

# 統計学I

早稲田大学政治経済学術院

西郷 浩

# 本日の目標

- 代表値
  - 算術平均、中央値、最頻値
  - 代表値と分布の歪みとの関係
- 散らばりの尺度
  - 範囲、四分位偏差
  - 分散、標準偏差、変動係数
  - 変数の標準化、変換
- 幹葉表示と箱ひげ図

# 代表値

- 代表値

- 分布の中心の位置

- 「代表」＝「集団全体の相場に対応する値」

- 対称分布について

- 中心の意味は明白

- » これから紹介する3つの代表値もほとんど等しくなる。

- 非対称分布について

- 中心を決めることが難しい

- » 分布の歪みと関連させて3つの代表値の位置関係を覚えておく。

# 算術平均

- 算術平均 (mean)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{ただし、} \quad \sum_{i=1}^N x_i = x_1 + x_2 + \cdots + x_N$$

- 代表値としての意味:

- 集団全員分の  $x$  を集めて、均等配分したときの構成員一人当たりの取り分。
- 分布の重心と解釈することもできる。

- 例: 東京都市区町村別世帯数の算術平均:

- 126,012.4 (世帯)

# 中央値(1)

- 中央値(または中位数)  $Me$  (median)
  - 集団の構成員を  $x$  の昇順に並べ替える。
    - $x_1, x_2, \dots, x_N \rightarrow x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$
  - $Me = 2$ つの「真ん中」の候補の平均
    - 昇順の順位が初めて  $N/2$  以上になる  $x$  の値
    - 降順の順位が初めて  $N/2$  以上になる  $x$  の値
      - $N$ : 奇数  $\rightarrow$  昇順で  $(N + 1)/2$  番目
      - $N$ : 偶数  $\rightarrow$  昇順で  $N/2$  番目と  $(N/2) + 1$  番目の平均

# 中央値(2)

## － 代表値としての意味：

- $Me$  以下の値をもつ構成員が半分、 $Me$  以上の値を持つ構成員が半分。
  - － ヒストグラムの柱の面積が、左右でちょうど等しくなる位置に相当する。
  - － 累積分布関数で、縦軸の0.5に対応する横軸の値。

