194
女性	105
合計	299

複数データ項目を
掛け合わせる

▼年代ごとの頻度集計

年代	人数
40	47
50	82
60	93
70	52
80	19
90	6
合計	299

▼性別×年代のクロス集計表

年代	性別		人数 合計
	男性	女性	
40	27	20	47
50	56	26	82
60	58	35	93
70	34	18	52
80	15	4	19
90	4	2	6
合計	194	105	299

▼応用編：他項目の「平均値」を表示

年代	性別		血清ナトリウム 平均値
	男性	女性	
40	137.1	135.6	136.4
50	137.3	136.9	137.1
60	136.0	137.1	136.4
70	136.7	136.6	136.7
80	135.5	138.0	136.1
90	133.0	141.0	135.7
血清ナトリウム 平均値	136.5	136.8	136.6

Excelの
ピボットテーブル
のようなイメージ

(参考)「データの理解」でよく用いる関数

確認内容		関数	使用例
基本情報	データの先頭/末尾確認	データフレーム.head(行数) / tail(行数)	df.head(10) / df.tail(10)
	データ件数確認	データフレーム.shape	df.shape
	各列のデータ型などの確認	データフレーム.info()	df.info()
	各列の欠損値確認	データフレーム.isnull().sum()	df.isnull().sum()
統計値	要約統計量の算出	データフレーム.describe()	df.describe()
	相関行列	データフレーム.corr()	df.corr(numeric_only=True) #文字列の列がある場合"numeric_only=True"の指定が必要
集計値	クロス集計	pd.crosstab(行見出し, 列見出し)	pd.crosstab(df['役割'], df[['成約']契約まで])
グラフ	散布図行列	sns.pairplot(df, hue='色分け対象列名') ※seabornのインポート必要	sns.pairplot(df, hue='[成約]契約まで')
	ヒストグラム	データフレーム.hist()	df.hist()
	箱ひげ図	データフレーム.boxplot()	df.boxplot()
	棒グラフ (縦棒/横棒/積み上げ)	縦棒: データフレーム.plot.bar() 横棒: データフレーム.plot.barh()	df.plot.bar(), df.plot.barh() ※積上棒グラフは df.plot.bar(stacked=True)
	その他	過去資料を参考に	

データ加工の基本

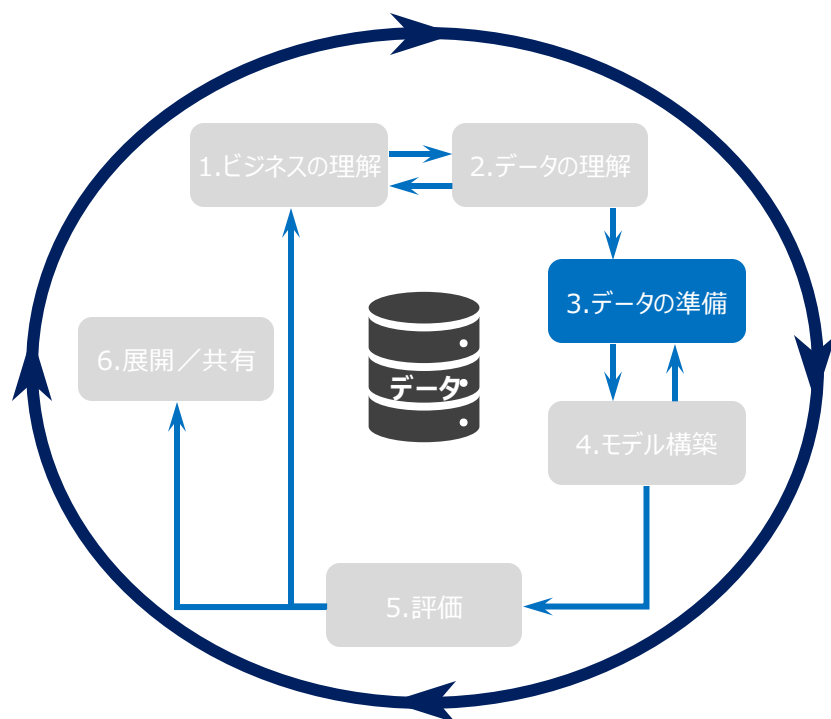
- ✓ 外れ値／異常値／欠損値の確認と対応
- ✓ カテゴリーデータとダミー変数化

データ分析の進め方

- データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(CRoss Industry Standard Process for Data Mining)



1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル構築

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

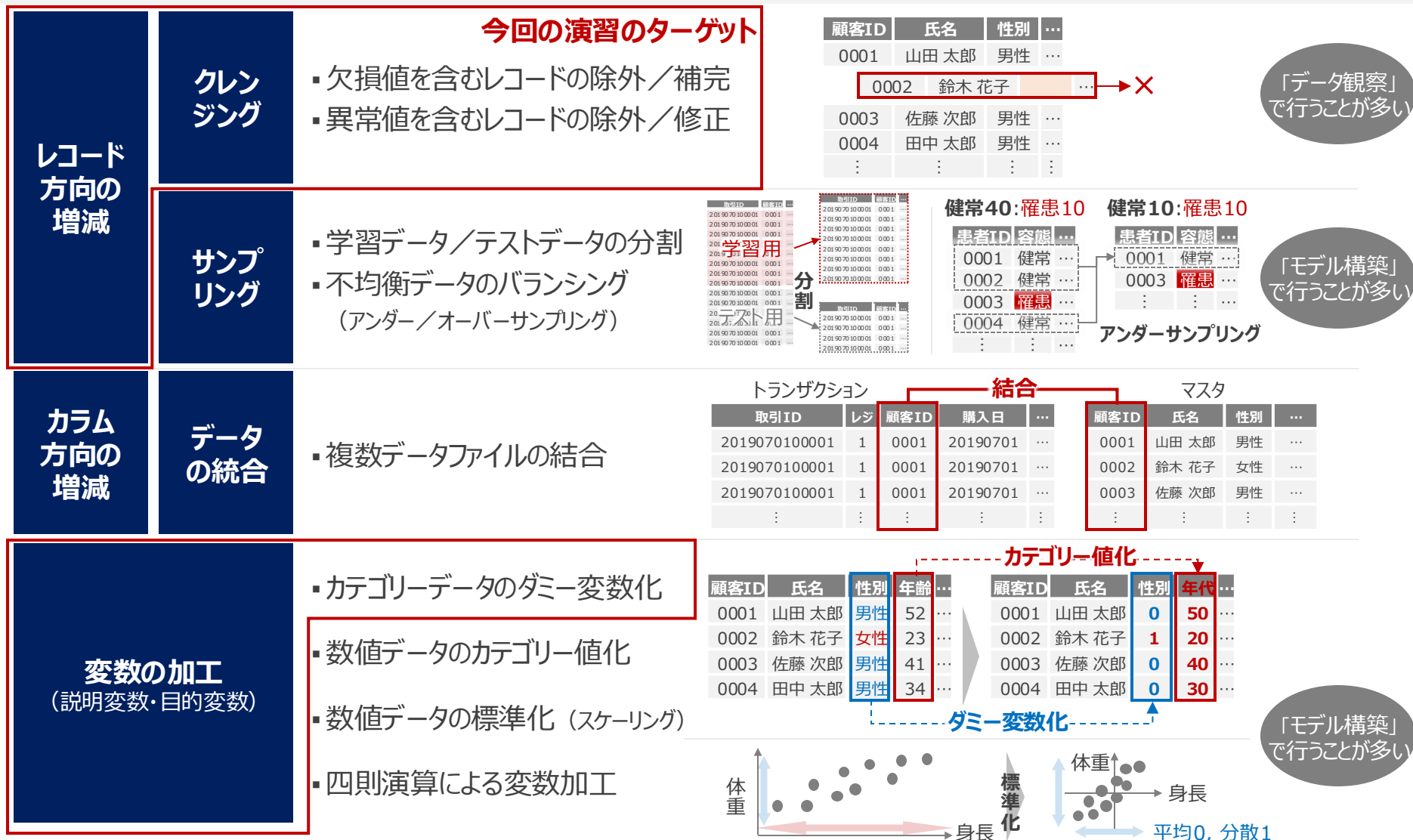
- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

