

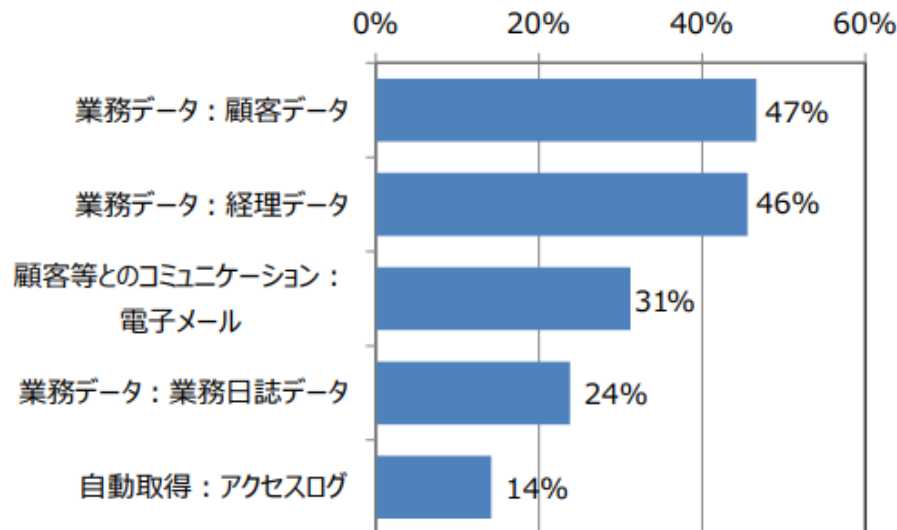
第04回 補助資料

1. データの見える化（可視化）の重要性

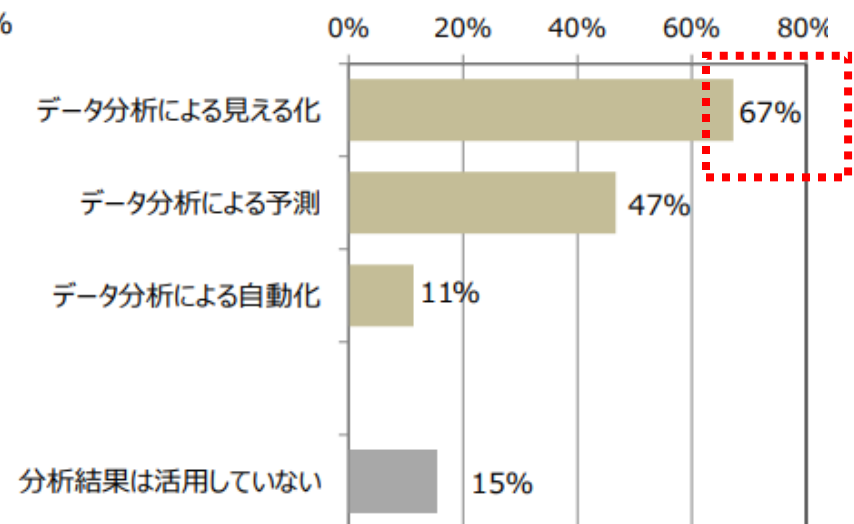
国内企業におけるデータ分析の実態

- 国内企業のデータ分析は「業務データ」の「見える化（可視化）」がスタート
- 分析に活用しているデータとして「顧客データ」、「経理データ」の割合が高くなっています。
 - いずれも意図的に取得したデータではなく、自然に集まる業務データとなっています。
- データ分析の活用方法として、最も割合が高いのは「データ分析による見える化（可視化）」の67%です
 - 「見える化（可視化）」とは、図表作成などを行うことでデータを分かりやすく示すことを指しています。

分析に活用しているデータの割合（複数回答：降順上位5位）



データ分析の活用方法（複数回答）



【出所】ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究〔総務省（調査委託先：株式会社 情報通信総合研究所）〕に基づき作成

http://www.soumu.go.jp/johotsusintokei/linkdata/h27_03_houkoku.pdf

データの可視化とは

- データ可視化はデータサイエンスの一領域
- 「可視化」とは人が直接見ることでできない現象、事象、関係性を見れるようにすること
(画像・グラフ・図・表など)
- 最新のデータ可視化は、以下に分類される
 - データビジュアライゼーション (グラフ等)
 - インフォグラフィックス
- 視覚表現の方法
 - 「長さ」「大きさ」「角度」「色」

