2-1-3. 代表値の性質の違い

2. データリテラシー
2-1.データを読む

この項目で学ぶこと

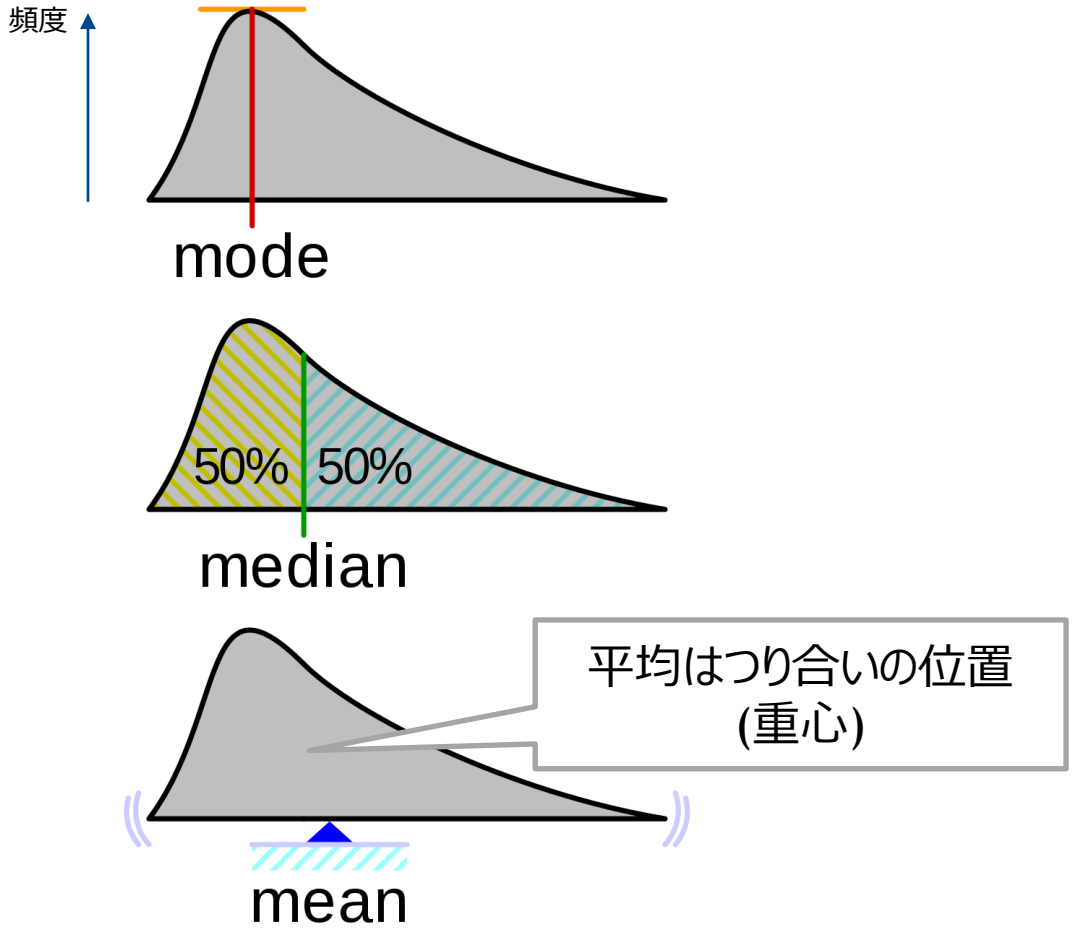
- 代表値（平均，中央値，最頻値）は性質が違う
  - ＝「代表」といっても，色々ありうる
- 代表値は（便利だが），使う場合には注意しよう
  - 上記のように性質が違うのだから，どの代表値を使うべきか考えよう
- 代表値が，「データの代表になっていない」可能性がある

この項目を学ぶ意義

- 「代表値＝平均でOKでしょ？」という考えを捨てられる
  - 平均がデータの代表としてふさわしくない場合もある
  - 「どの代表値が自分の分析にふさわしいか」を考えよう！
- 代表値に対する「はずれ値」の影響がわかる
  - はずれ値＝例外的な値＝データ分析の厄介者
- はずれ値に強い代表値があることを知る