## － 例：東京都市区町村別世帯数の中央値：

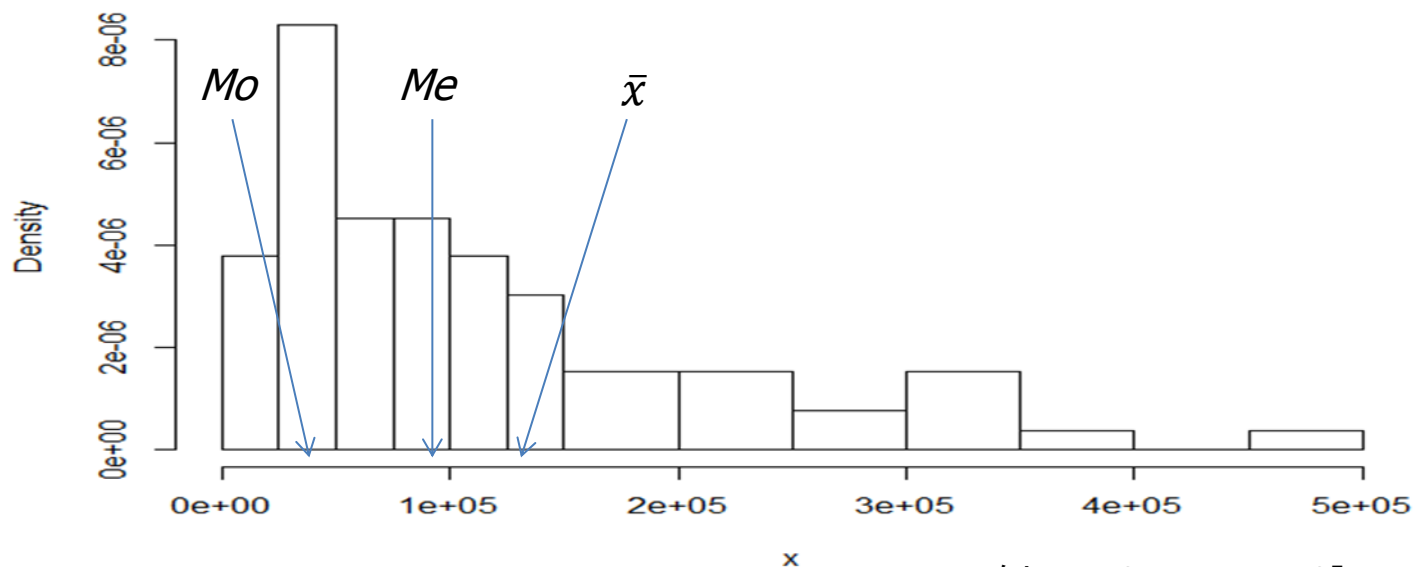
- 昇順で27番目の値＝ 89,676（世帯）
  - － 西東京市の世帯数

# 最頻値

- 最頻値  $Mo$  (mode)
  - ヒストグラムの峰に対応する階級値
    - つまり、ヒストグラムの頂点に対応する横軸の値
  - 代表値としての意味：
    - 「その近辺の値をもつ構成員がもっとも多い」という意味で人並みの値
  - 例：東京都市区町村別世帯数の最頻値
    - 37,500 (世帯)
      - 階級幅 25,000 の度数分布表を使用した場合)

# 分布の歪みと3つの代表値の位置関係(1)

図1: 代表値の位置



例:  $1e+05 = 1 \times 10^5$

資料: 総務省「平成27年国勢調査」人口速報集計結果

## 分布の歪みと3つの代表値の位置関係(2)

- 対称分布の場合

$$Mo \approx Me \approx \bar{x}$$

- 右に歪んだ分布(裾が右に長い)

$$Mo < Me < \bar{x}$$

– 算術平均よりも小さい値をもつ市区町村: 33

- 算術平均は外れ値の影響を受けやすい。
- 外れ値: 極端に大きい or 極端に小さい値

# 散らばりの重要性(1)

- 例

- 2つのクラスA, B

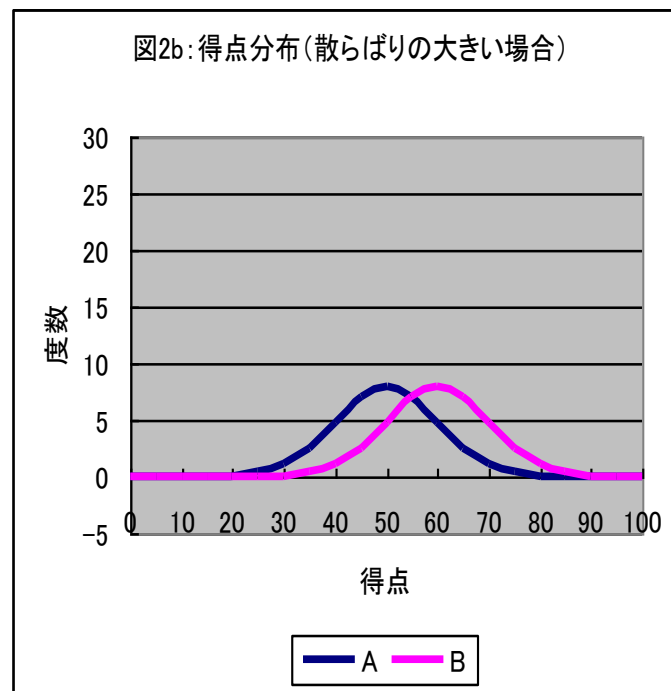
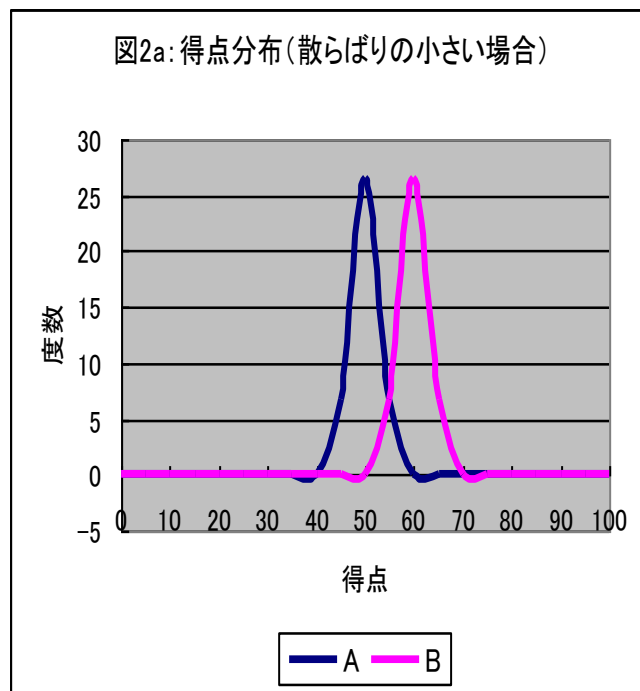
- クラスAの平均点 : 50点