「データの準備」で主に行うこと

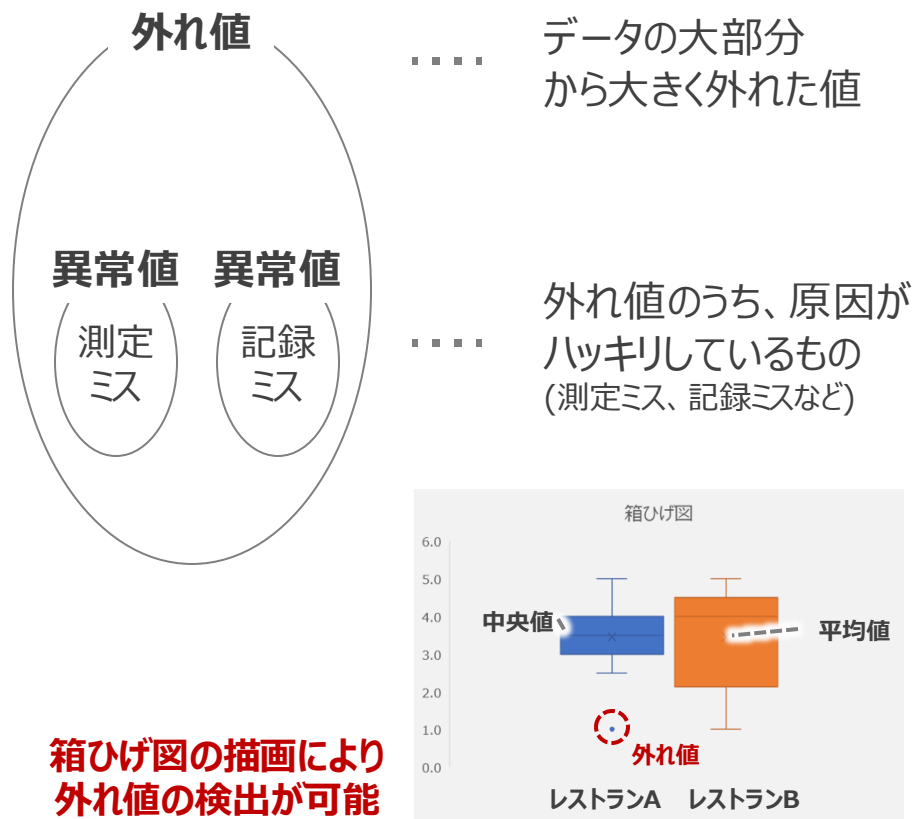
- 前段のデータの理解や後段のモデル構築と並行しながら、必要に応じてデータ加工を行う