2-1-2の復習：平均値，中央値，最頻値

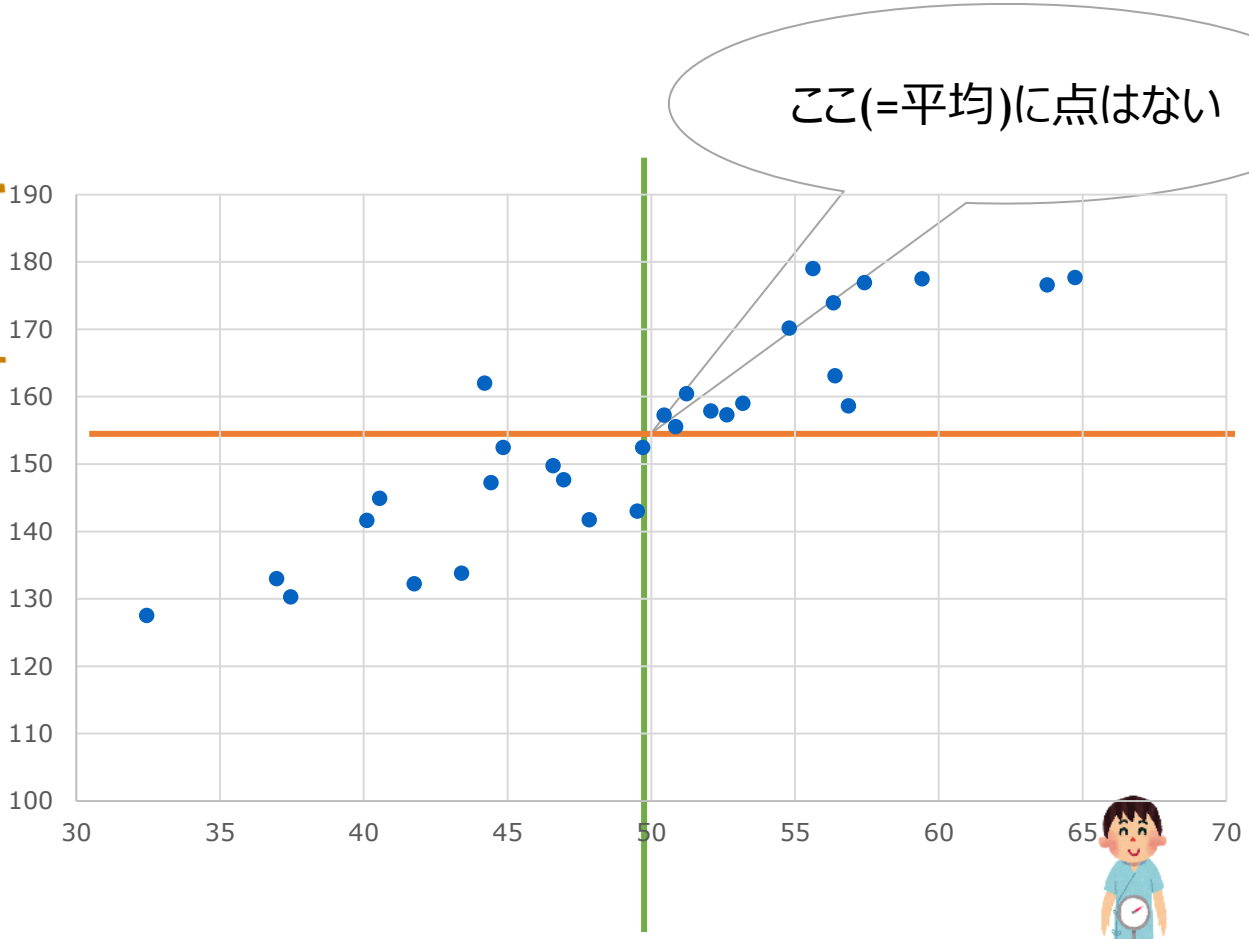
- 最頻値
  - 一番頻度の高い値
- 中央値
  - 大きい順に並べた時にちょうど真ん中に来る値
- 平均
  - 全データを合計して，データ数で割った値



https://ja.wikipedia.org/wiki/中央値#/media/ファイル:Visualisation_mode_median_mean.svg

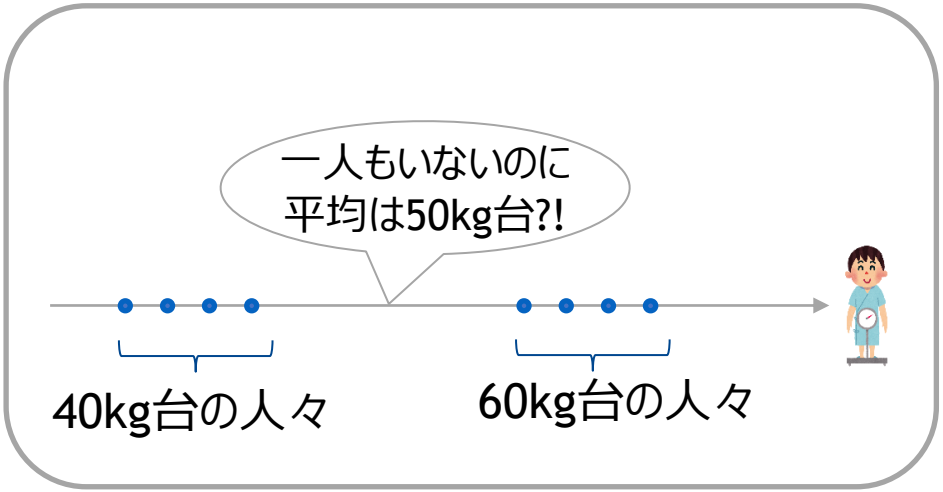
平均は代表値としてふさわしくない場合がある (1/3)
平均と全く同じデータがあるわけではない

