総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第1回:Analysis(分析)とは

講師名:今津　義充

講座内容

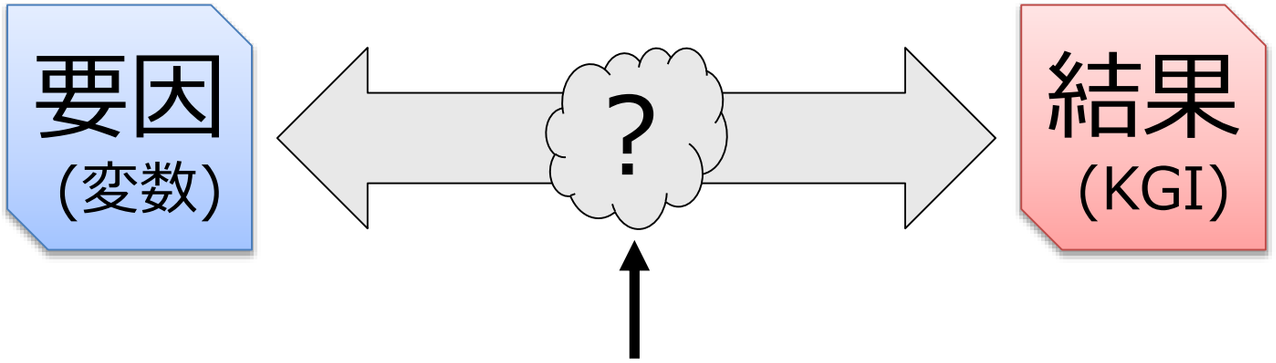
第1週	● データサイエンスとは
第2週	● 分析の概念と事例 ビジネス課題解決のためのデータ分析基礎(事例と手法)①
第3週	● 分析の具体的手法 ビジネス課題解決のためのデータ分析基礎(事例と手法)②
第4週	● ビジネスにおける予測と分析結果の報告 ビジネス課題解決のためのデータ分析基礎(事例と手法)③
第5週	● ビジネスでデータサイエンスを実現するために

第2週の内容紹介

第1回	● Analysis（分析）とは
第2回	● 1変数の状況の把握①（可視化の活用）
第3回	● 1変数の状況の把握②（代表値の活用）
第4回	● 比較して2変数の関係を見る
第5回	● ビジネスにおける比較①（概要）
第6回	● ビジネスにおける比較②（適切なA/Bテストの活用）

Analysis（分析）とは

- 分析＝複雑な事柄を要因に分け、その構造・関係を解明
- 仮説に基づいて、各要因と結果(KGI)の関係を調査する



分析の第一歩は、1変数による状況把握と
要因と結果を2変数の関係として解明すること

変数の尺度

- **分析手法を理解する前提として必ずおさえたい知識**

名義尺度（質的）	カテゴリに分類するための特性を表す尺度	順序無し	• 順序に意味がない 例：性別、都道府県、血液型など
		順序付き	• 順序に意味がある 例：満足度、順位など
連続尺度（量的）	数値で表し測れる大小の関係がある尺度	間隔尺度	• 順序及び和差の演算が意味がある 例：年齢、セ氏度など
		比率尺度	• 順序及び和差積商の演算が意味がある 例：体重、金額、速度など

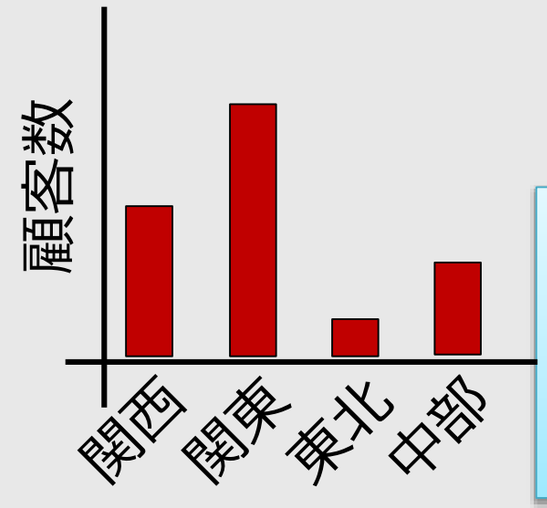
変数の尺度により分析手法を変える必要がある

1変数の状況を把握 (データチェック)

- 分析の第一歩としては、可視化と代表値により、
  各要因 (1変数) の状況を把握

棒グラフ

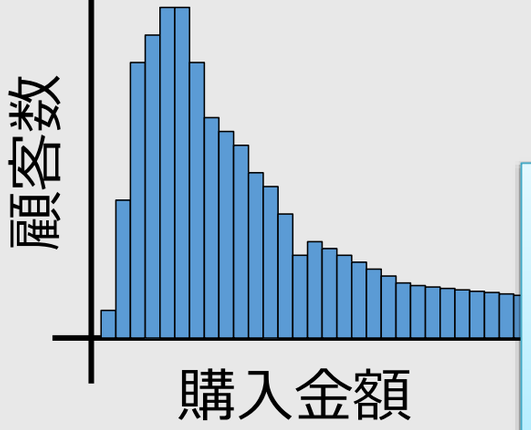
可視化



代表値
- カテゴリ数
- 最頻値
- など

ヒストグラム

可視化



代表値
- 平均値
- 標準偏差
- 最頻値
- など

1変数の可視化と代表値の算出は鳥瞰的な状況把握と分析の次のステップを計画するのに重要

２変数の関係を調査

- KGIと要因の関係を調査するために、尺度によって様々な手法がある

比較

名義 vs 名義：クロス集計を用いて、離散分布を比較する

名義 vs 連続：ヒストグラムを用いて、連続分布を比較する

傾向

連続 vs 連続：散布図を用いて、片方の変数に対して
もう片方の変数の傾向を見る (片方は時間だと、時系列と呼ぶ)

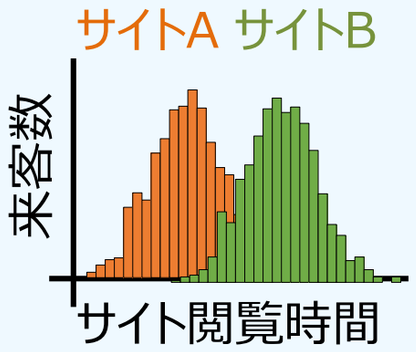
比較

名義 vs 名義

来客数	男	女
サイトA	18	3
サイトB	4	16

クロス集計

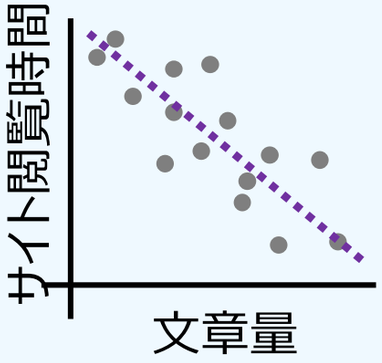
名義 vs 連続



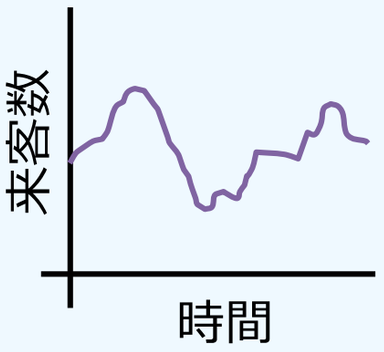
ヒストグラム

傾向

連続 vs 連続



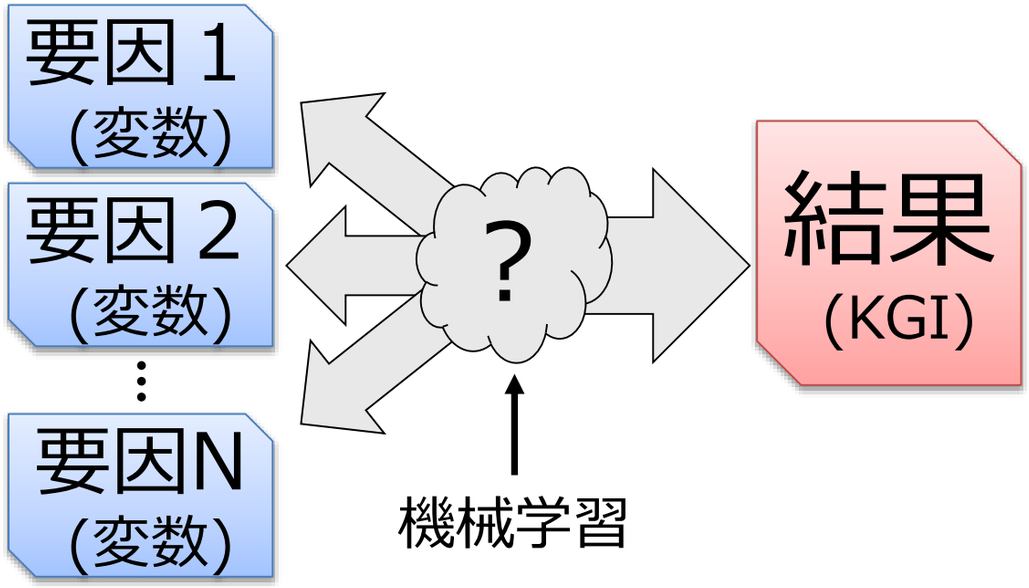
散布図



時系列

複数変数の関係を調査したい場合は？

- 要因が複数の時、要因間の相互作用も考慮すべきであるが、
  変数が３～４個以上になると、前述の手法だけでは困難
- 要因と結果を示すデータをコンピューターに与え、自動的にその
  関係を学習させる機械学習などが有効となる