- クラスBの平均点 : 60点

- 2つのクラスの平均点の差は意味があるか(クラスBの方が優秀か)？

- 得点分布の散らばりによって答が異なる。

# 散らばりの重要性(2)



# 散らばりの尺度：観点

- 観点
  - 度数分布の幅を捉える：
    - 範囲
    - 四分位範囲、四分位偏差
  - 中心からの乖離（偏差）の程度
    - 分散、標準偏差
    - 変動係数

# 散らばりの尺度：範囲

- 範囲(レンジ)  $R$

- $R = x_{max} - x_{min} = x_{(N)} - x_{(1)}$

- 例：東京都市区町村別世帯数

- $R = 462,335 - 837 = 461,498$  (世帯)

- 散らばりとしての意味

- 全部の  $x$  が存在する範囲

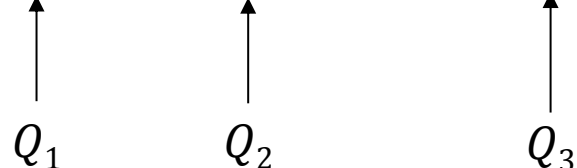
- 長短：

- 長所：わかりやすい。
    - 短所：極端な値(分布の端)だけで決まる。

# 散らばりの尺度：四分位範囲(1)

- 四分位範囲 IQR

- 四分位点： $Q_1, Q_2, Q_3$

- $x_{(1)} \leq x_{(2)} \leq \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \leq x_{(N)}$   


- 第1四分位点

- 以下の2つの値の平均

- » 昇順の順位が初めて $N/4$ 以上になる $x$ の値

- » 降順の順位が初めて $3N/4$ 以上になる $x$ の値

# 散らばりの尺度：四分位範囲(2)

－ 四分位範囲  $IQR = Q_3 - Q_1$

- ・ 例：東京都市区町村別世帯数

- －  $IQR = 186,376 - 39,520 = 146,856$  (世帯)