データビジュアライゼーション	数字や単語が並んだデータをプログラムによって統計処理し、意味ある情報を見つけ出しやすくするもの
インフォグラフィックス	既に見つかっている意味ある情報を整理し、わかりやすく多くの人に興味を持ってもらうために表現するもの

既に見つかっている意味ある情報を整理し、わかりやすく多くの人に興味を持ってもらうために表現するもの



日刊スポーツのプロ野球一球速情報
(日刊スポーツサイトより)



	1	2	3	4	5	6	7	8	9	計	安	失
中日	0	0	0	0	0	0	0	0	0	0	4	0
阪神	0	0	0	2	1	0	0	0	x	3	6	1

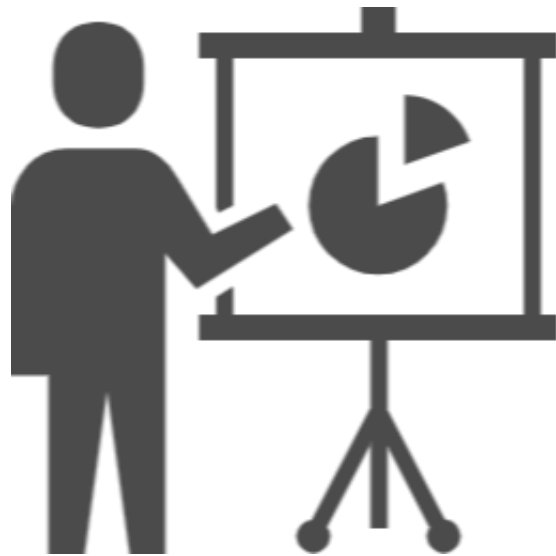
中日		
1	中	大島 洋平
2	遊	京田 陽太
3	左	福田 永将
4	一	堂上 直倫
5	三	高橋 周平
6	二	阿部 寿樹
7	右	平田 良介
8	捕	加藤 匠馬
9	投	三ツ間 卓也

阪神		
1	中	近本 光司
2	遊	北條 史也
3	左	福留 孝介
4	三	大山 悠輔
5	二	糸原 健斗
6	一	陽川 尚将

投手	青柳 晃洋	打者	高橋 周平
球数	69球	本日	2打数1安打
今季成績	9勝 9敗 0S	今季成績	.293 本7 打59
捕手	梅野 隆太郎		
5回表			
球数	12球目		
結果	ストライク(ファウル)		
球種	スライダー		
球速	124km/h		

なぜデータの可視化が重要か

- 専門家ではない人々（例えば、ビジネスでは経営層）に直感的にデータの特徴を伝えることができる
- データによるエビデンス（証拠）を効率的かつ効果的に把握することが可能になり、意思決定（判断）のスピードが高まる
- トレンドの洞察を得られ、予測が立てられる



見る/する：80%

読む：20%

聞く：10%

注目されるデータストーリーテラー

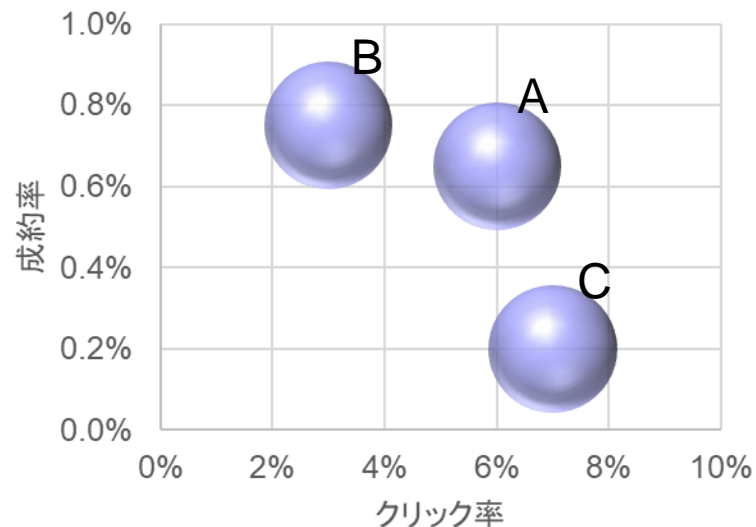
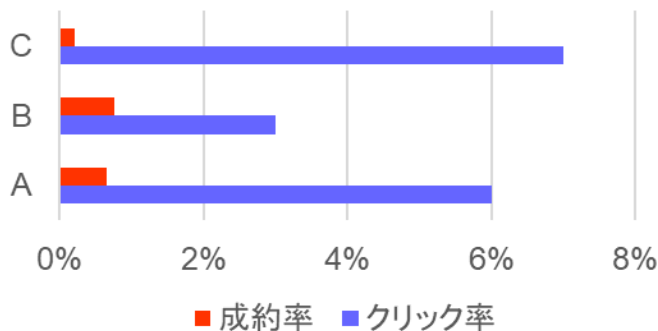
- データストーリーテリングは、事実を提示するだけでなく、「物語」として伝えることで、相手により強い印象を与えることができる手法
- データストーリーテリングにおいてデータ可視化は有用な手段。冷たい数字とファクトを多彩な色、図形、チャートに表示することで、メッセージの共感度が高まると考えられている
- データストーリーテラーはデータストーリーテリングのプロ。データサイエンスが社会に浸透した近年において、データストーリーテラーの役割が注目されている
 - ビジネス領域では経営層へデータ分析の価値を訴える役割