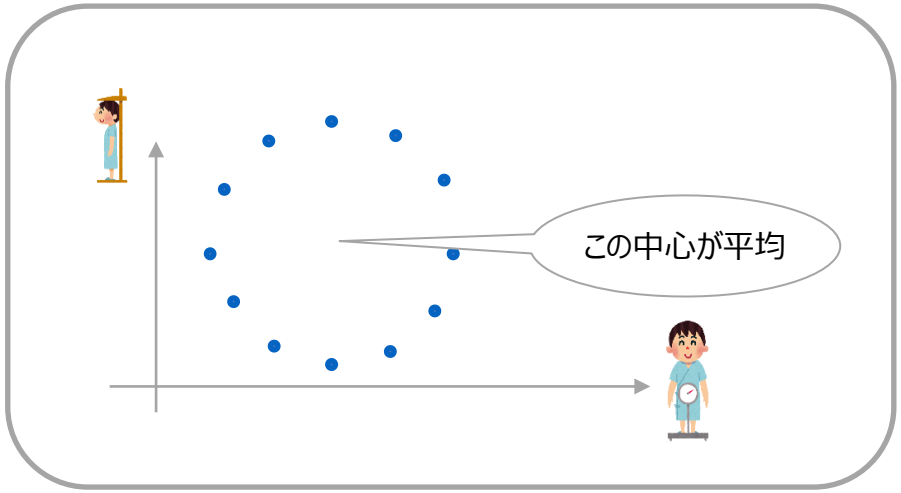
約154.7cm



約49.4kg

平均は代表値としてふさわしくない場合がある (2/3)
「よくある値」ですらないことも





なので，2-1-2のこの話は
いつもそうとは限らない

- A3: 平均すると約49.4kg

ああ，それぐらいの
体重の人が多いのね

平均は代表値としてふさわしくない場合がある (3/3)
はずれ値に弱い

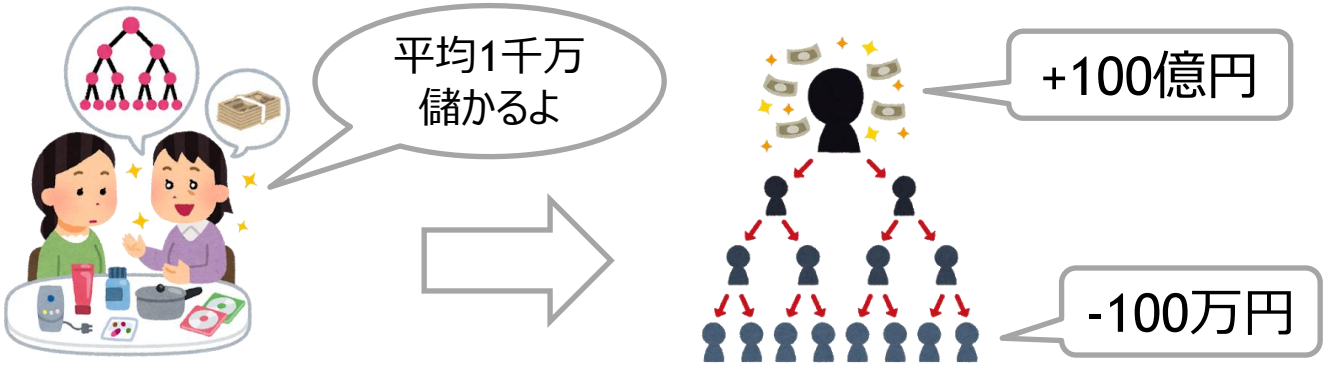
- はずれ値＝例外的なデータ(次スライド)

- 例：N = 5人の体重{62, 50, 49, 53, 550000}

1人は巨大ロボだった…

  - 平均でおよそ110043kg (!?)