- »  $53 \times 1/4 = 13.25 \rightarrow Q_1 = x_{(14)} = 39,520$

- »  $53 \times 3/4 = 39.75 \rightarrow Q_3 = x_{(40)} = 186,376$

－ 散らばりとしての意味：

- ・ 昇順で中央部  $1/2$  の存在範囲

# 散らばりの尺度：四分位範囲(3)

- 参考

- 四分位偏差： $Q = \{(Q_3 - Q_2) + (Q_2 - Q_1)\}/2$

- » 中央値 ( $Me = Q_2$ ) から上下1/4の存在範囲の平均

- 長短：

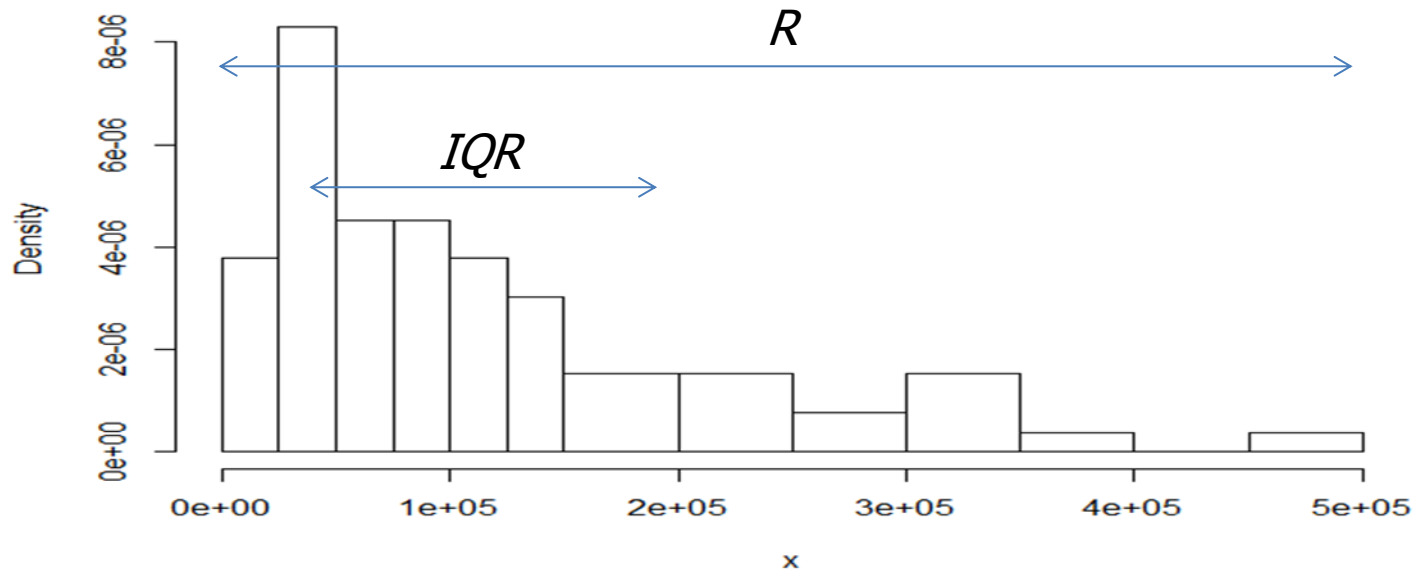
- 長所：極端な値の影響を排除している。

- 短所：使用頻度が低い。

- 後に解説する分散・標準偏差の方が使用頻度が高い。

# 散らばりの尺度：範囲と四分位範囲

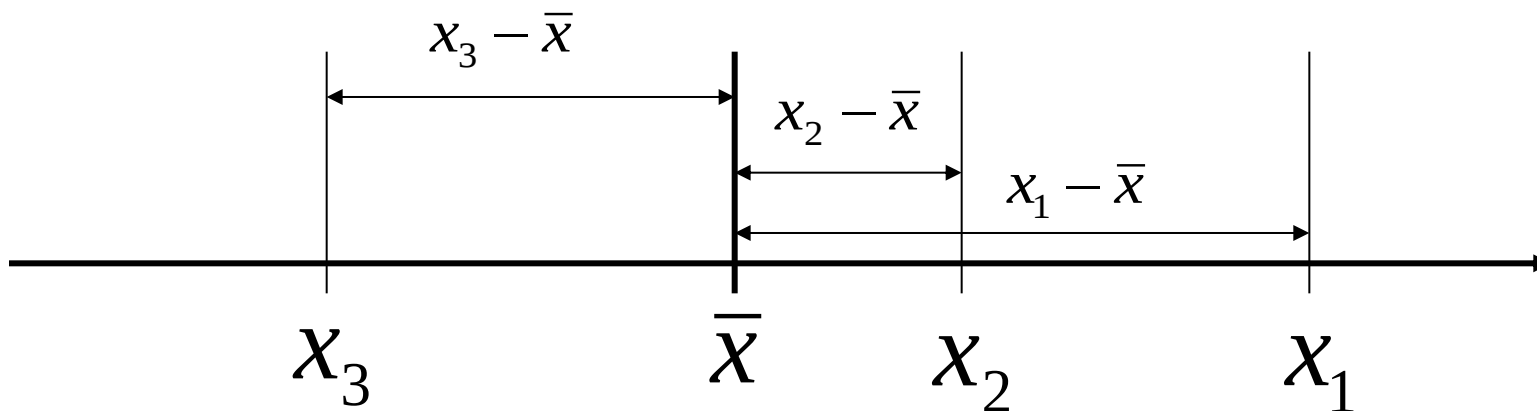
図2：範囲と四分位範囲



資料：総務省「平成27年国勢調査」人口速報集計結果

# 散らばりの尺度：平均からの偏差

- 偏差：  $x_i - \bar{x}$   
– 各  $x_i$  の中心（算術平均）からのズレ



# 散らばりの尺度：平均からの偏差

## －散らばりとの関連：

- 平均からの偏差  $x_i - \bar{x}$ 
  - － 偏差が 0 に近いものが多い。
    - 全体的なズレが小さい。
    - 散らばりが小さい。

## －性質：

- $\sum_{i=1}^N (x_i - \bar{x}) = 0$

# 散らばりの尺度：分散(1)

- 分散  $s^2$

- 分散：
$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

- 例：東京都市区町村別世帯数

- $s^2 = 11,423,734,102$  (世帯<sup>2</sup>)