データストーリーテリングで意識すべきこと

- 実現すべきゴールを明確にする
- データを伝わるように表現する

クリック率と成約率からどのサイトに広告を出すかを検討し、あなたは上司にAがよいことを伝えたい。どの表現が伝わりやすいでしょうか？

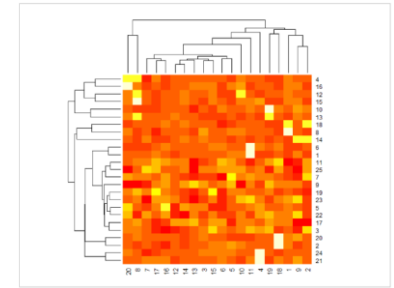
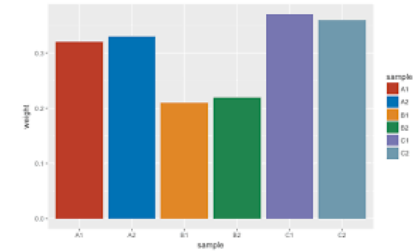
	クリック率	成約率
A	6.00%	0.65%
B	3.00%	0.75%
C	7.00%	0.20%



2. データの可視化方法

データ可視化方法の分類

①大きさによる可視化	特定の図形の長さ、高さまたは面積によって、異なる指標に対応する数値と数値間の差を表現 棒グラフ、円グラフ、折れ線グラフなど
②色による可視化	データの強弱や密度を色や濃淡として表現 ヒットマップなど
③画像による可視化	実際の意味を持つ画像やアイコンを用いれば、データとチャートをよりリアルに表現
④地図による可視化	地域間のデータを比較する場合に有効 GIS (Geographic Information System) とも呼ばれる



大きさによる可視化のためのグラフの種類

グラフの種類	使う場面
棒グラフ	棒の高さで、量の大小を比較
折れ線グラフ	量の増減の変化の推移を見たい
円グラフ	全体の中での構成比を見る
帯グラフ	構成比を比較
レーダーチャート	複数の指標をまとめる
散布図	2種類のデータの関係（相関）を見る
ヒストグラム（密度プロット）	データの散らばり具合を見る
箱ひげ図	データの散らばり具合を見る（比較）