外れ値／異常値／欠損値の確認と対応

- データに外れ値や異常値、欠損値が存在した場合、**分析結果に悪影響を与える可能性があるため、事前に存在を確かめ、対応方法を検討しておくことが重要**

外れ値と異常値



欠損値

欠損値

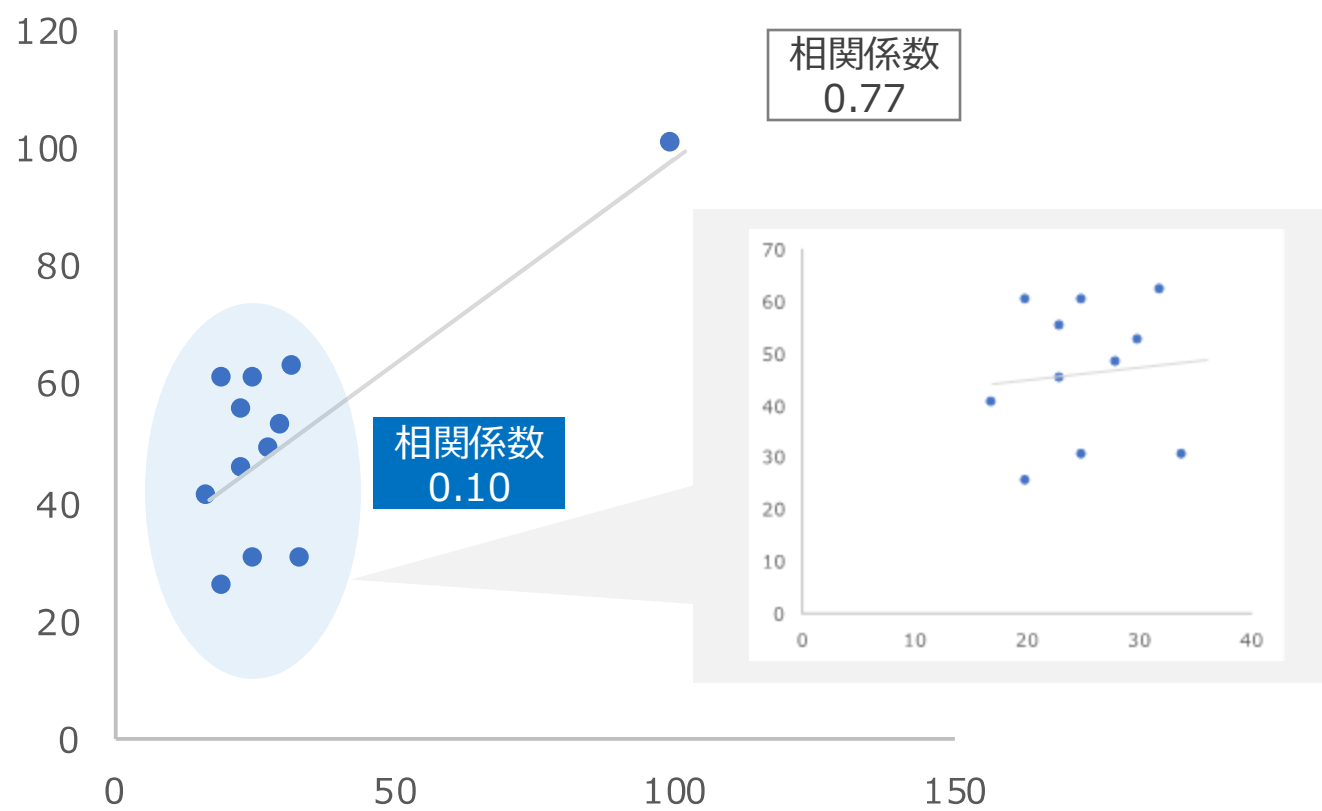
	A	B	C
	121	15	187
		21	175
	120	12	177
	115	25	169
	⋮	⋮	⋮

①行ごと除外
※あまりに欠損の多い列は列ごと除外

②補完
(平均値／中央値、予測値など)

外れ値の影響例

- 相関係数は外れ値の影響を大きく受けるため、数字だけに惑わされぬよう、散布図の確認も併せて行うことが重要である

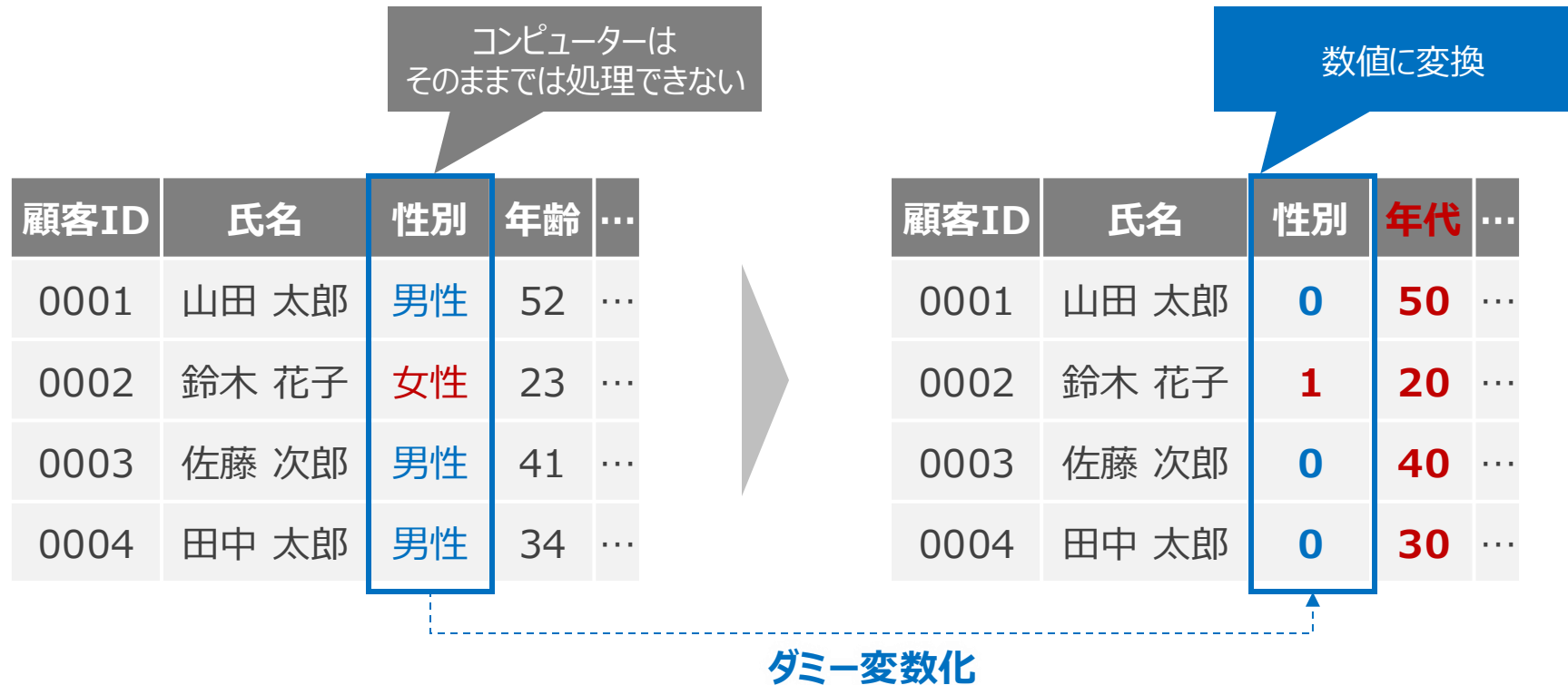




Google Colaboratory上での レクチャー&演習

カテゴリデータ（文字列データ） と ダミー変数化

- 通常の数値データに対して、血液型（A, B, O, AB）や部署（経営企画部、DX推進部、・・・）、性別など、分類上の区別に用いられる変数を「**カテゴリデータ**」と呼ぶ
- カテゴリデータは通常文字列で表されるため、擬似的に数値変換する「**ダミー変数化**」が必要



ダミー変数化の種類

- ダミー変数化のメジャーな考え方は2つあり、男性／女性、契約／未契約のような**2カテゴリのデータは連続値化**をし、職種や部署のような**カテゴリ間に連続性がない場合には横持ち化**を行う

① 連続値化 (Label/Ordinal Encoding)

2カテゴリ（男／女）やアンケートのような連続的な尺度に有効

元データ		ダミー変数化
▪ 男性	----->	0
▪ 女性	----->	1
▪ 不満である	----->	0
▪ どちらでもない	----->	1
▪ 満足している	----->	2

② 横持ち化 (One Hot Encoding)

カテゴリ間に直接的な連続性がない場合に有効

元データ		ダミー変数化
...	職種	...
...	会社員	...
...	自営業	...
...	会社員	...
...	会社員	...
...	会社員	...
...	会社員	...
...	医師	...
...	...	...