- 散らばりとしての意味

- 「平均からの偏差の二乗(ズレ)」の平均

# 散らばりの尺度：分散(2)

## －長短

- 長所：理論的な性質が導きやすい。
  - －多用されるひとつの理由。
- 短所：
  - －元の測定単位の2乗の単位をもつ。
  - －外れ値の影響が大きい。

# 散らばりの尺度：標準偏差(1)

- 標準偏差  $s$

- 標準偏差：  $s = \sqrt{s^2}$

- 例：東京都市区町村別世帯数

- $s = 106,881.9$  (世帯)

- 散らばりとしての意味：

- 分散の平方根

- 分散とともに多用される。

# 散らばりの尺度：標準偏差(2)

## －長短

- ・長所：

- －元と同じ測定単位

分布の型

- －便利な性質

対称単峰      一般

- »  $\bar{x} \pm s$  に含まれる個体の割合   約2/3

- »  $\bar{x} \pm 2s$  に含まれる個体の割合   約95%    3/4以上

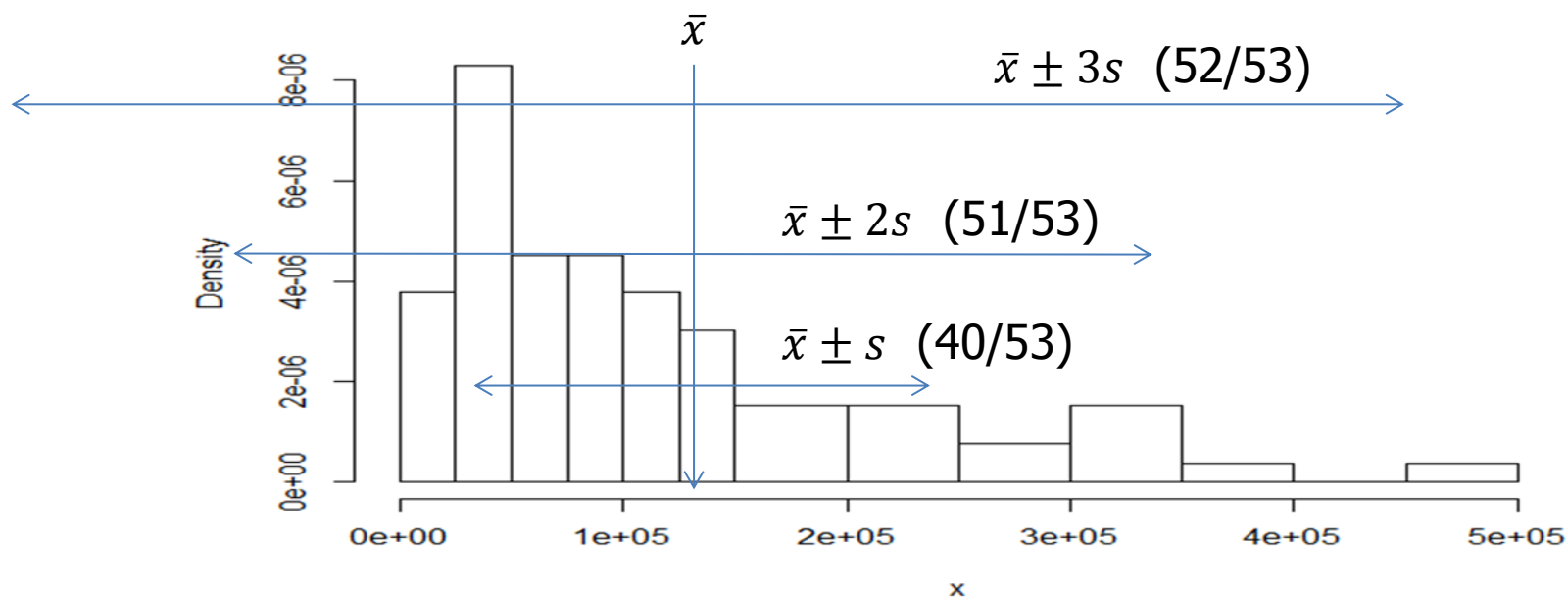
- »  $\bar{x} \pm 3s$  に含まれる個体の割合   約99%    8/9以上