- 「平均」を悪用

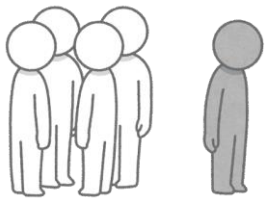


  - ウソではないかもしれないが，ごく少数の人だけが莫大な利益を上げ，残り大多数は大損している可能性も

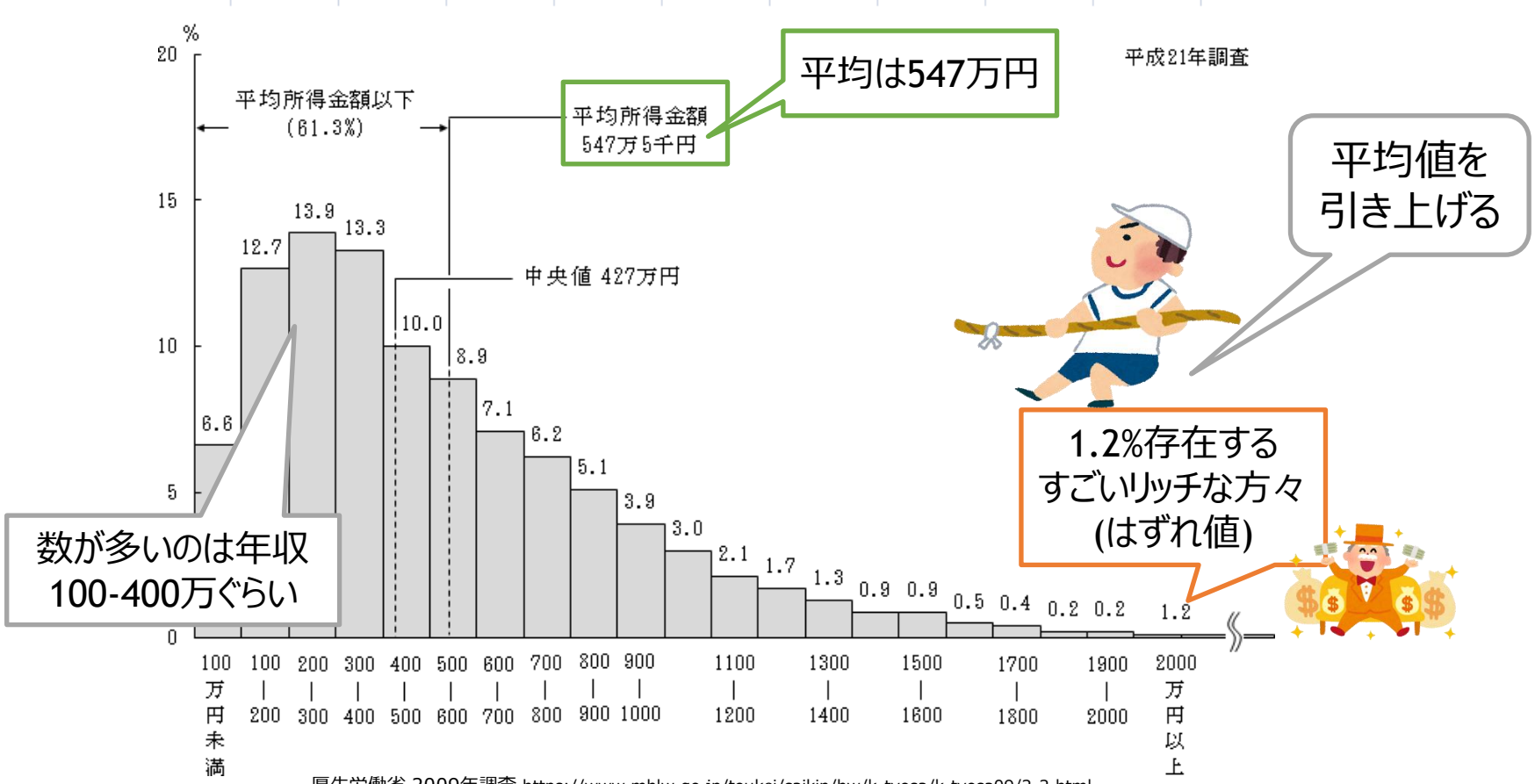
はずれ値＝例外的な値

- 一般的なデータと著しく異なる値を持つデータ
  - 結構よくある

- はずれ値は色々な原因で発生する
  - 測定ミス
  - 測定機器の故障
  - イイカゲンな回答者
  - 異常現象
  - 希少現象(めったに起きない現象)
  - 想定外の現象・初めて発生した現象