機械学習は、第４週で紹介

次回のテーマ

次回は

「1変数の状況の把握①（可視化の活用）」

お疲れ様でした！

総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第2回:1変数の状況の把握①（可視化の活用）

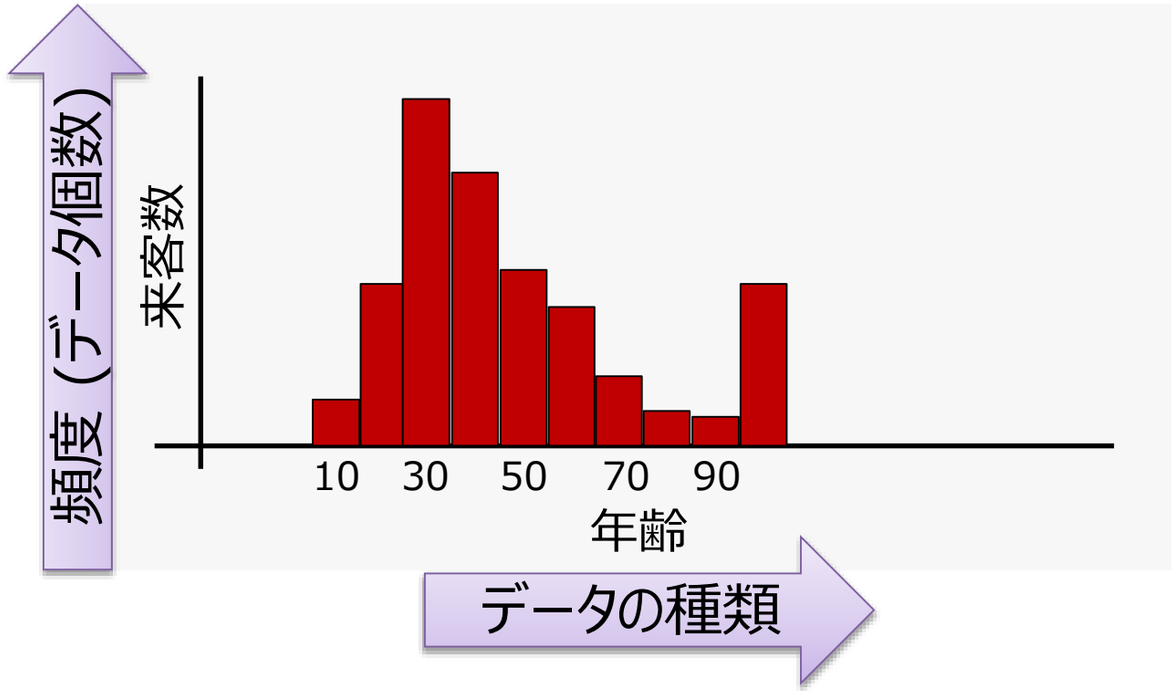
講師名:今津　義充

第2週の内容紹介

第1回	● Analysis (分析) とは
第2回	● 1変数の状況の把握① (可視化の活用)
第3回	● 1変数の状況の把握② (代表値の活用)
第4回	● 比較して2変数の関係を見る
第5回	● ビジネスにおける比較①(概要)
第6回	● ビジネスにおける比較②(適切なA/Bテストの活用)

可視化の重要性

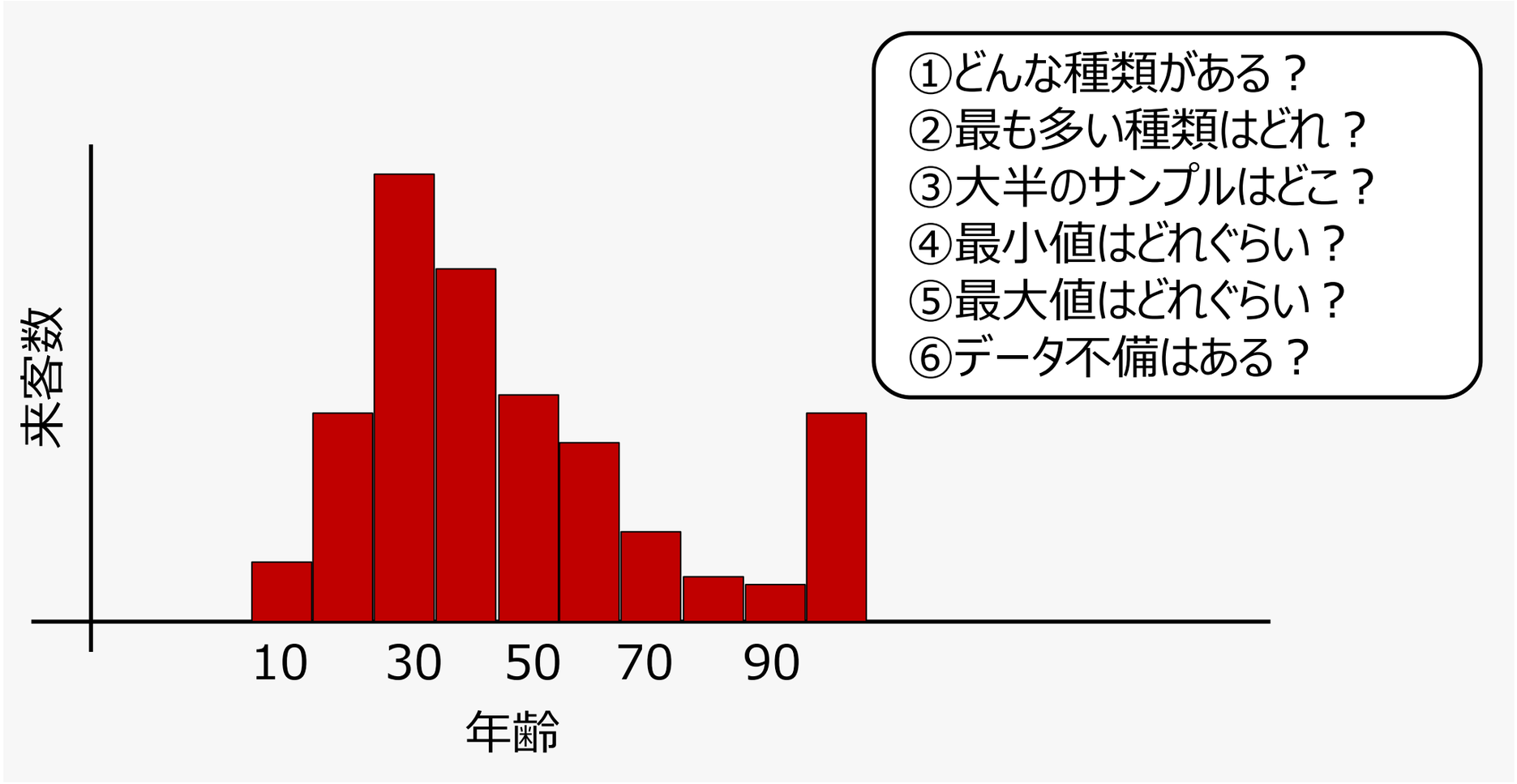
- 可視化では様々な情報を一目で把握できる
- 1変数の状況把握のために、ヒストグラムを用いる