...	自営業	医師	会社員
...			1
...	1		
...			1
...			1
...			1
...		1	
...	⋮		

▪ 自営業	----->	0
▪ 会社員	----->	1
▪ 医師	----->	2

この場合、連続値化は不適



Google Colaboratory上での レクチャー&演習

データ分析（モデル構築） の基本

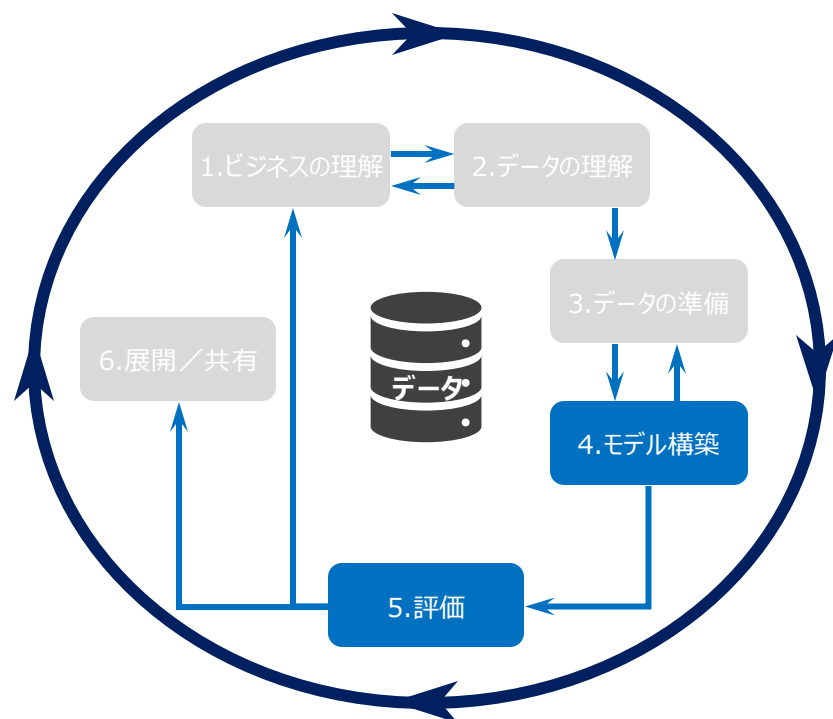
- ✓ 機械学習手法の基本
- ✓ 回帰分析による数値予測と影響因子の調査

データ分析の進め方

- データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(CRoss Industry Standard Process for Data Mining)



1. ビジネスの理解

- ビジネス、データマイニング目標の決定
- プロジェクトの立ち上げ

2. データの理解

- データの収集
- データの調査
- データ品質の検証

3. データの準備

- データの選択や除外
- データのクリーニング
- データの構築や統合

4. モデル構築

- モデリング手法の選択
- モデルの作成
- モデルの評価

5. 評価

- データマイニングの結果の評価
- プロセスの見直し
- 実行可能なアクションリストの作成

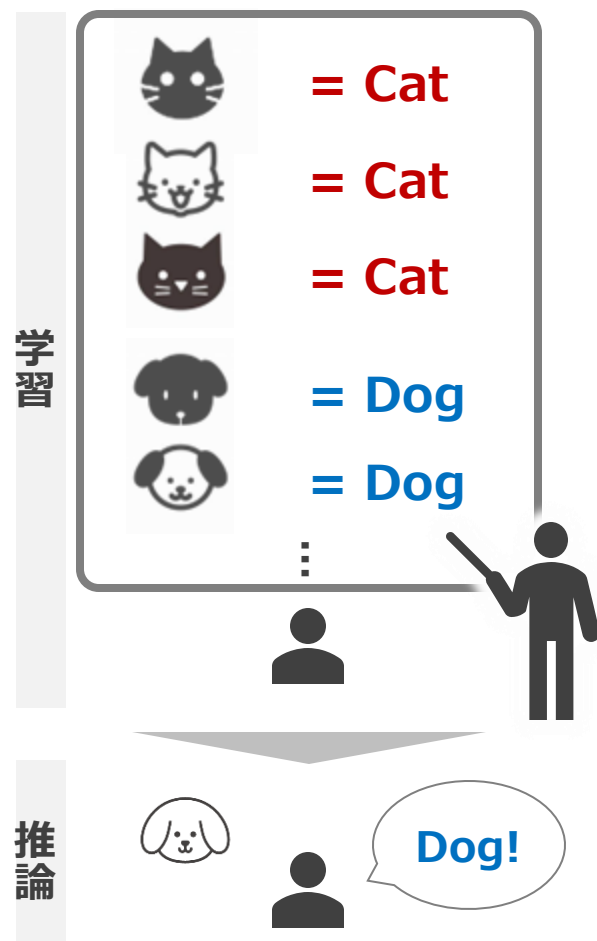
6. 展開／共有

- 業務への導入計画
- モニタリング、メンテナンスの計画

機械学習の分類

教師あり学習

あらかじめ「正解」を与えて
各データと正解の関係を学習させる



教師なし学習

「正解」を与えずに、各データ
のパターンから自分で学習する



※分類されたグループの意味づけは人が行う

教師あり学習のイメージ（数値予測とクラス分類）

- 各顧客レコードに対して数値もしくはカテゴリー値（クラス）の**解答**を与え関係性を学習

数値予測

（回帰）

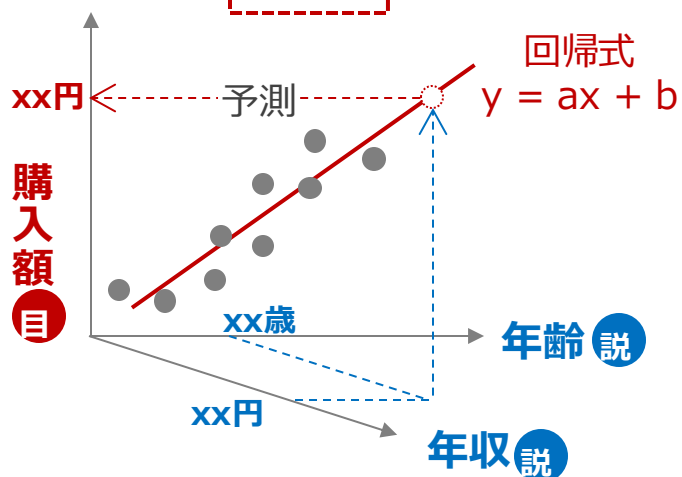
説明変数 目的変数

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...