統計解析の観点では、データ間の関係やデータの分布（散らばり）を確認することは重要

グラフィイメージ①

実務におけるグラフの作り方

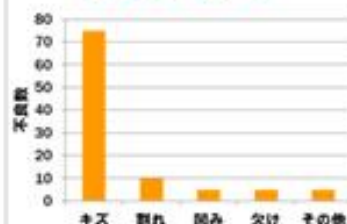


2. グラフの種類と特徴

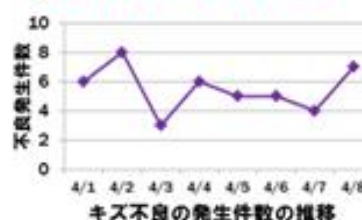
円グラフ



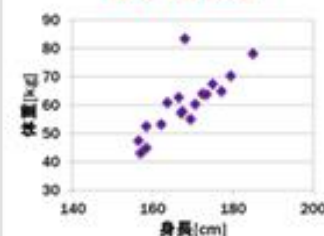
棒グラフ



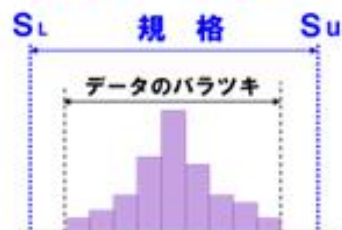
折れ線グラフ



散布図



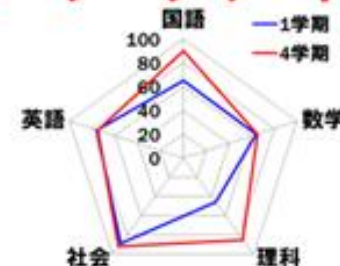
ヒストグラム



帯グラフ

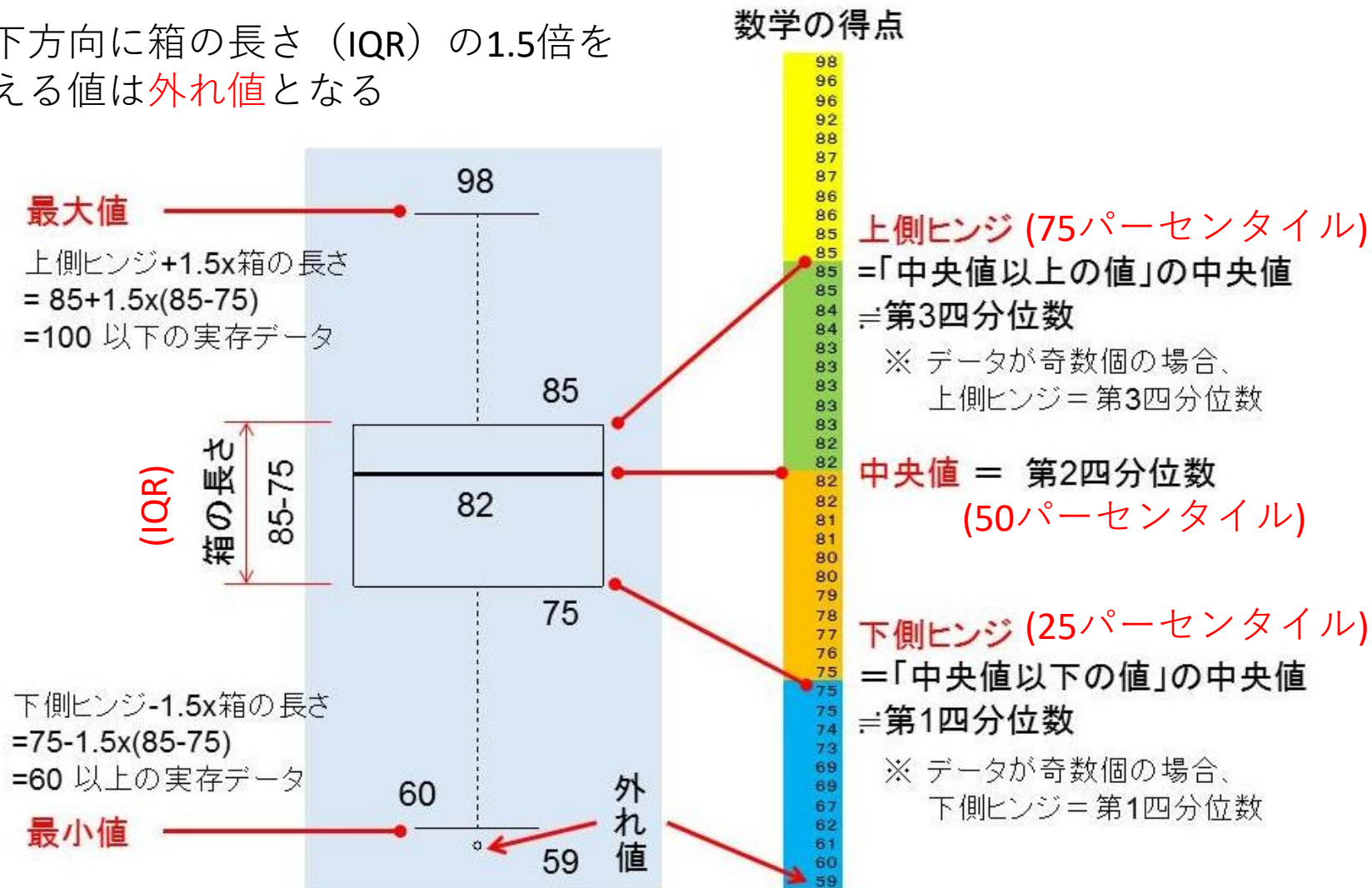


レーダーチャート



グラフィイメージ②：箱ひげ図

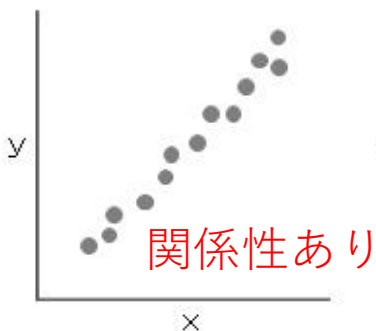
上下方向に箱の長さ (IQR) の1.5倍を超える値は**外れ値**となる



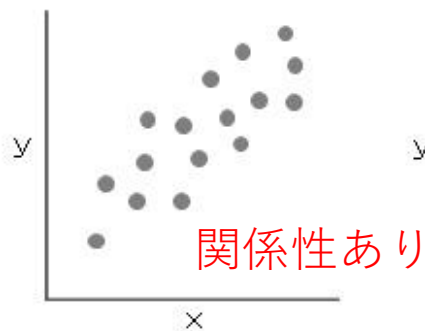
散布図と相関

- 散布図とは、2つの変数の間の関係を見るために、縦軸と横軸に目盛りを設けてデータをプロットした図
- 相関関係性が強いほど**直線状**に点が並ぶ(**相関がある**)

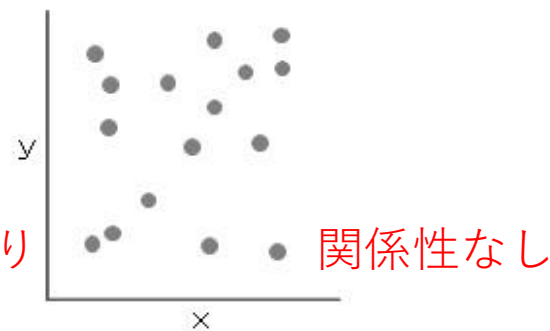
(1) 強い正の相関がある