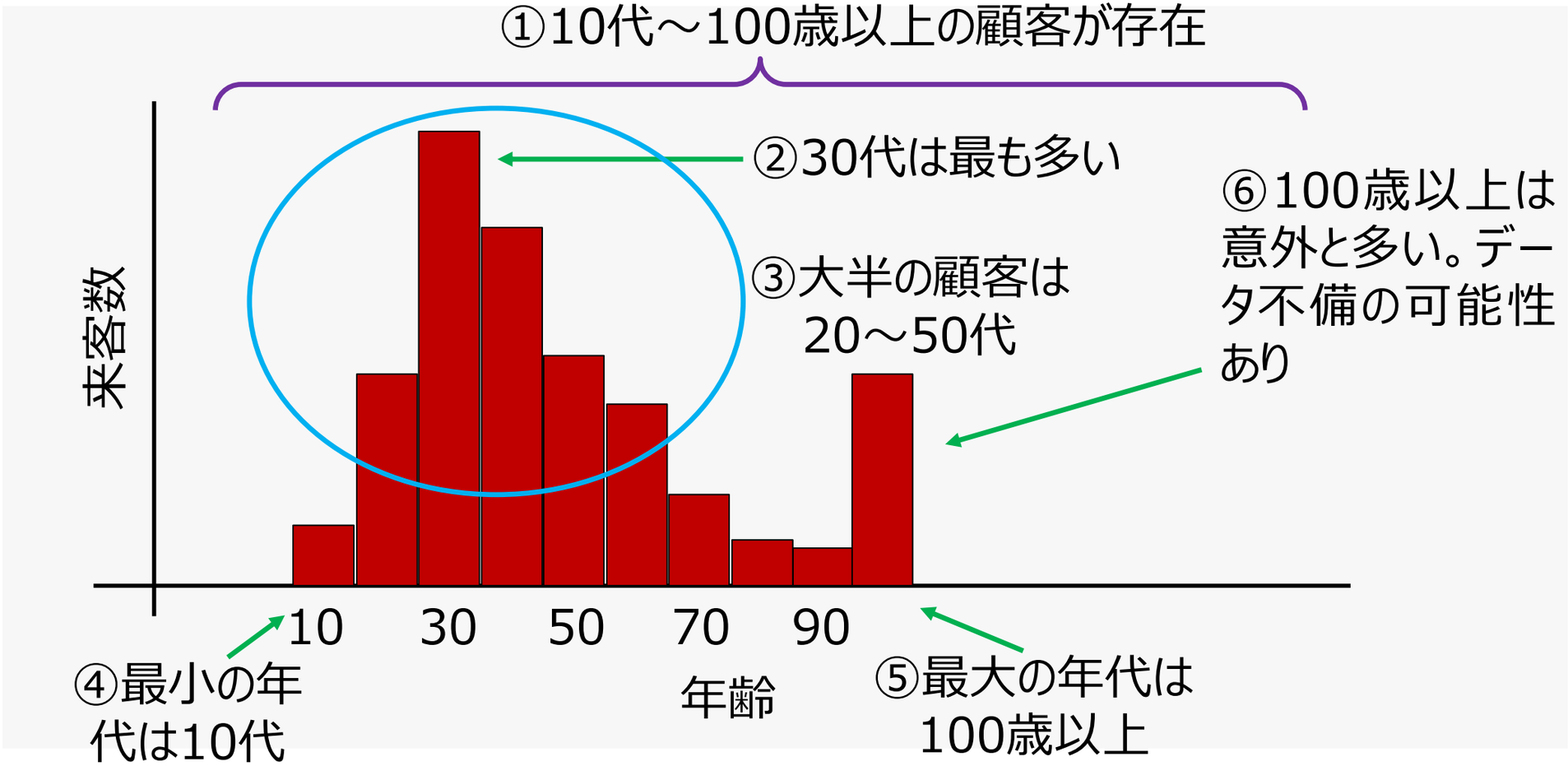
一枚の絵は一千語に匹敵する

分布の見方①

- 下図は、ある店の年代別来客数のヒストグラムです。
  グラフから何が読み取れるでしょうか



分布の見方②

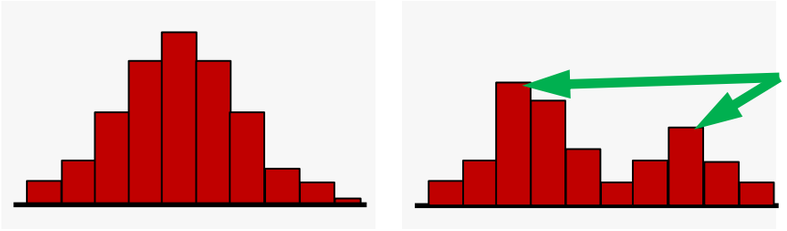


可視化することで様々な情報を一目で把握できる

分布の見方③

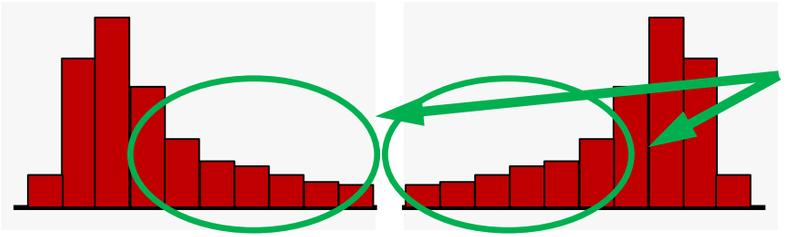
- 変数の性質によって特徴の異なる様々な分布がある

ピーク（峰）の数



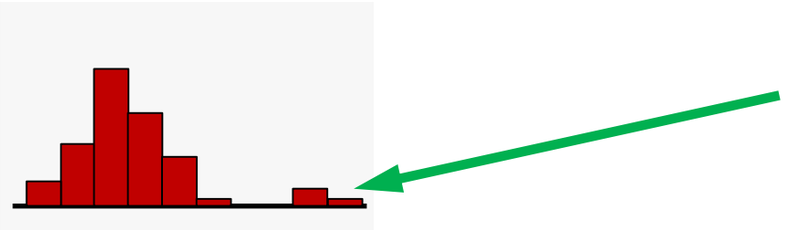
異種データの
混在の可能性

ピーク（峰）の
偏り



平均値を見る
際に注意

外れ値の有無



データ不備や
異常値の可能性

変数の性質を把握するのに分布特徴に注意すべき

次回のテーマ

次回は

「1変数の状況の把握②（代表値の活用）」

お疲れ様でした！

総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第3回:1変数の状況の把握②（代表値の活用）

講師名:今津　義充

第2週の内容紹介

第1回	● Analysis (分析) とは
第2回	● 1変数の状況の把握① (可視化の活用)
第3回	● 1変数の状況の把握② (代表値の活用)
第4回	● 比較して2変数の関係を見る
第5回	● ビジネスにおける比較①(概要)
第6回	● ビジネスにおける比較②(適切なA/Bテストの活用)

代表値の重要性

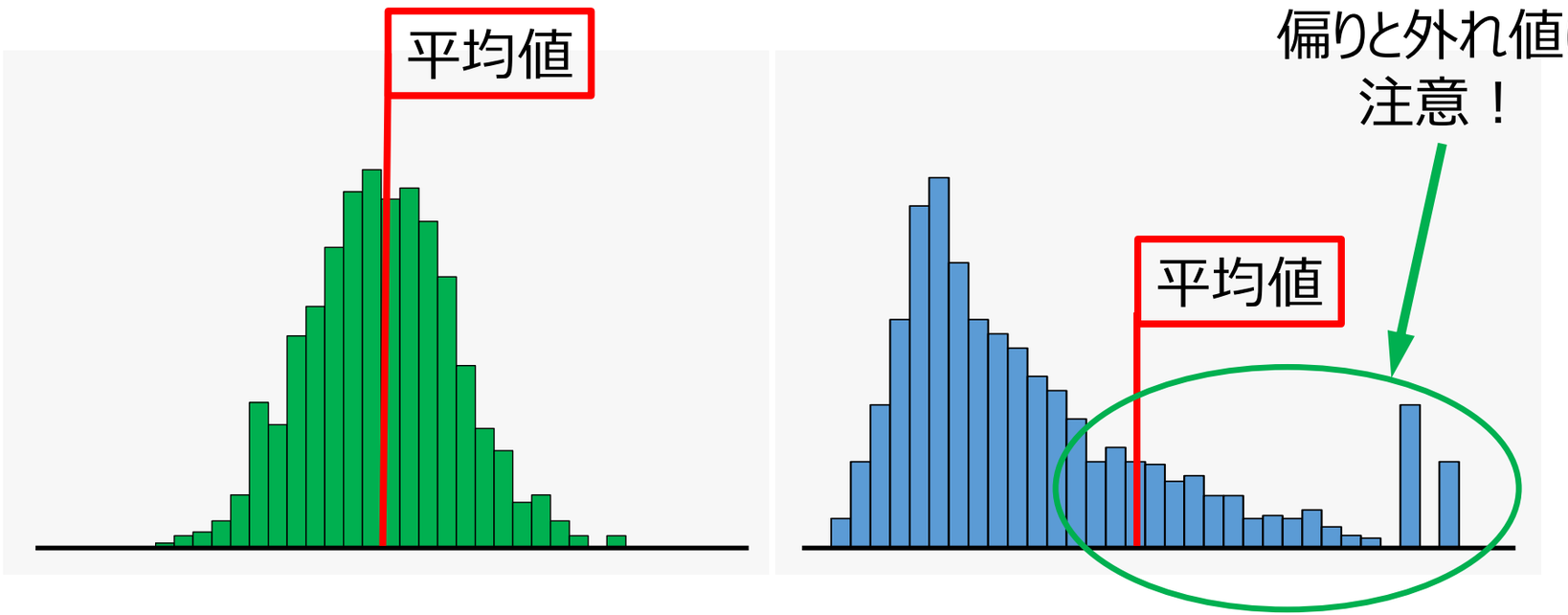
- 代表値（統計量）は分布の特徴を数値にまとめるもの
- 代表値では分布を見なくても、分布の特徴を把握できる
- 一般的には、以下の代表値がよく用いられる