はずれ値の影響:
日本人の年収(ヒストグラム)を例に



厚生労働省 2009年調査 https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html

中央値（メディアン）は「はずれ値」に強い

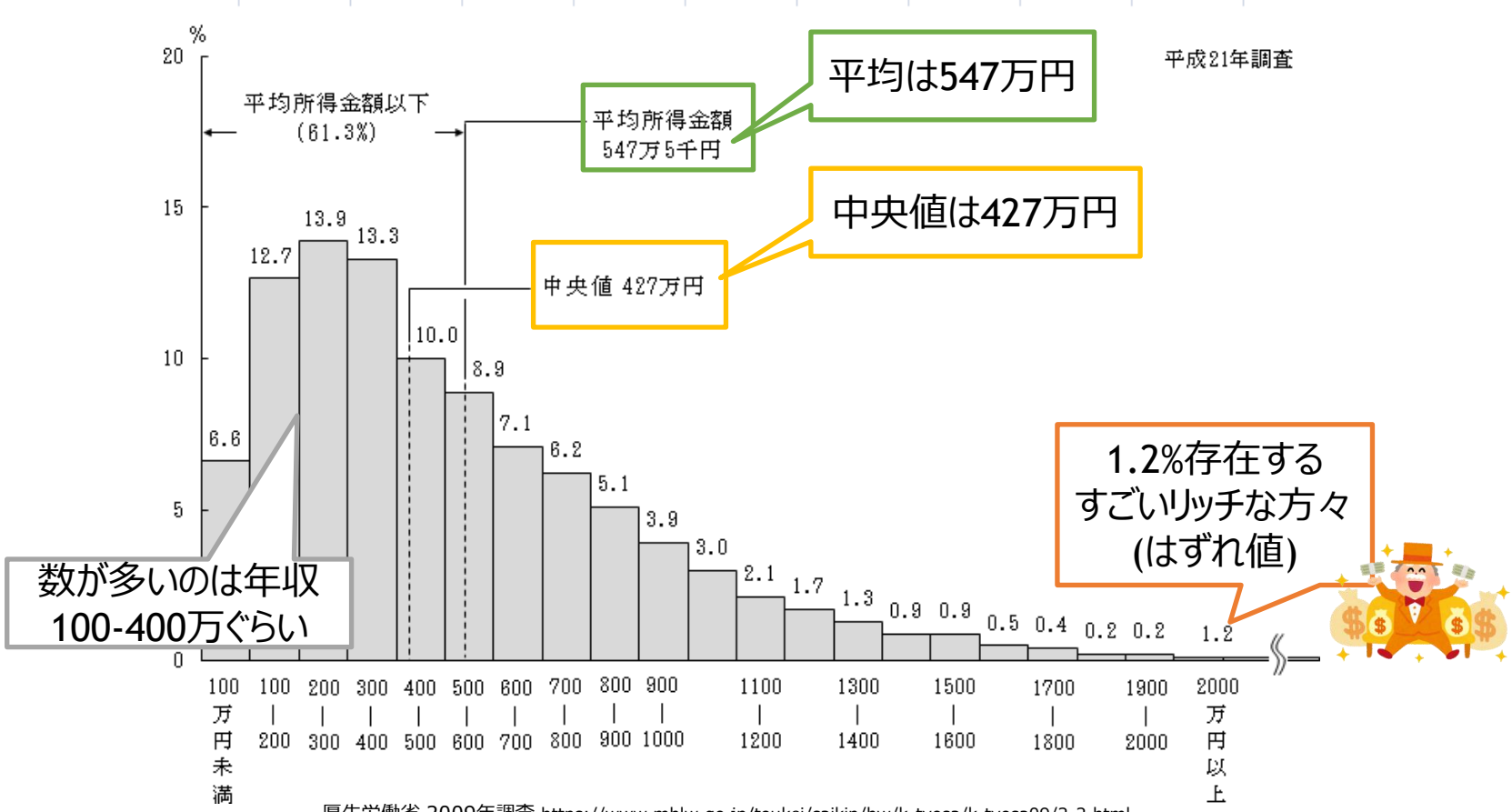
- 例： N = 5人の体重{62, 50, 49, 53, 550000}の場合
  - 並べ替えると，49, 50, 53, 62, 550000
  - なので，中央値は53

並みはずれた体重

- はずれ値に全く影響されていない!