- ・短所：

- －外れ値の影響を受けやすい。

# 散らばりの尺度：標準偏差(3)

図2：範囲と四分位範囲



資料：総務省「平成27年国勢調査」人口速報集計結果

# 散らばりの尺度：変動係数(1)

- 変動係数  $CV = s/\bar{x}$ 
  - 例：東京都市区町村別世帯数
    - $CV = 0.85$
  - 散らばりとしての意味：
    - 平均を1単位とした標準偏差の大きさ
  - 長短：
    - 長所：無名数（異なる単位をもつものの比較が可能）
    - 短所：外れ値の影響を受けやすい。

# 散らばりの尺度：変動係数(2)

## － 相対化する理由

- 「平均が大きくなると、散らばりが平均に比例して大きくなる」ということが多い。

－ 例：

» 年齢別にみた身長や体重

# 平均・標準偏差・分散の調整(1)

- 変数の標準化

$$Z_i = \frac{x_i - \bar{x}}{s}$$

- $z$  の算術平均=0
- $z$  の分散 = 1
- $z$  の標準偏差 = 1.

# 平均・標準偏差・分散の調整(2)

- さらに新しい変数  $t$ :

$$t_i = a + bz_i = a + b \left( \frac{x_i - \bar{x}}{s} \right)$$

—  $t$  の

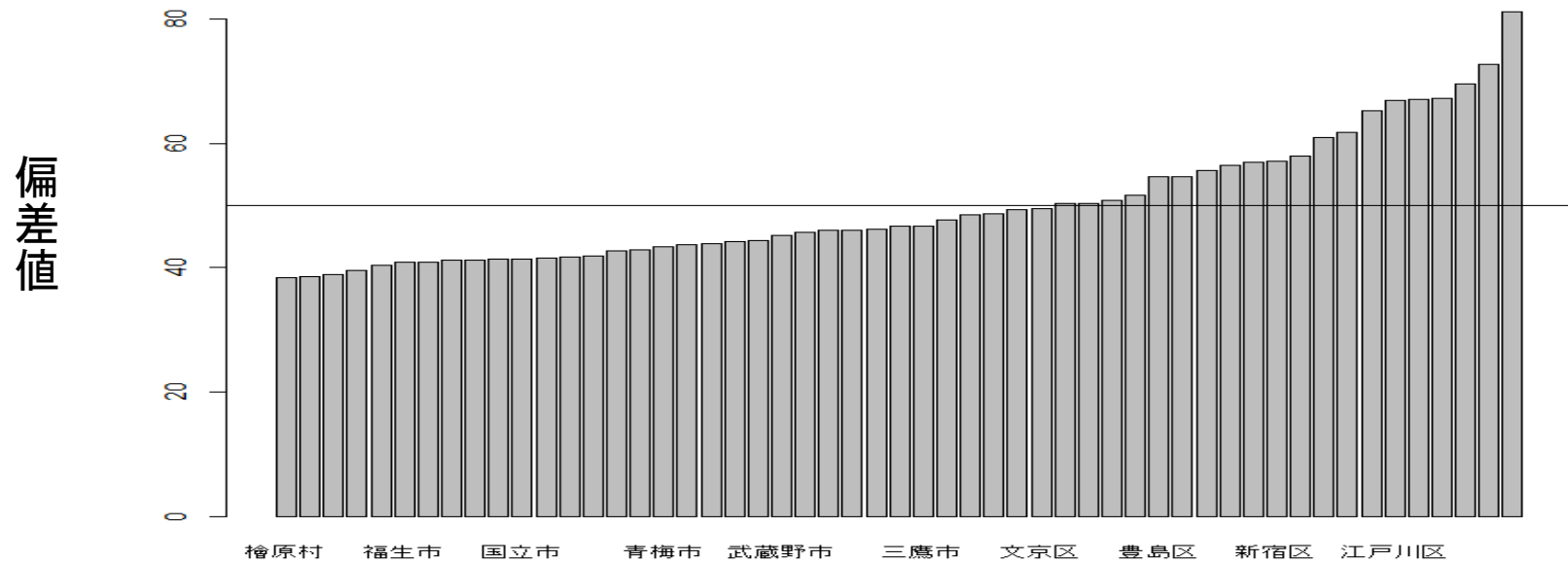
- 算術平均 =  $a$
- 分散 =  $b^2$
- 標準偏差 =  $|b|$