位置を示す代表値	• 平均値 • 中央値 • 最頻値
ばらつきを示す代表値	• 標準偏差（分散）
分布の形を示す代表値	• 尖度 • 歪度

代表値では分布の特徴を少ない情報で伝えられる

位置を示す代表値①

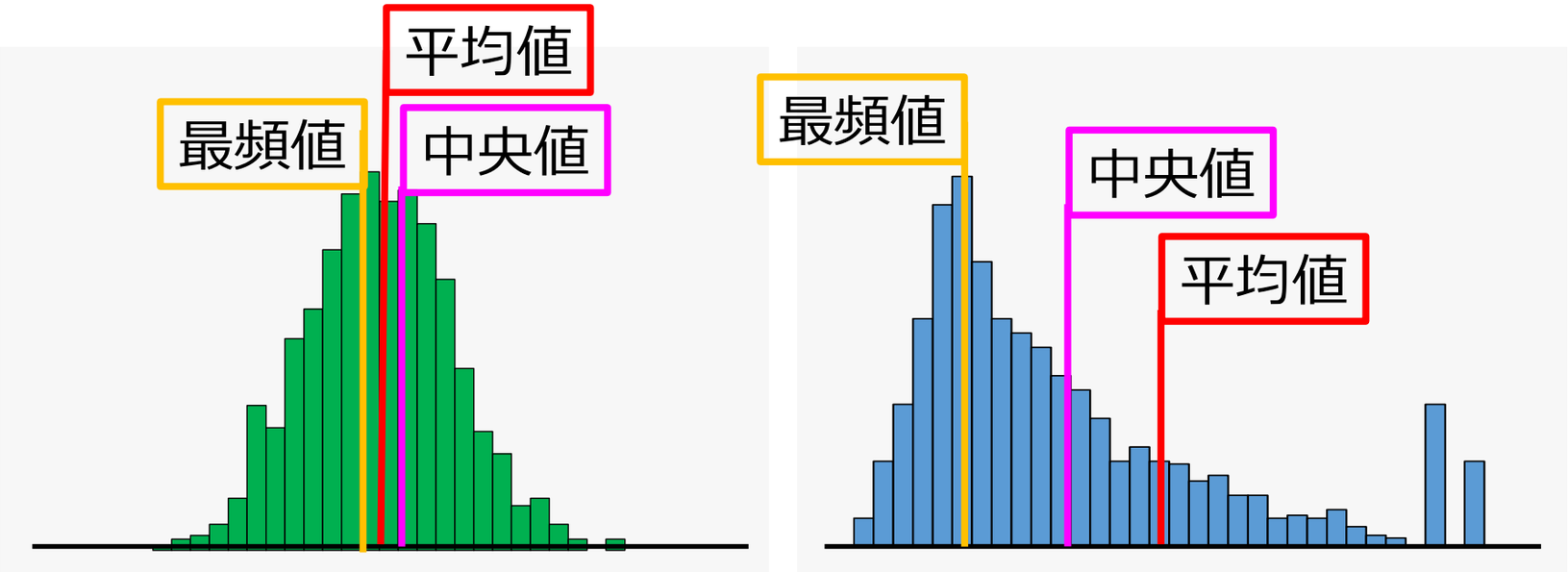
- 平均値：分布の中心傾向を表す値
- 但し、分布が偏っている場合や、外れ値が存在する場合には
  平均値を解釈する際に注意



平均値では分布の中心を推定できる

位置を示す代表値②

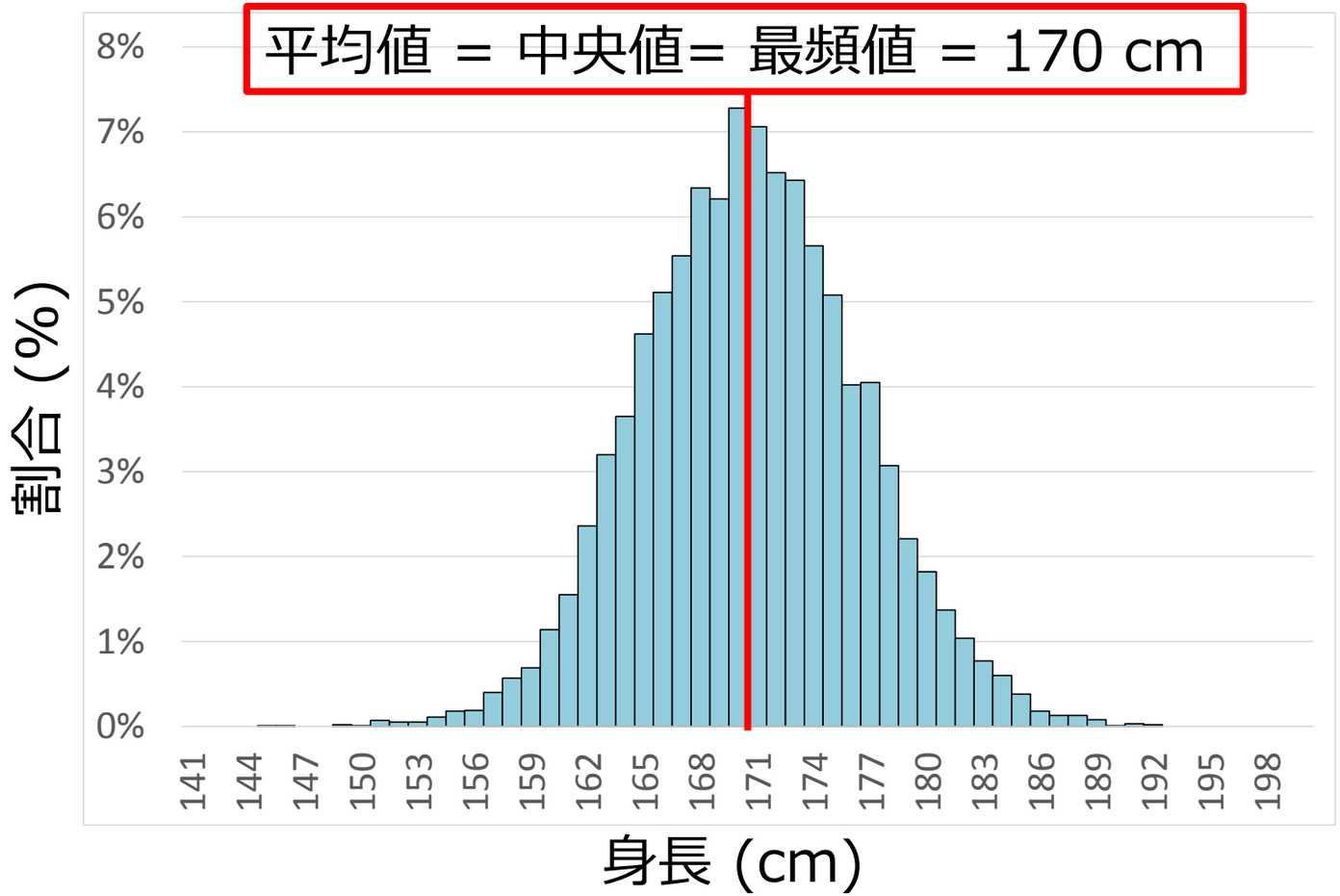
- 中央値：分布を下半分と上半分に分ける値
- 最頻値：頻度が最も高い値



偏りや外れ値がある場合、
中央値と最頻値は平均値より有意義であることがある

位置を示す代表値の例①

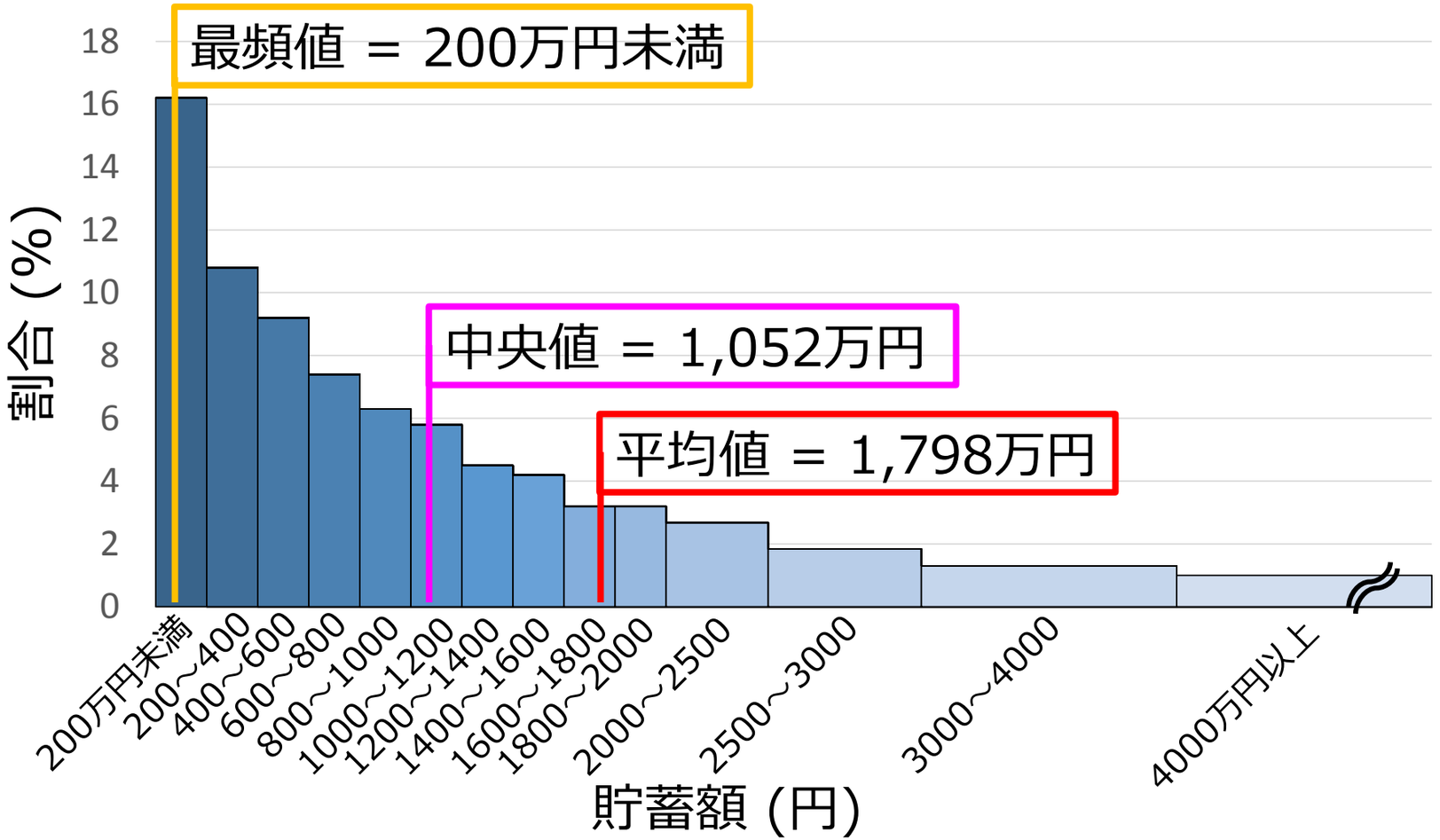
17歳の男子の身長分布 (平成26年度)