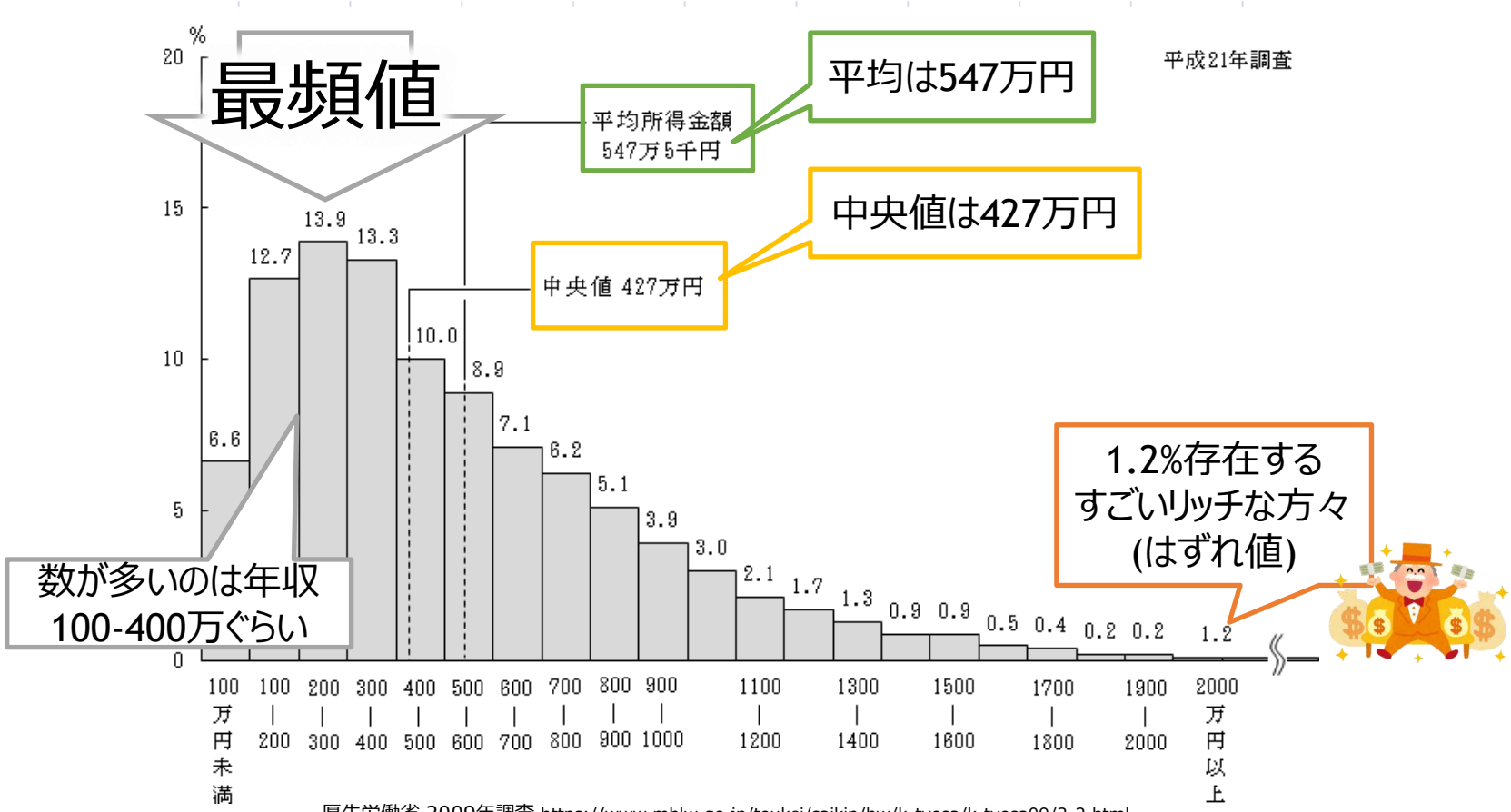
中央値（メディアン）は「はずれ値」に強い: 日本人の年収（ヒストグラム）を例に

- 中央値のほうがはずれ値に影響されにくそう



厚生労働省 2009年調査 https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html

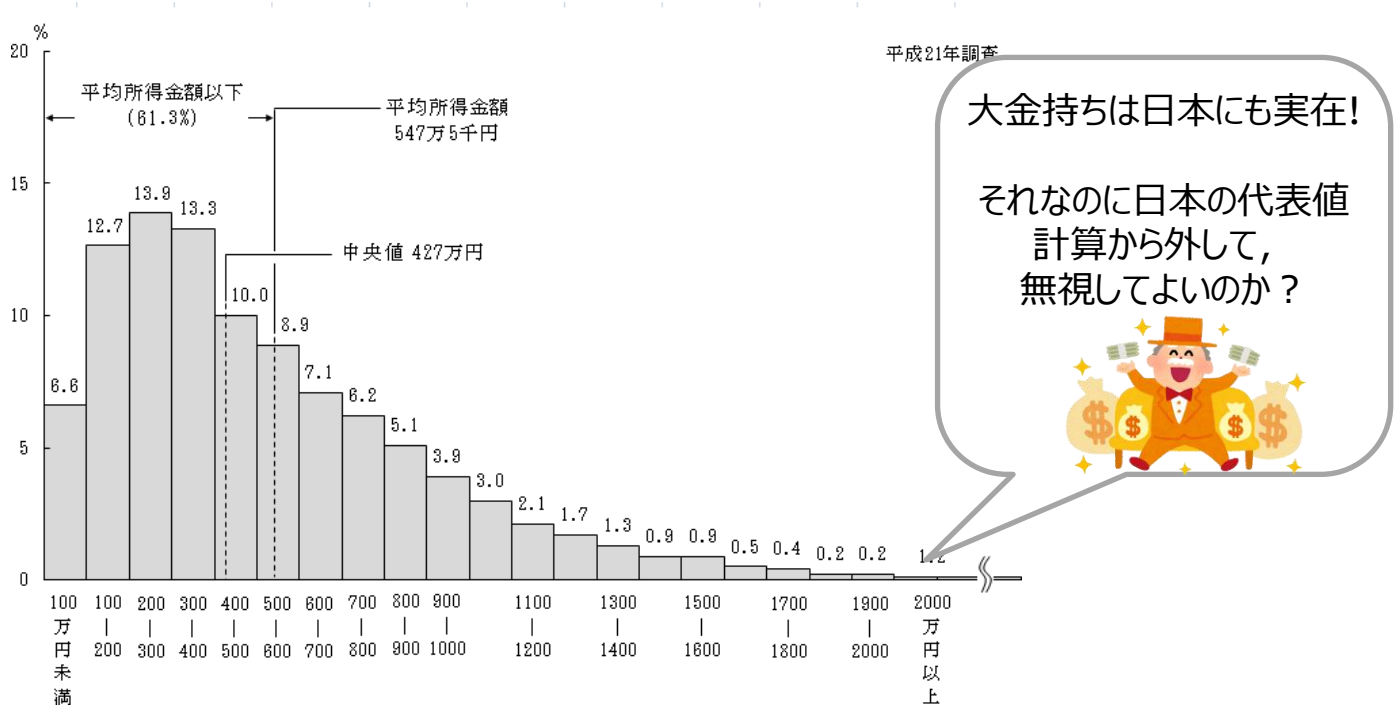
最頻値 （モード）も「はずれ値」に強い：日本人の年収（ヒストグラム）を例に



厚生労働省 2009年調査 https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html

確かに，中央値や最頻値は「はずれ値」に強い：しかし本当にそれでいいのか？(1/2)

- はずれ値でも「本当のデータ」の場合もある



厚生労働省 2009年調査 https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html