データ例

イメージ

分析例



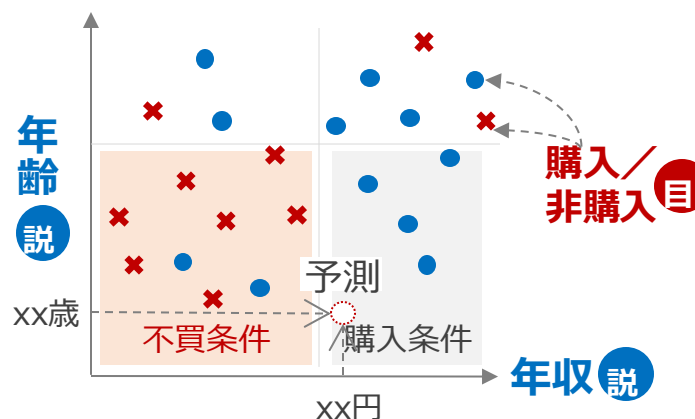
- 小売：売上の予測
- 小売：需要予測
- 製造：不良発生率の予測

クラス分類

（2クラス or 多クラス分類）

説明変数 目的変数

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...



- 小売：購入／非購入顧客の分類
- 医療：生存条件の分析
- 製造：故障種類の分類

教師なし学習のイメージ（クラスタリング）

- 各データ間の距離に基づき、近接データ（＝類似度が高いデータ） 同士のグループ（**クラスタ**）を作り、データを分類する手法。**学習データなし**でデータを大きく層別したい場合に有効

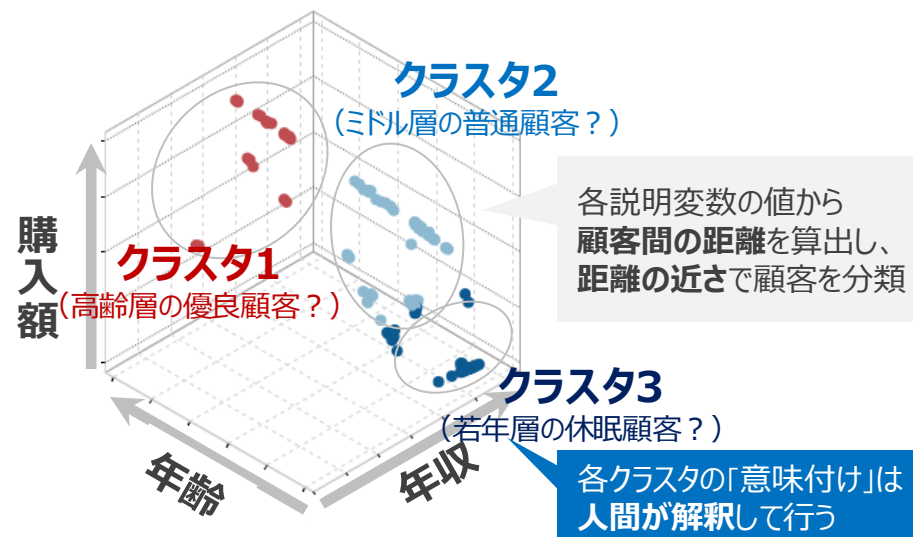
データ例

顧客ID	名前	年齢	年収	購入額	購入有無	...
0001	xx	25	300万	35,000	購入	...
0002	xx	35	600万	68,000	購入	...
0003	xx	18	120万	0	非購入	...
0004	xx	42	820万	85,000	購入	...
⋮	⋮	⋮	⋮	⋮	⋮	...

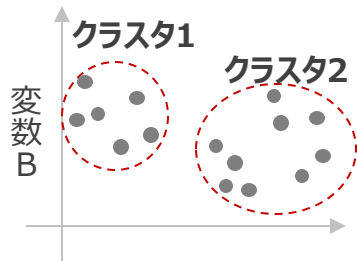
説明変数

※目的変数無し

クラスタリング



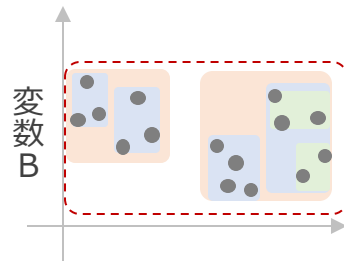
非階層的クラスタリング



主な手法

- **k-means法**
(k平均法)
- 混合ガウス

階層的クラスタリング



主な手法

- 最短距離法
- 最長距離法
- 群平均法
- ウォード法

(参考) 代表的な機械学習手法

- 目的やデータの特徴に合わせて手法を選ぶが、適宜、**手法の試行錯誤**が必要となることも多い

手法			概要	主なアルゴリズム
教師あり (答え合わせあり)	数値予測	数値予測	連続的な数値に対して、いくつかの影響因子を用いてある関数モデル (y=f(x)) を当てはめる	■ 線形回帰／非線形回帰 (2変数：単回帰、多変数：重回帰) ■ 回帰木 ■ ニューラルネットワーク
		時系列予測	継時的な変動に対して、過去の履歴やトレンドから将来の動きを予測	■ ARモデル (自己回帰モデル) ■ ARIMAモデル (自己回帰移動平均モデル)
	分類	クラス分類 (多クラス/二クラス)	あらかじめ分類のラベル付けがされたデータに対して、いくつかの影響因子を用いて、分類境界を抽出する	■ ロジスティック回帰 ■ 決定木分析 ■ 判別分析 ■ ニューラルネットワーク ■ サポートベクタマシン (SVM)
クラスタリング (グルーピング)		データに含まれる各変数の特徴から、類似データをグルーピングする	■ K-means ■ 階層的クラスタリング	
教師なし (答え合わせなし)		異常値検出	データの分布に基づき、分布から外れる異常な値を検出する	■ k-近傍法 (k-NN) ■ Local Outlier Factor (LOF)
	パターン抽出 (アソシエーションルール分析)		頻出するデータのパターン (データの組み合わせ) を抽出する (例：オムツとビールの併売)	■ アプリアリ ■ シーケンス
	データ要約 (次元削減)		高次元 (多変数) のデータに対して、より少ない変数で説明できる変数や軸を抽出する	■ 主成分分析 ■ 因子分析
強化学習 (報酬)	最適化		試行錯誤を通じてある目的達成のための最適な行動を学習 (例：自動運転、アルファ碁)	■ Q-Learning ■ SARSA (State-action-reward-state-action) ■ モンテカルロ法