出典：平成26年度 学校保健統計調査結果(文部科学省)
http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001058732&cycode=0

位置を示す代表値の例②

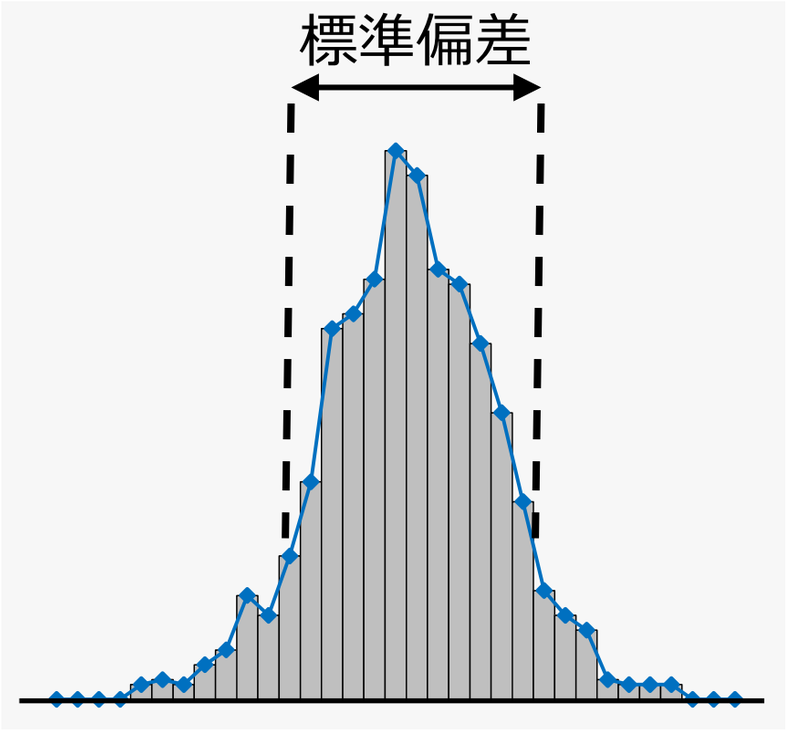
貯蓄現在高階級別世帯分布 (二人以上の世帯) (平成26年)