確かに，中央値や最頻値は「はずれ値」に強い：しかし本当にそれでいいのか？(2/2)

- 「はずれ値」として見捨てていいかは場合による
  - ex. 毎月のインフルエンザ死亡者数（10万人あたり）

1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2.1	7.3	9.1	4.6	1.6	0.5	0.1	0.1	0.1	0.1	0.3	1.4

小→大の順に並べ替え

7月	8月	9月	10月	11月	6月	12月	5月	1月	4月	2月	3月
0.1	0.1	0.1	0.1	0.3	0.5	1.4	1.6	2.1	4.6	7.3	9.1

中央値 0.5 or 1.4

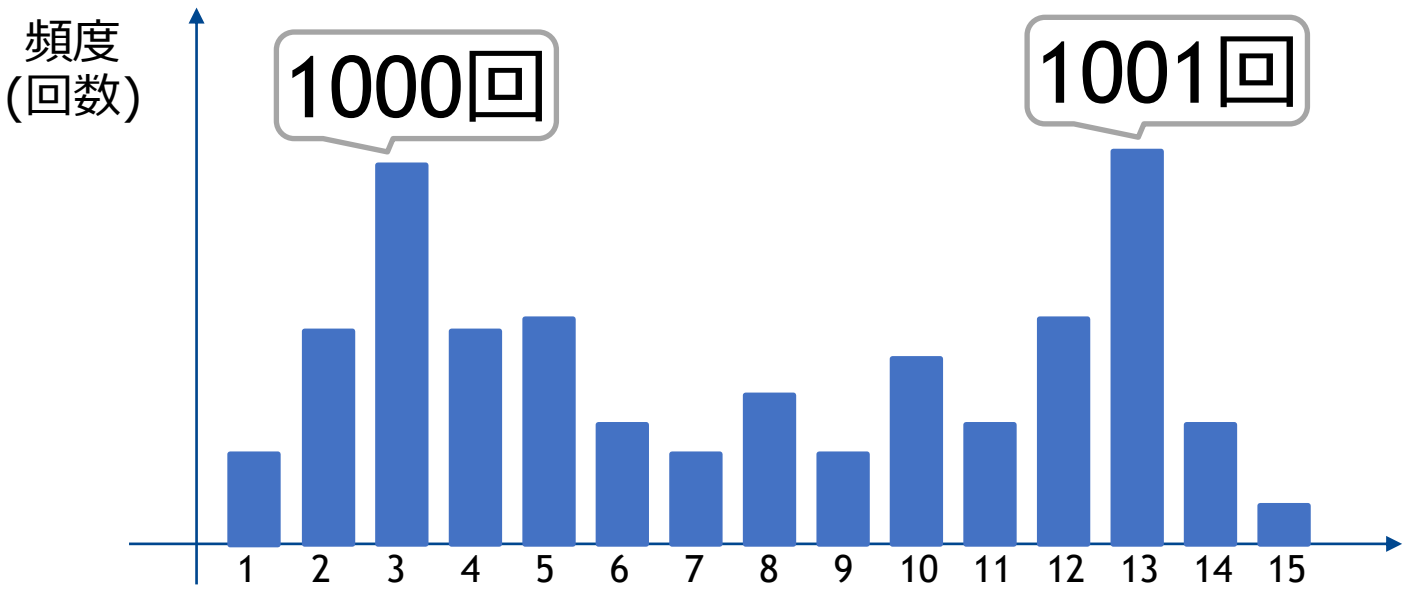
※データが偶数個なので2つある

はずれ値として
中央値に影響なし
→いいのか？

平成21年 人口動態調査 月別にみた死因簡単分類別死亡率（人口10万対）
https://www.e-stat.go.jp/dbview?sid=0003214726