(参考) 代表的な機械学習手法

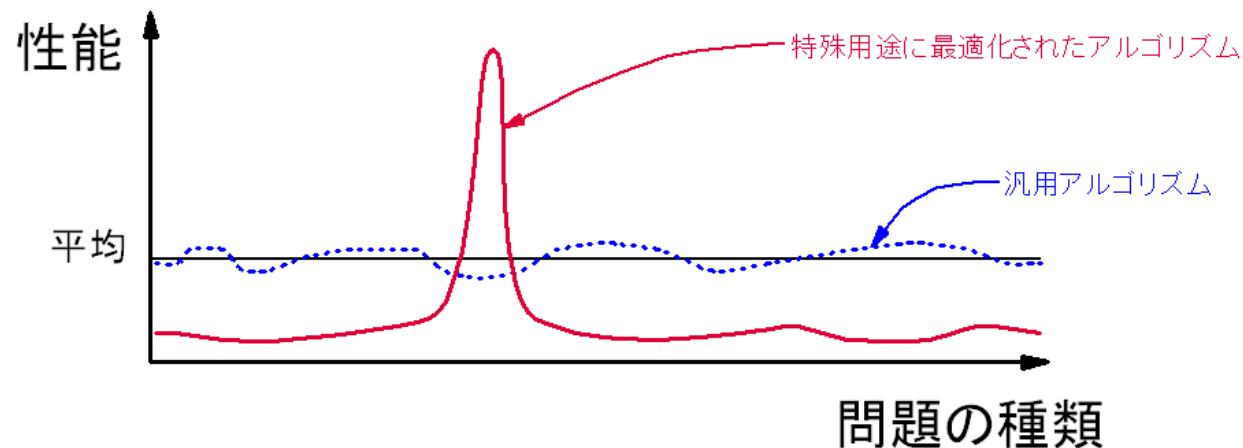
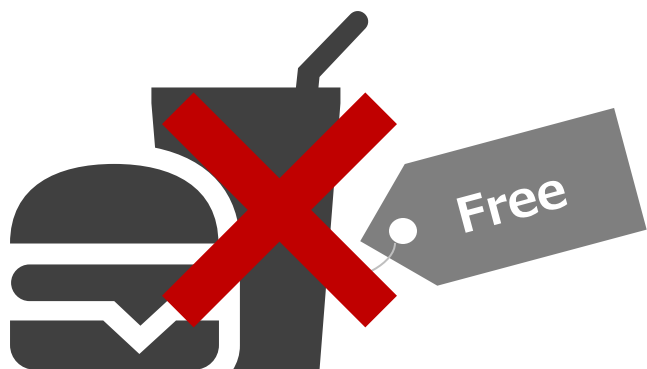
- 目的やデータの特徴に合わせて手法を選ぶが、適宜、**手法の試行錯誤**が必要となることも多い

手法			概要	主なアルゴリズム
教師あり (答え合わせあり)	数値予測	数値予測	連続的な数値に対して、いくつかの影響因子を用いてある関数モデル (y=f(x)) を当てはめる	■ 線形回帰 / 非線形回帰 (2変数：単回帰、多変数：重回帰) ■ 回帰木 ■ ニューラルネットワーク
		時系列予測	継続的な変動に対して、過去の履歴やトレンドを用いて予測する	■ ARモデル (自己回帰モデル) ■ ARIMAモデル (自己回帰移動平均モデル)
	分類	クラス分類 (多クラス/二クラス)	あらかじめ分類のラベル付けがされたデータに対して、いくつかの影響因子を用いて、分類境界を抽出する	■ ロジスティック回帰 ■ 決定木分析 ■ 判別分析 ■ ニューラルネットワーク ■ サポートベクタマシン (SVM)
		クラスタリング (グルーピング)	データに含まれる各変数の特徴から、類似データをグルーピングする	■ K-means ■ 階層的クラスタリング
教師なし (答え合わせなし)		異常値検出	データの分布に基づき、分布から外れる異常な値を検出する	■ k-近傍法 (k-NN) ■ Local Outlier Factor (LOF)
		パターン抽出 (アソシエーションルール分析)	頻出するデータのパターン (データの組み合わせ) を抽出する (例：オムツとビールの併売)	■ アプリオリ ■ シーケンス
		データ要約 (次元削減)	高次元 (多変数) のデータに対して、より少ない変数で説明できる変数や軸を抽出する	■ 主成分分析 ■ 因子分析
		最適化	試行錯誤を通じてある目的達成のための最適な行動を学習 (例：自動運転、アルファ碁)	■ Q-Learning ■ SARSA (State-action-reward-state-action) ■ モンテカルロ法

本セミナーで扱う手法

No Free Lunch定理：タダ飯は無い！

- あらゆる問題で高性能を示すモデルは理論上、構築不可能であることが示されている
- 問題設定や分析目的に応じて、最適なアルゴリズムの検討が必要である



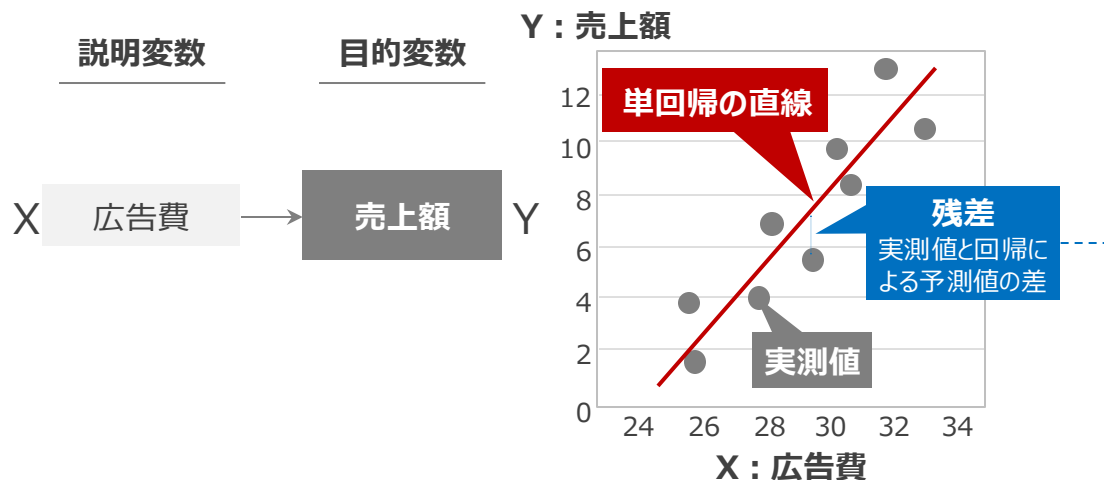
*画像出典：Wikipedia

No-free-lunch theorem: コスト関数の極値を探索するあらゆるアルゴリズムは、全ての可能なコスト関数に適用した結果を平均すると同じ性能となる (Wolpert and Macready, 1995年)