# 偏差値(1)

- とくに、 $a = 50, b = 10$  : 偏差値
  - 元の点の平均点  $\rightarrow$  偏差値 50点
  - 元の点の平均点 +  $s$   $\rightarrow$  偏差値 60点
  - 元の点の平均点 +  $2s$   $\rightarrow$  偏差値 70点
  - 元の点の平均点 +  $2.5s$   $\rightarrow$  偏差値 75点
  - 100点満点のイメージに合うように  
 $a = 50, b = 10$  という数字を選んだ。

# 偏差値(2)

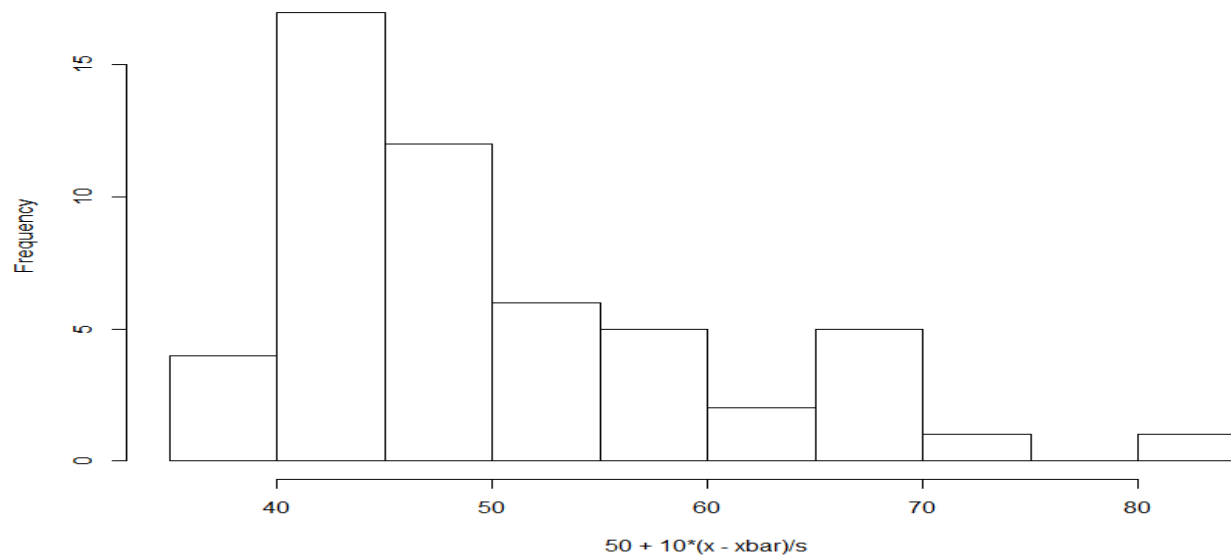
図5: 東京都市区町村別世帯数の偏差値



資料: 総務省「平成27年国勢調査」人口速報集計結果

# 偏差値(3)

図6: 偏差値のヒストグラム



資料: 総務省「平成27年国勢調査」人口速報集計結果

# 幹葉表示(1)

- 幹葉表示

- ヒストグラムの改善

- 視覚的な分布のイメージ
    - 元のデータの情報の保存

- 数値で表したヒストグラム

- 例: 東京都市区町村別世帯数

- 説明の簡単のため、1万世帯単位に変換する。

- » 3 8 13 20 12 11 13 24 21 14 37 46 14 20 31 18 18 10 29 34 31 20  
31 25 8 7 9 5 12 5 11 19 6 8 8 6 6 3 3 4 4 3 5 3 7 4 2  
3 9 1 1 0 0