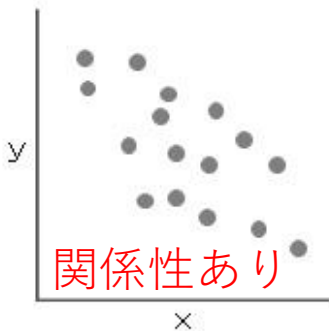
(2) 正の相関がある



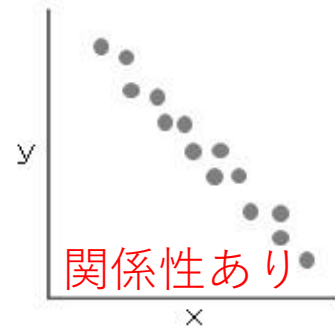
(3) 無相関



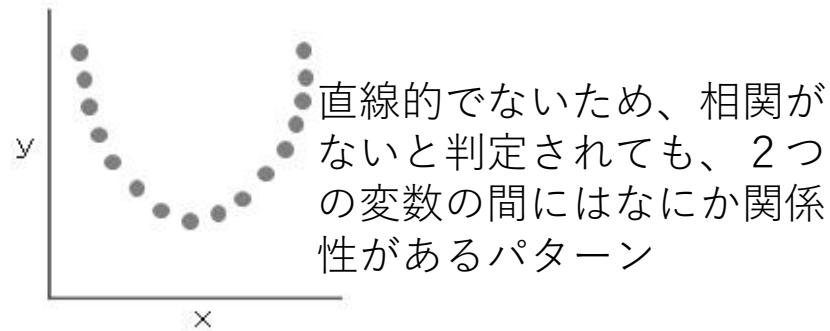
(4) 負の相関がある



(5) 強い負の相関がある



(6) ?



相関係数

- 直線的な相関関係の強さを表す指標
- 相関係数 r_{xy} の範囲： $-1 \leq r_{xy} \leq 1$

2つの要素xとyの相関係数

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

平均値

分母：xとyの共分散（2変数の関係を表す）

分子：xの標準偏差 × yの標準偏差（単位の違いを調整）

相関係数	相関の強さ	解釈
$0.7 < r_{xy} \leq 1.0$	強い 正 の相関	xが増加すれば yも増加する関係
$0.4 < r_{xy} \leq 0.7$	適度な 正 の相関	
$0.2 < r_{xy} \leq 0.4$	弱い 正 の相関	
$-0.2 < r_{xy} \leq 0.2$	相関はほぼなし	xとyの間に関係性はない
$-0.4 < r_{xy} \leq -0.2$	弱い 負 の相関	
$-0.7 < r_{xy} \leq -0.4$	適度な 負 の相関	
$-1.0 < r_{xy} \leq -0.7$	強い 負 の相関	