最頻値（モード）には別の弱点も

- 似たピークが複数ある場合，わずかな差異で全く異なる最頻値になる



- 最頻値は13だけど，わずかな頻度の差異で3にもなりうる

平均値，中央値，最頻値の比較

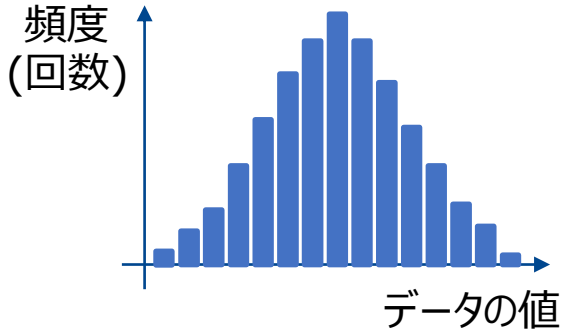
	対象となるデータ	はずれ値への強さ	データの中の一つの値かどうか？	その他
平均値	比率データ 間隔データ	弱い	そうとは限らない	データの広がりの「重心」
中央値	比率データ 間隔データ 順序データ	強い	データの一つが中央値として選ばれる(※)	並べ替えという処理がはいるので，数学的には扱いにくい
最頻値	比率データ 間隔データ 順序データ カテゴリデータ	強い	データの一つが最頻値として選ばれる	データが連続的な値を持つ場合は使いづらい 頻度に差のないデータの場合，わずかな頻度差に影響される

どんな方法も万能ではない！
メリット・デメリットを見極めて，
適切な方法を選択すること！

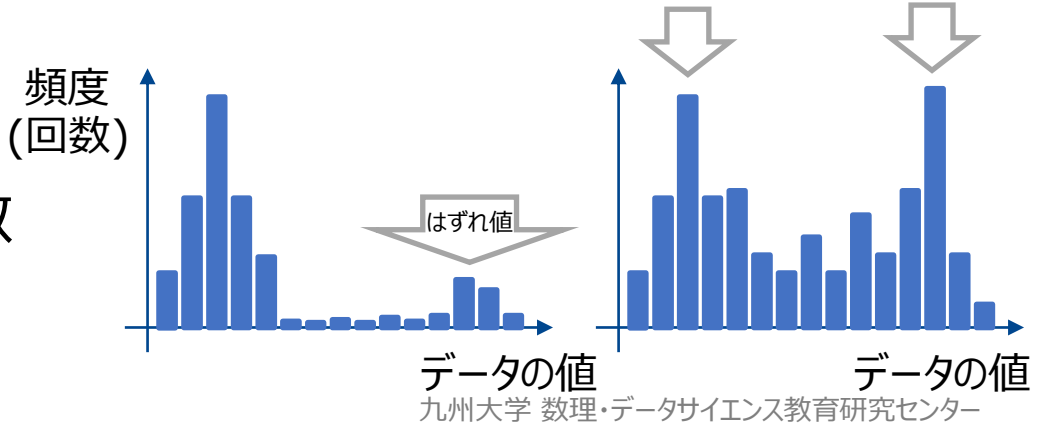
※データが偶数個の場合は，中央順位2つの平均（N=6なら第3,4位の値の平均）を利用するので，これは成り立たない．

代表値の使い方に関する「お勧め」：全ての代表値(平均/中央値/最頻値)を出してみる

- もし，全部がだいたい同じなら
  - はずれ値があまりない
  - ヒストグラムで書くと，左右対称な山が一つ



- 値が結構違うなら
  - はずれ値がある
  - ヒストグラムで書くと山が複数
    - そもそも「1つの代表値」だけで表すべきでないデータの可能性あり



- 以上に加え，分散(→2-1-4)を出してみるのも効果的

まとめ

- 性質を理解して代表値を使おう
  - それによって，代表値が本当にデータを代表しているか，リスクはないか，理解できる
- 代表値には限界もある
  - 代表値は，一つの数であり，それだけで全データを代表するのは無理な場合もある
- 「はずれ値」を意識しよう
  - 大規模なデータには，よく入っている
  - 無視するべきか，考慮すべきかも，データ分析の目的次第