# 幹葉表示(2)

図7: 東京都市区町村別世帯数の幹葉表示

0 | 0011233333444

0 | 55566677888899

1 | 011223344

1 | 889

2 | 00014

2 | 59

3 | 1114

3 | 7

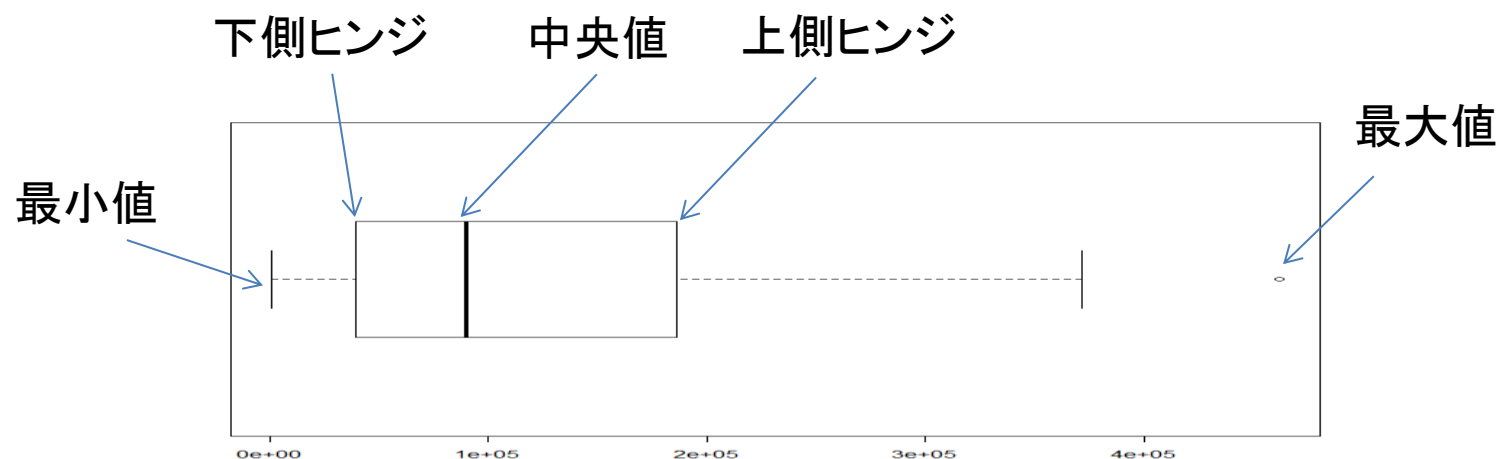
4 |

4 | 6

資料: 総務省「平成27年国勢調査」人口速報集計結果

# 箱ひげ図(1)

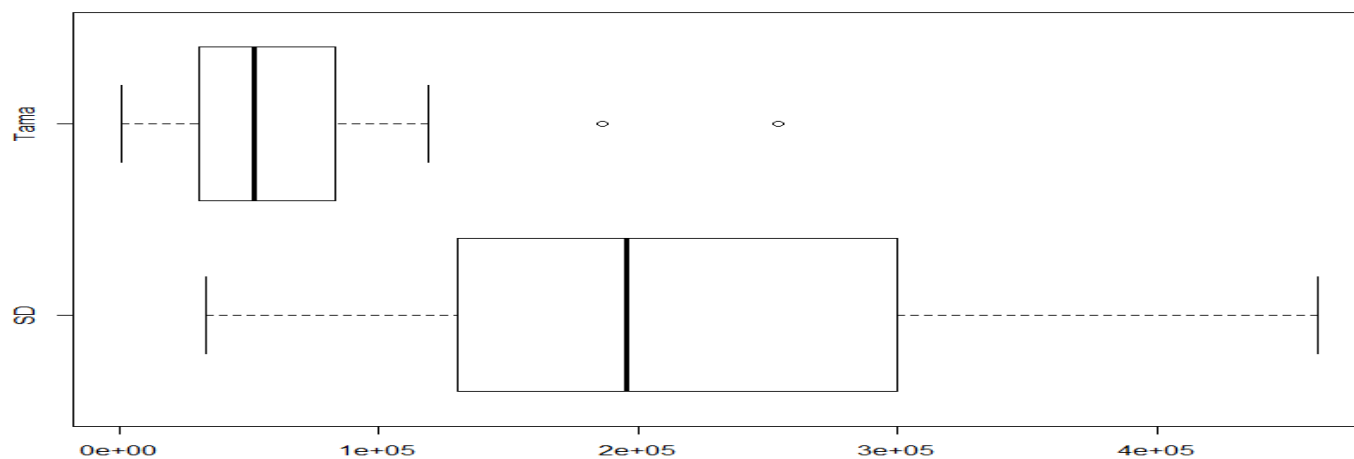
図8: 東京都市区町村別世帯数の箱ひげ図



資料: 総務省「平成27年国勢調査」人口速報集計結果

# 箱ひげ図(2)

図9: 東京都市区町村別世帯数(特別区・多摩地域別)



資料: 総務省「平成27年国勢調査」人口速報集計結果