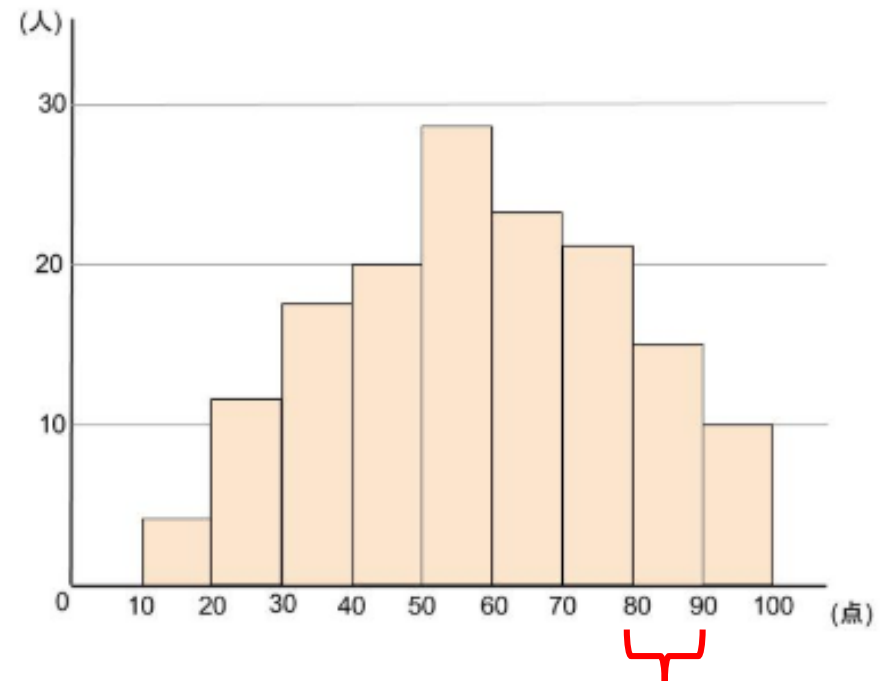
ヒストグラムとデータ分布

- 縦軸に度数、横軸に階級をとった統計グラフで、データの分布を視覚的に読み取るもの

階級 度数

階級数=10個

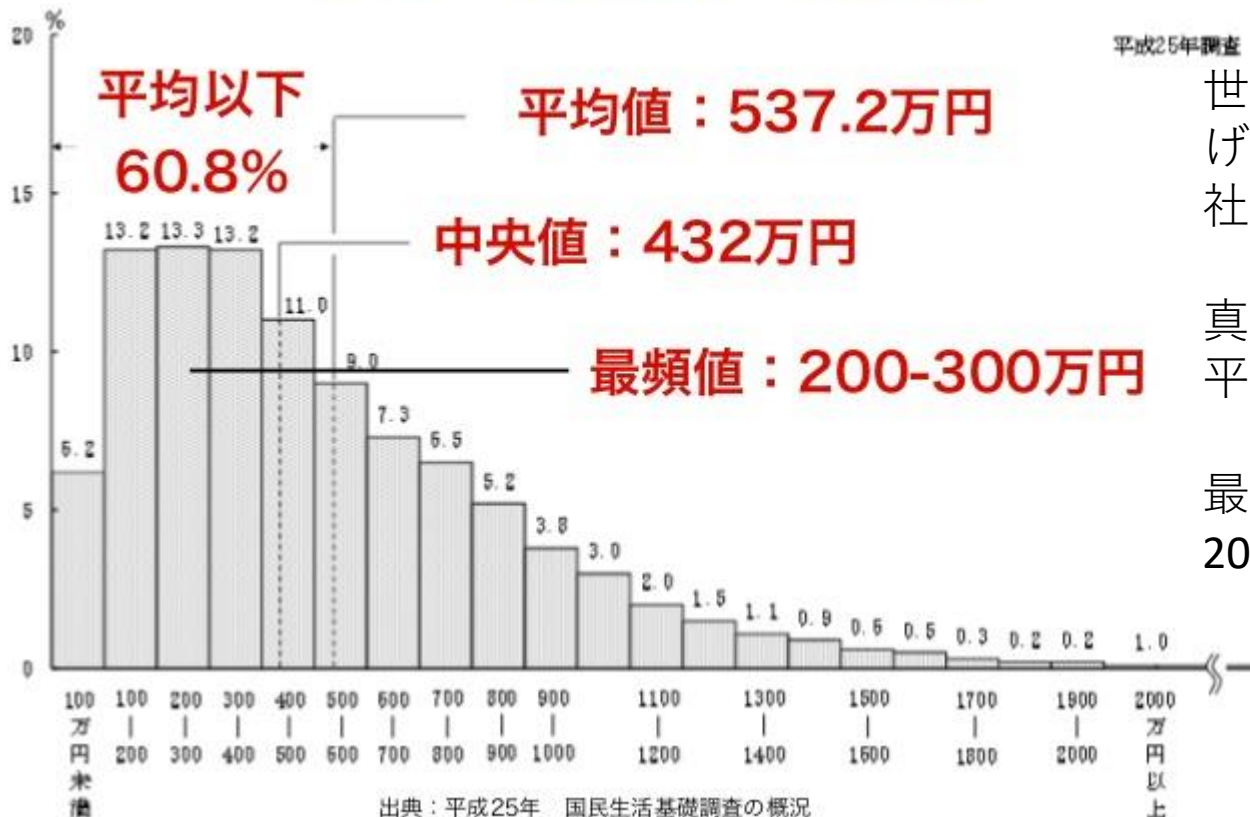
得点(点)	人数(人)
以上 未満 90~100	10(100点も含む)
80~90	15
70~80	21
60~70	23
50~60	28
40~50	20
30~40	17
20~30	12
10~20	4
0~10	0
計	150