線形回帰分析による数値予測

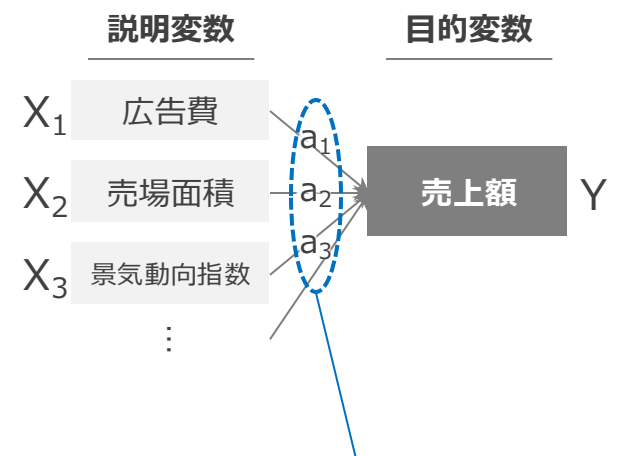
- 目的変数Yとその影響因子である説明変数Xとで関数 ($Y=f(X)$) に当てはめることを**回帰**という
- 特に $f(X)$ が線形関数の場合、**線形回帰**と呼び、1変数では**単回帰**、多変数では**重回帰**という

単回帰



残差の2乗和が最小になるように
係数a, 定数項bを決める (最小二乗法)

重回帰

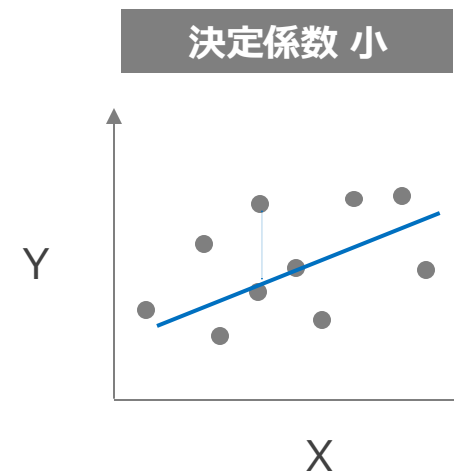
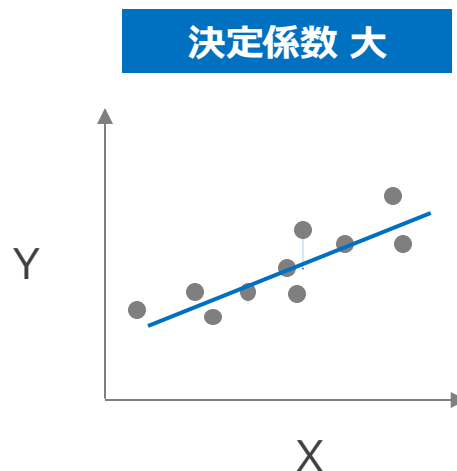


$$\text{売上額} = a \times \text{広告費} + b$$

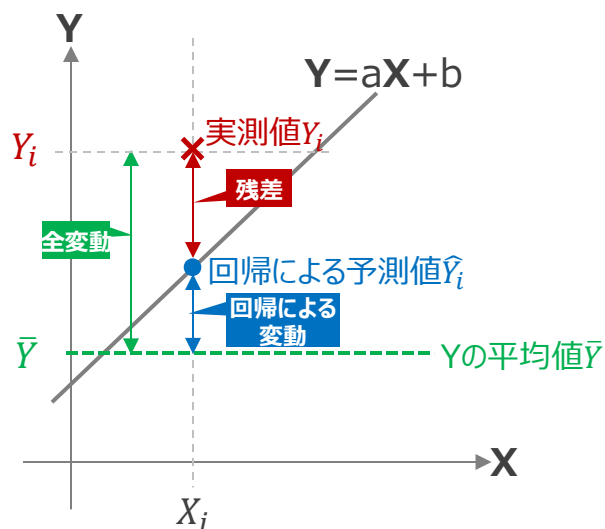
目的変数 (Y) 結果となる変数	回帰係数 (傾き) Xの影響の大きさ	説明変数 (X) 原因となる変数	定数項 (Y切片) 広告費=0のときの売上額
----------------------------	------------------------------	----------------------------	----------------------------------

決定係数によるモデルの当てはまり評価

- モデルの当てはまりの良さは、「決定係数」によって表すことが多い



▼参考：決定係数の定義イメージ



$$\text{決定係数 } R^2 = \frac{\text{回帰による変動 (平方和)}}{\text{全変動 (平方和)}}$$

(全変動のうち、回帰式で説明できる変動の割合)