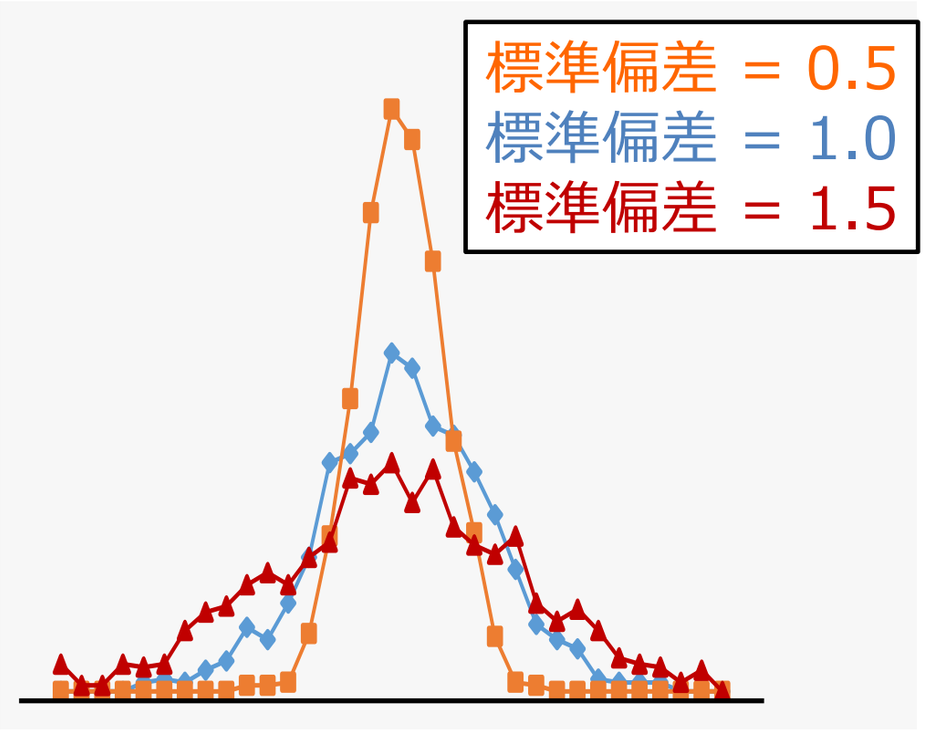


出典：家計調査結果(総務省)
http://www.stat.go.jp/data/kakei/family/05.htm

ばらつきを示す代表値

- 標準偏差：分布が平均値からの散らばりを示す値

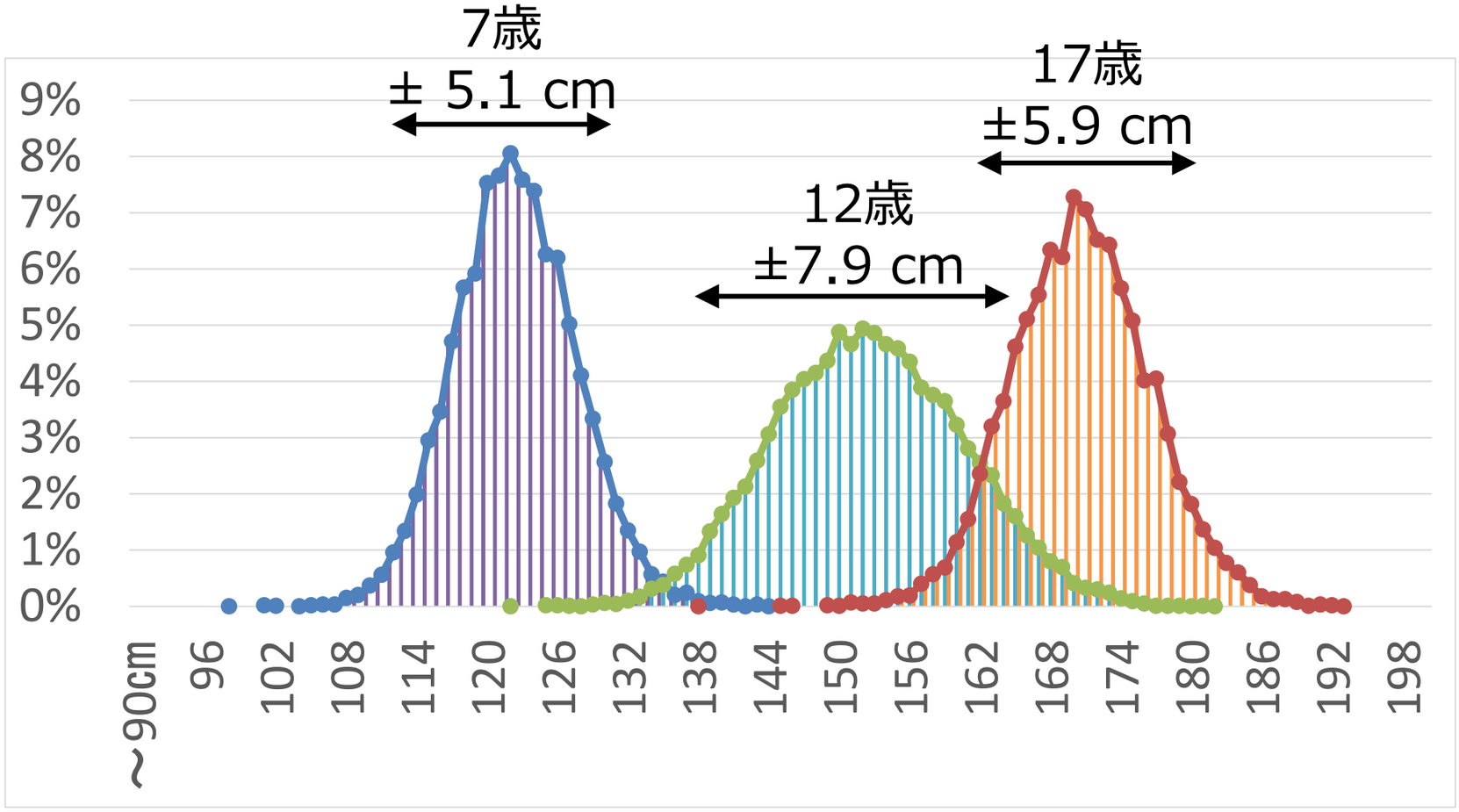




分布のばらつきが広いほど、標準偏差が高い

ばらつきを示す代表値の例

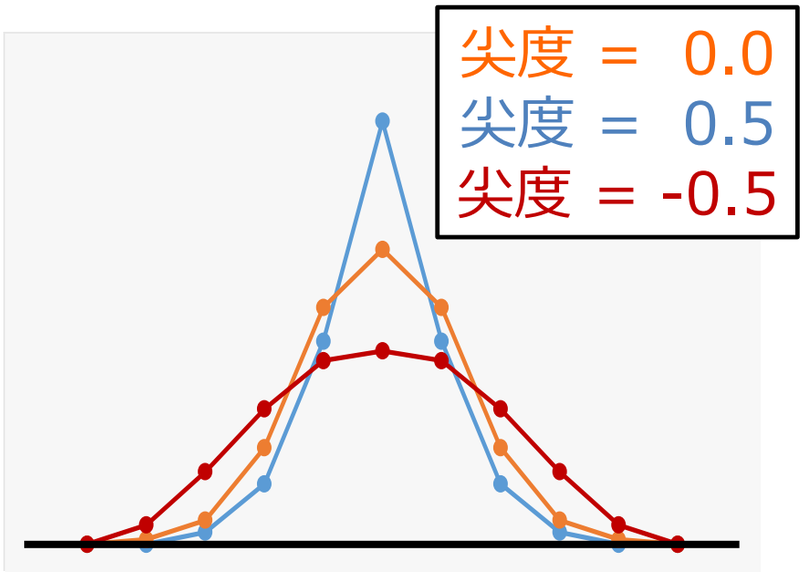
男子の身長分布 (平成26年度)

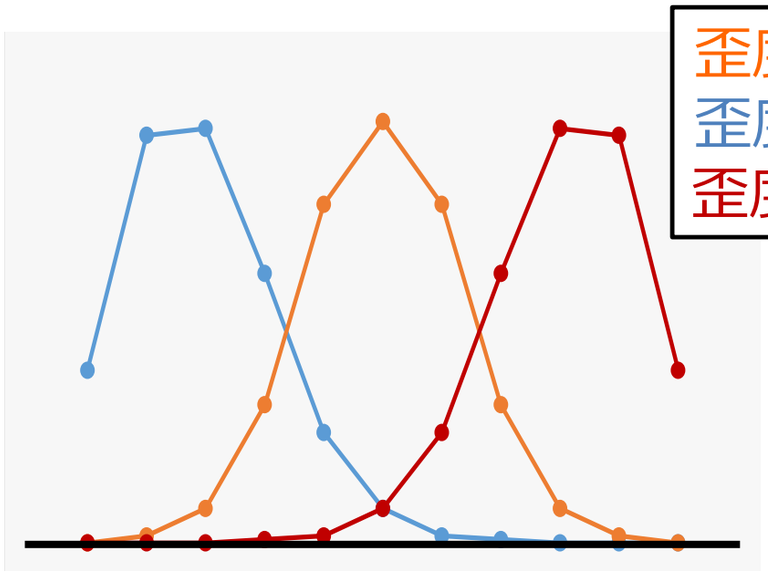


出典：平成26年度 学校保健統計調査結果(文部科学省)
http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001058732&cycode=0

分布の形を示す代表値

- 尖度：ピーク（峰）への集中度合いを示す値
- 歪度：左右へのピーク（峰）の偏りを示す値





次回のテーマ

次回は

「比較して２変数の関係を見る」

お疲れ様でした！

総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第4回:比較して2変数の関係を見る

講師名:今津　義充

第2週の内容紹介

第1回	● Analysis (分析) とは
第2回	● 1変数の状況の把握① (可視化の活用)
第3回	● 1変数の状況の把握② (代表値の活用)
第4回	● 比較して2変数の関係を見る
第5回	● ビジネスにおける比較①(概要)
第6回	● ビジネスにおける比較②(適切なA/Bテストの活用)

比較とは

- 比較する変数の尺度により手法を変える必要がある

名義 vs 名義：クロス集計を用いて、離散分布を比較する

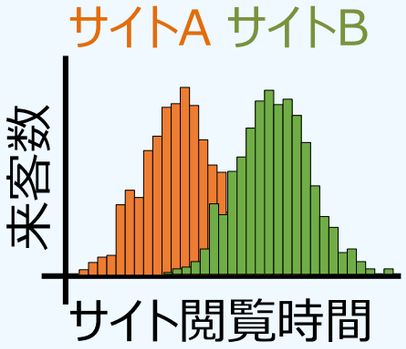
名義 vs 連続：ヒストグラムを用いて、連続分布を比較する

名義 vs **名義**

来客数	男	女
サイトA	18	3
サイトB	4	16

クロス集計を用いて
離散分布を比較する

名義 vs **連続**



ヒストグラムを用いて
連続分布を比較する

名義変数 vs 名義変数：クロス集計

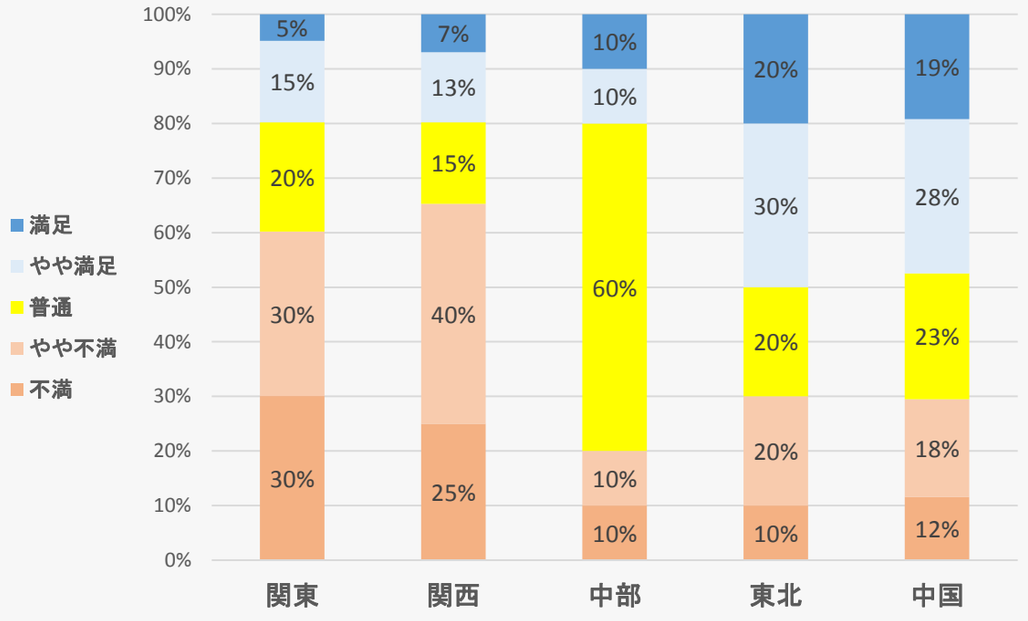
- ２変数のカテゴリの組み合わせでデータの個数を集計
- 横カテゴリにより縦カテゴリの構成が変化するかを調査する

あるネット銀行の地域別顧客満足度の構成比

KGI：顧客満足度 (5カテゴリ)
要因：地域 (5カテゴリ)

地域別顧客満足度 (万人)

	関東	関西	中部	東北	中国
満足	17	20	20	24	15
やや満足	52	37	20	36	22
普通	70	43	120	24	18
やや不満	105	116	20	24	14
不満	105	72	20	12	9

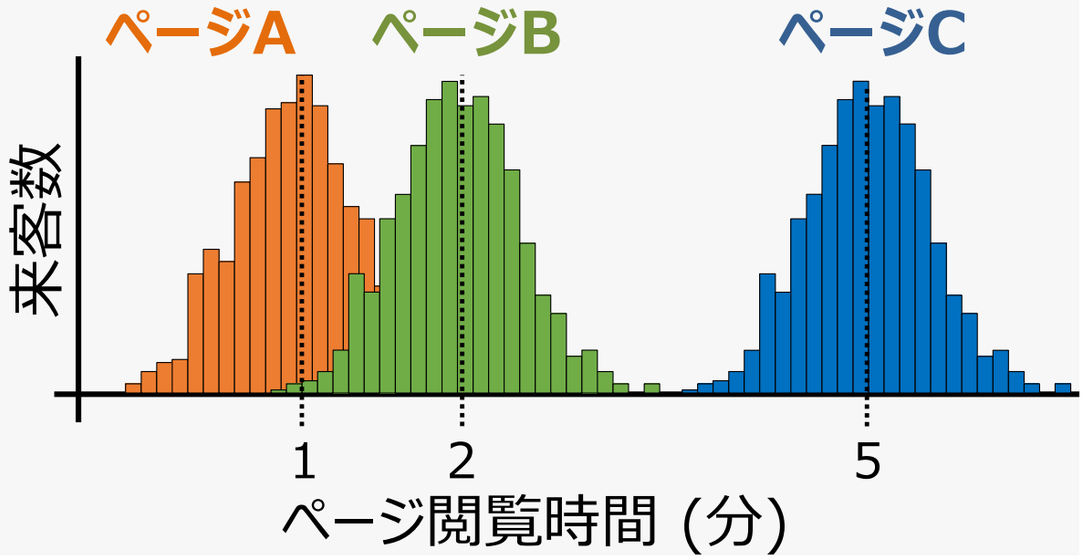


クロス集計で一目で比率の違いを把握できる

連続変数 vs 名義変数：ヒストグラムの比較

- 「平均値や分布の形はカテゴリによって違うか」を調査するために、ヒストグラムの比較を行う

KGI：ページ閲覧時間
要因：ページ名 (3カテゴリ)