データの分布を見ることの重要性

- 以下のようなヒストグラムから何が読み取れるか

参考：世帯所得の統計値



世帯所得の平均値くらい稼げたらよいと思いますか。社会はそんなに甘くない

真ん中の世帯（中央値）と平均は約100万円も違う

最も多い層（最頻値）は、200-300万円

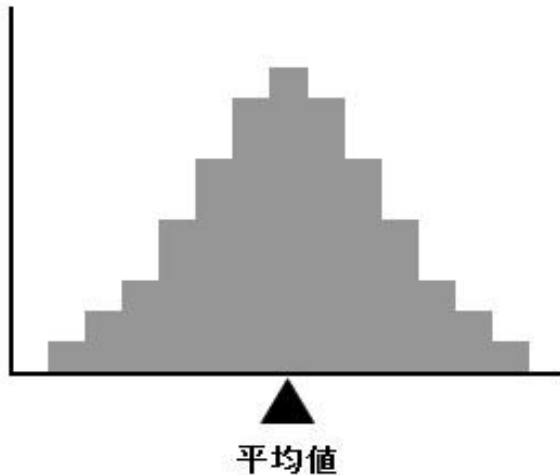
ヒストグラムの読み取り方

データの分布の仕方によって、代表値（データの特徴を表す指標）は異なる

図表1 主な代表値

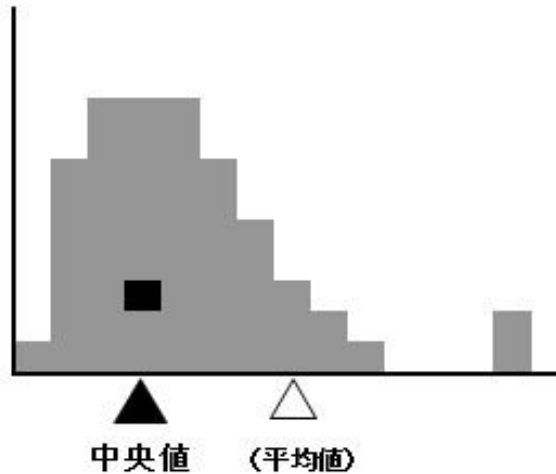
平均値が適している場合

- ・ 極端に高い（低い）データがないとき
- ・ データの分布に極端な偏りがないとき



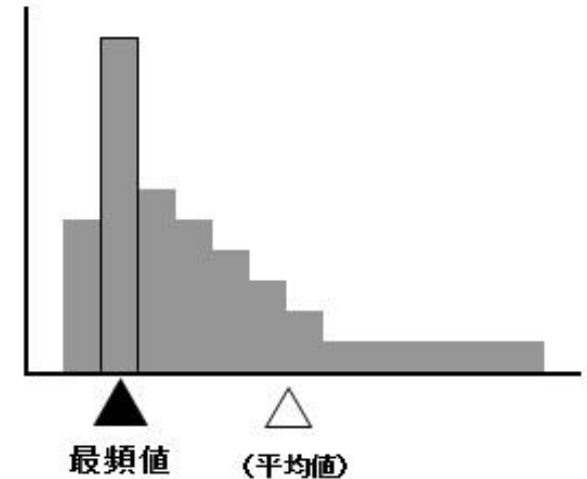
中央値が適している場合