- ・ 0~1の値をとる
- ・ (一般的な定義では) 相関係数の2乗に等しい

(参考) 相関係数の目安

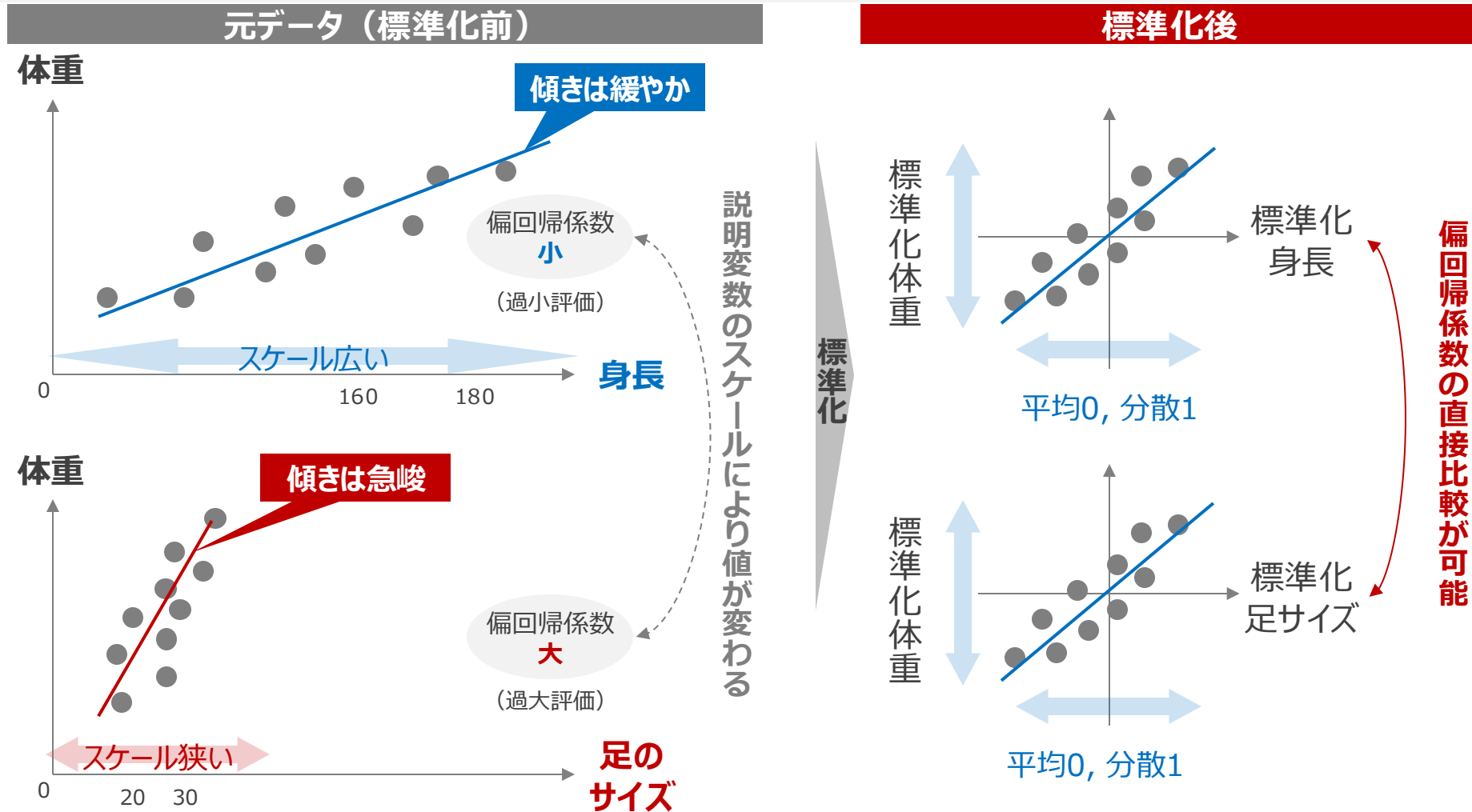
相関係数 (絶対値)	相関の強弱
1.0	強い相関
0.7	やや相関あり
0.4	弱い相関
0.2	ほぼ相関なし
0	



Google Colaboratory上での レクチャー&演習

回帰の注意点：データの標準化

- 重回帰分析の偏回帰係数は、目的変数に対する各説明変数の「傾き」を表しており、**その大きさは各変数のスケールに依存**するため、そのままでは直接比較が困難
- 予めデータを標準化しておくことで、偏回帰係数の直接比較が可能となる

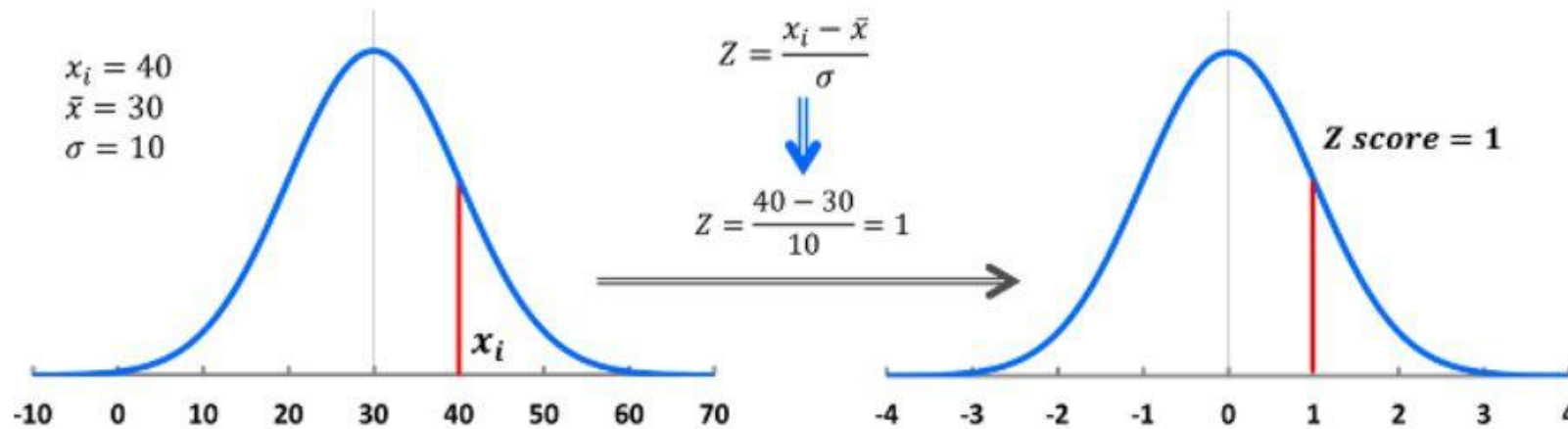


標準化の方法 (Zスコア)

- 重回帰分析の偏回帰係数は、各変数のスケールの大小に依存するため、直接比較が困難
- **Zスコア**を算出することでデータの標準化が行え、偏回帰係数の直接比較が可能となる

$$Z\ score = \frac{x_i - \mu}{\sigma}$$

x_i = 母集団を構成する要素iの値
 μ = 母平均
 σ = 標準偏差



出典 : <https://www.monodukuri.com/gihou/article/703>

(参考) 偏差値

偏差値は、「平均が50、標準偏差が10」になるように変換したもの : $50 + 10 \times Z\text{スコア}$



Google Colaboratory上での レクチャー&演習