ヒストグラムの比較でカテゴリによって連続変数の分布が変わるかを一目で把握できる

次回のテーマ

次回は

「ビジネスにおける比較①(概要)」

お疲れ様でした！

総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第5回:ビジネスにおける比較① (概要)

講師名:渋谷 直正

第2週の内容紹介

第1回	● Analysis (分析) とは
第2回	● 1変数の状況の把握① (可視化の活用)
第3回	● 1変数の状況の把握② (代表値の活用)
第4回	● 比較して2変数の関係を見る
第5回	● **ビジネスにおける比較①(概要)**
第6回	● ビジネスにおける比較②(適切なA/Bテストの活用)

ビジネスにおける比較の事例

- **ビジネスにおいて、「比較」は施策の効果検証のためによく用いられる**

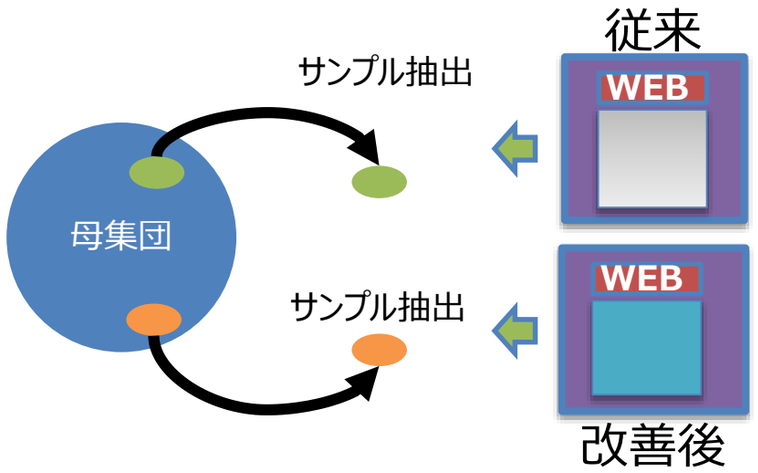
  - 広告デザインの売上への効果
  - ウェブサイト・コンテンツのクリック率への効果
  - ワクチンの感染病予防率への効果

  など

比較による効果検証のために
A/Bテストを行うことが多い

A/Bテストの事例

あるウェブサイトは**会員登録ボタンのクリック率**を向上させたい。そのために、ウェブページのデザインを改善した



A/Bテストの実施

1. １ヶ月間の来客を2群に分けた
2. 2デザインをそれぞれの群に出した
3. 各群におけるクリック率を記録した
4. 2分布を比較した結果、改善デザインによりクリック率が上がったと分かった

	クリックあり	クリックなし	計	クリック率
従来	100	9,900	10,000	1.0%
改善	150	9,850	10,000	1.5%

- **要因：デザイン（従来、改善）**
- **KGI：クリック率**

A/Bテストの紹介

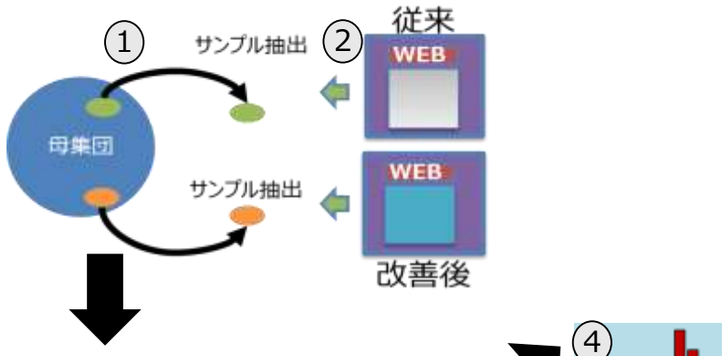
- **A/BテストはKGIと施策の間の関係（施策効果）を調査する手法。以下の流れにより行う**

① 対象の集団から小集団を2つ取り出す。小集団は「標本」と呼ぶ

② 効果検証をしたい施策Aと施策Bをそれぞれの標本に適用する

③ それぞれの標本においてKGIを測る

④ 両施策によるKGIの分布を比較し、有意な効果があるかを判断する



	クリックあり	クリックなし	計	クリック率
従来	100	9,900	10,000	1.0%
改善後	150	9,850	10,000	1.5%

要因：施策A又は施策Bのカテゴリをとる

名義変数

KGI：施策の効果を受ける値

（連続変数又は名義変数）

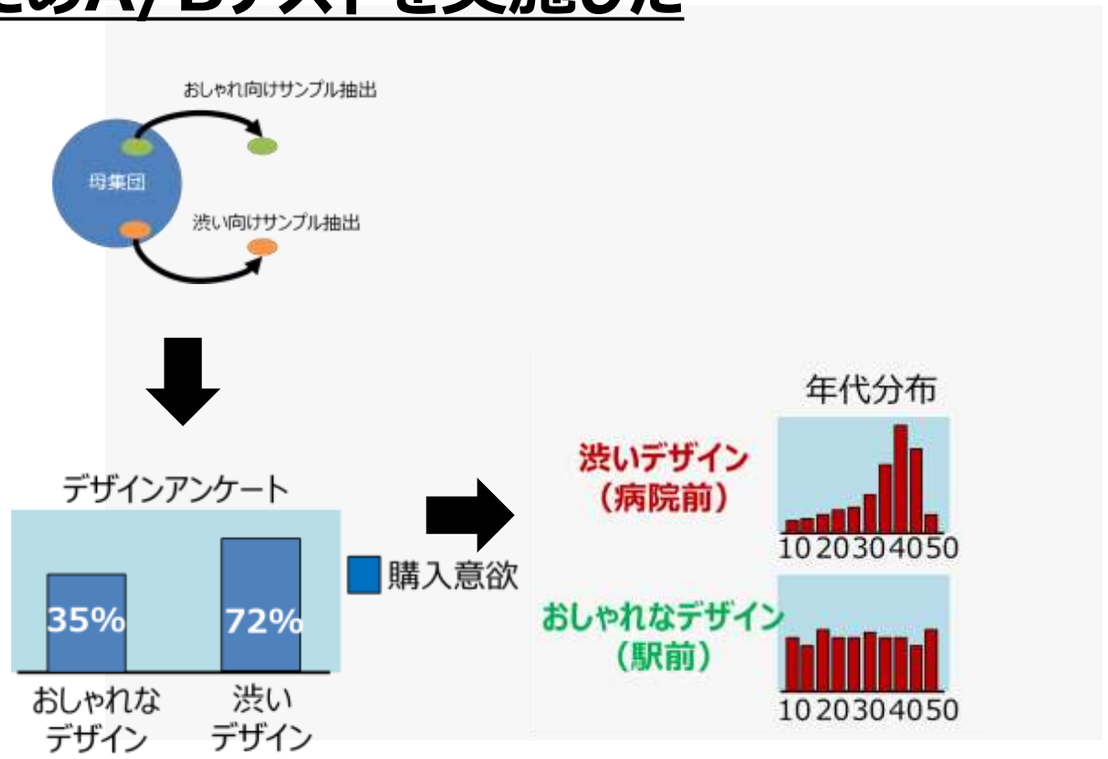
不適切なA/Bテストの事例

- 比較が公平であるようにテストを適切に設計すべき

(事例)ある広告会社は“渋いデザイン”と“おしゃれなデザイン”2デザインの売上効果を図るためA/Bテストを実施した

標本Aに渋いデザインを、標本Bにオシャレなデザインを設定しアンケートをとった結果、渋いデザインが最も売上を増やすと見られた

ただし、渋いデザインの年代分布は年配層に偏っており、おしゃれなデザインと分布が異なっていた。この場合渋いデザインはベストだと言えるか？



両標本は全ての要因について同一である必要がある

次回のテーマ

次回は

「ビジネスにおける比較②

(適切なA/Bテストの活用）」

お疲れ様でした！

総務省統計局
統計研修所
Statistics Japan

社会人のためのデータサイエンス演習

第2週:分析の概念と事例

第6回:ビジネスにおける比較② (適切なA/Bテストの活用)