- ・ 極端に高い（低い）データがあって、平均値がその影響を強く受けるとき



最頻値が適している場合

- ・ データが特定の階級に集中しているとき
- ・ データの分布に極端な偏りがあるとき



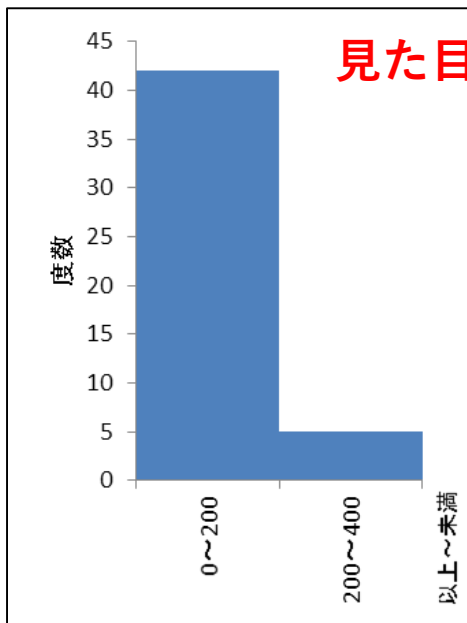
ヒストグラムの作図の注意：

階級幅・階級数

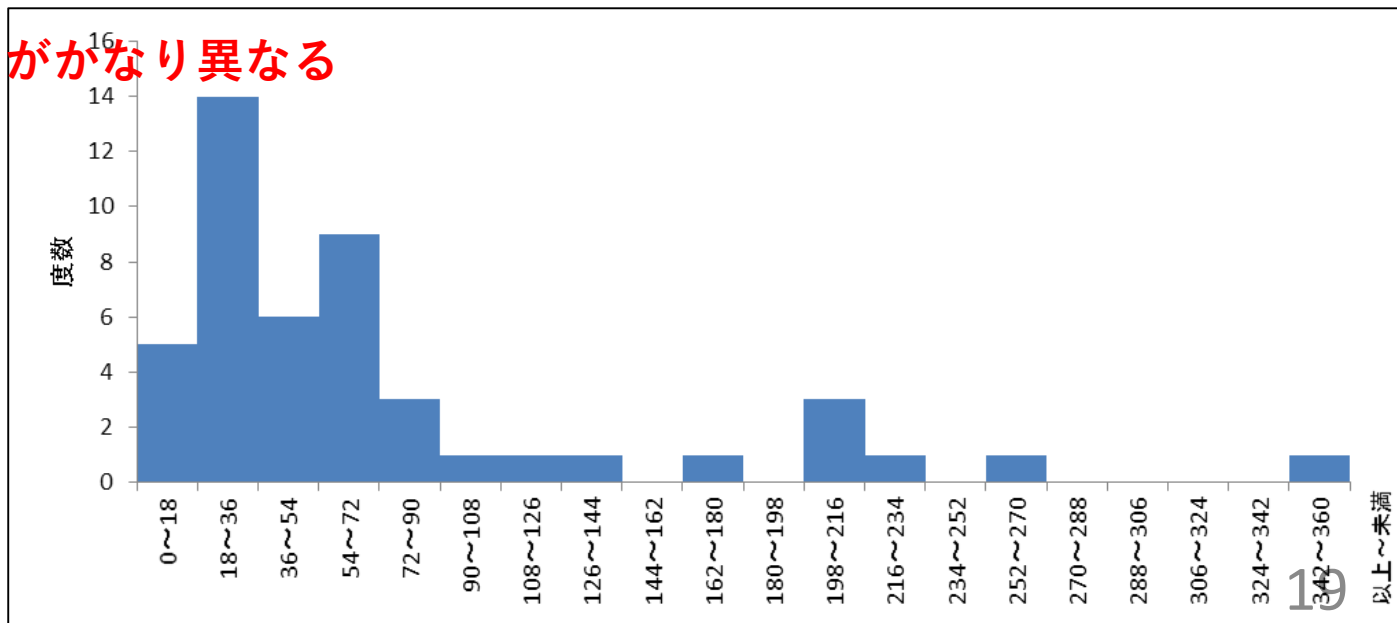
- 階級幅の取り方によって見た目の印象が大きく変わること
- 階級幅が大きすぎても、逆に小さすぎてもデータの分布が分かりづらくなる
- 階級幅の決め方に決まったルールはないが、困った場合は**スタージェスの公式**で決めた階級数から階級幅を決める方法がある

スタージェスの公式: 階級数 = $1 + \log_2 N$ (N: データ数)

階級幅=200



階級幅=18



見た目がかなり異なる