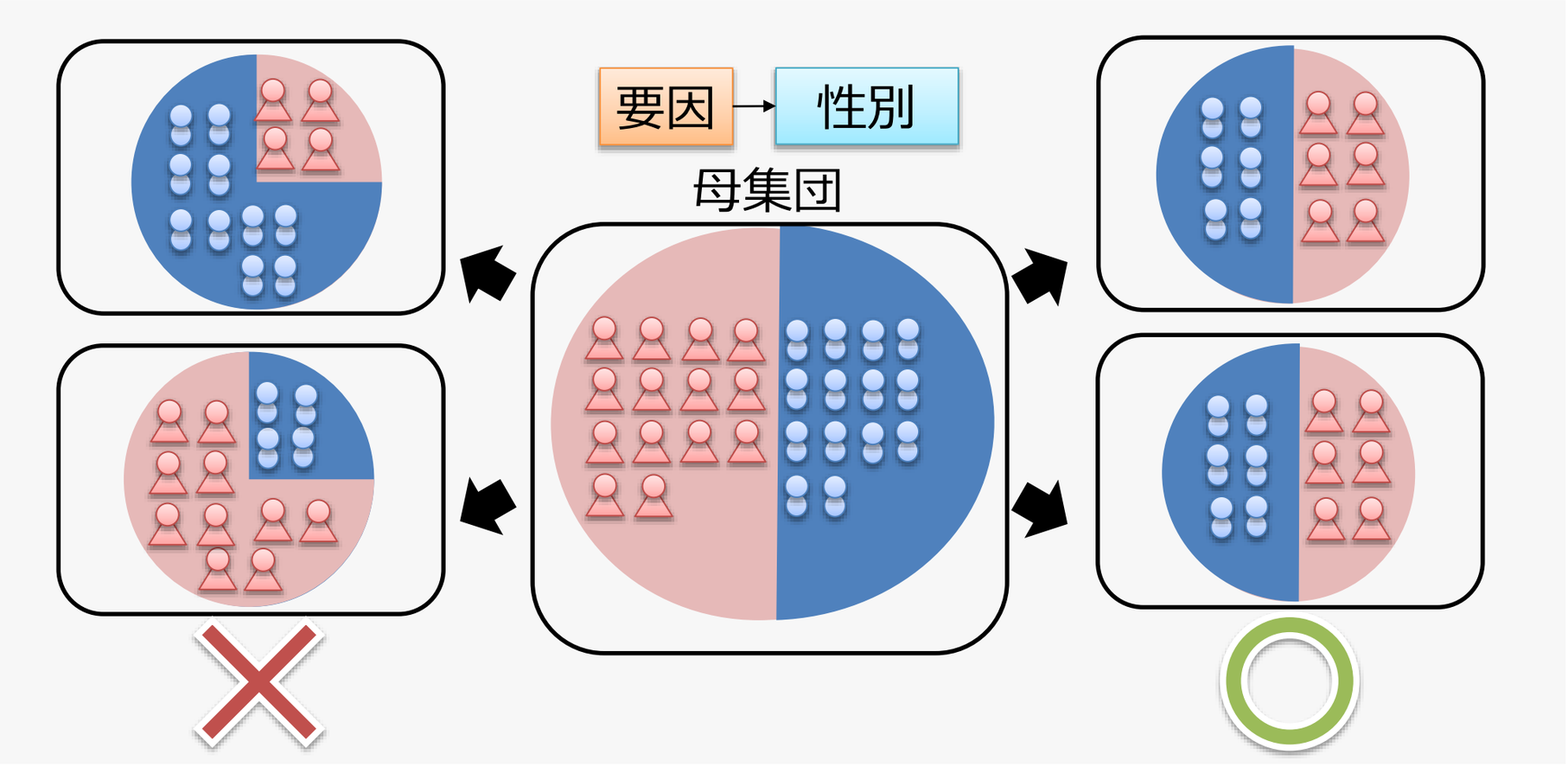
講師名:渋谷 直正

第2週の内容紹介

第1回	● Analysis (分析) とは
第2回	● 1変数の状況の把握① (可視化の活用)
第3回	● 1変数の状況の把握② (代表値の活用)
第4回	● 比較して2変数の関係を見る
第5回	● ビジネスにおける比較①(概要)
第6回	● ビジネスにおける比較②(適切なA/Bテストの活用)

公平な比較を行うためのロジック

- 全ての要因について両標本が等しい必要がある

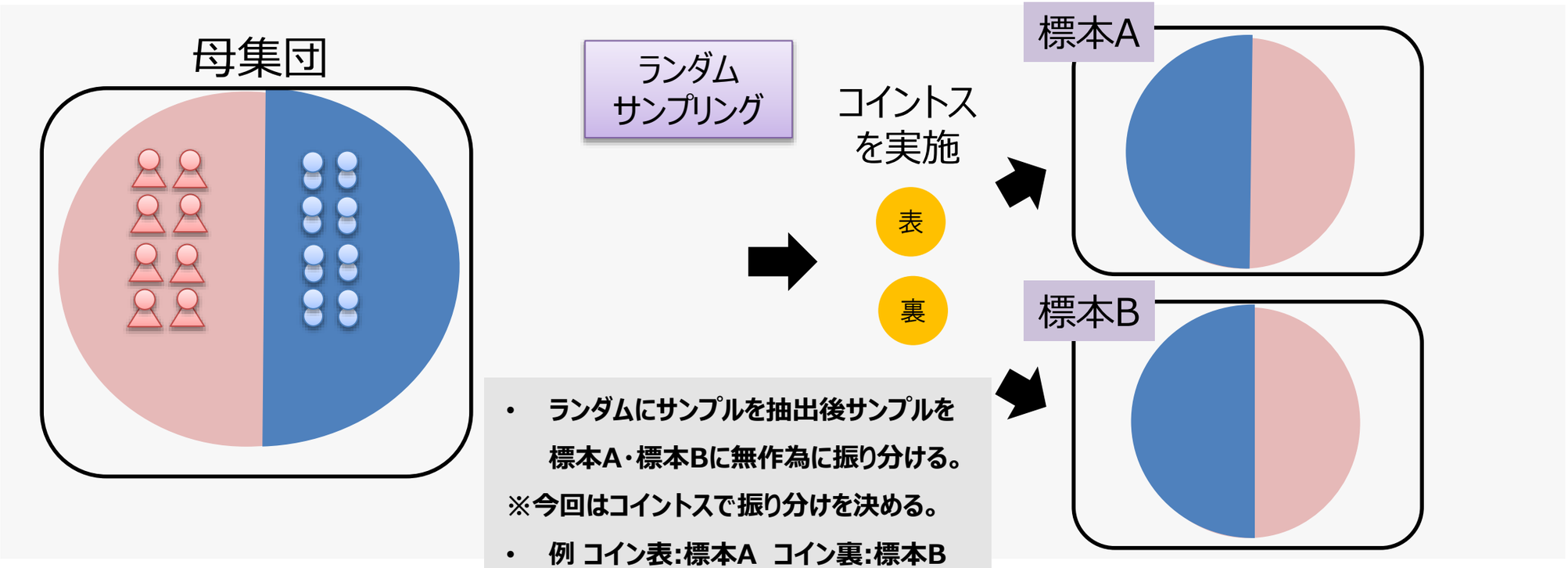


データの全種類が両標本に同率で含まれるようにする

ランダムサンプリングの紹介

- データから標本をランダムに（無作為に）抽出すること

データの各種類が選択される確率はそれぞれの頻度と等しい → **全種類が両標本に同率で含まれることを確保できる**



- ランダムにサンプルを抽出後サンプルを標本A・標本Bに無作為に振り分ける。

※今回はコイントスで振り分けを決める。

- 例 コイン表:標本A　コイン裏:標本B

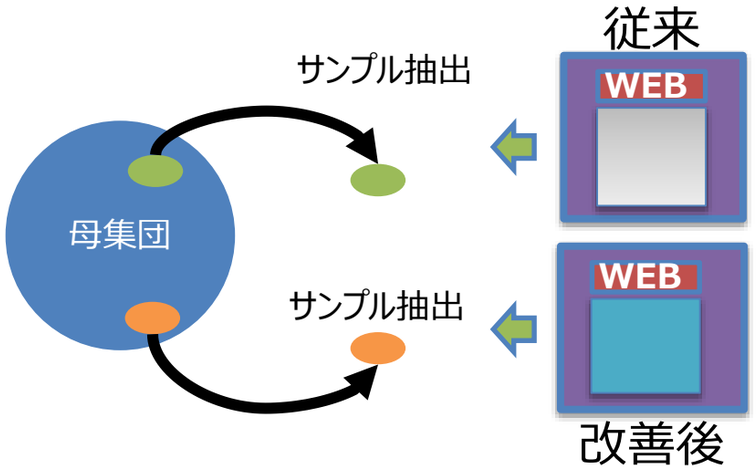
ランダムサンプリングにより、公平な比較を実現できる

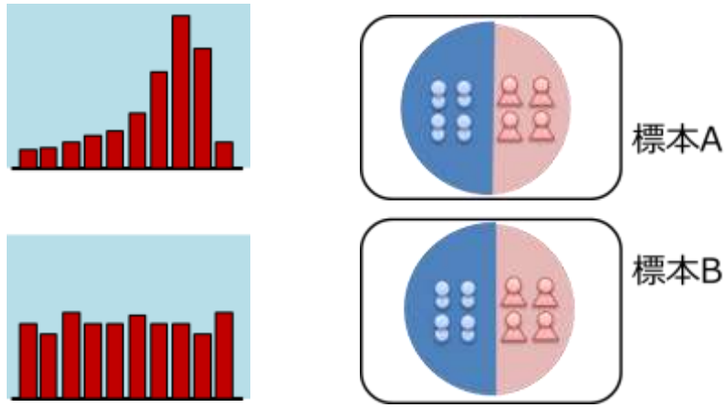
ビジネスにおける比較まとめ

- 全ての要因について両標本が等しい必要がある

標本はすべての要因について同一である必要がある

分布が偏らないようランダムサンプリングを心がける





正しいサンプリングと比較を実施することで

より正確にA/Bテストの効果を測定できる

次週のテーマ

次週は

「分析の具体的手法」

お疲れ様でした！