統計学入門 第3回

早稲田大学政治経済学部
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：都道府県別基幹的農業従事者数：
    - 43.7千人

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。

$$x_1, x_2, \ldots, x_n \underset{\text{sort}}{\Longrightarrow} x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

  - *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→昇順で (*n*+1)/2 番目
      - *n* : 偶数→昇順で *n*/2 番目 と (*n*/2)+1 番目の平均

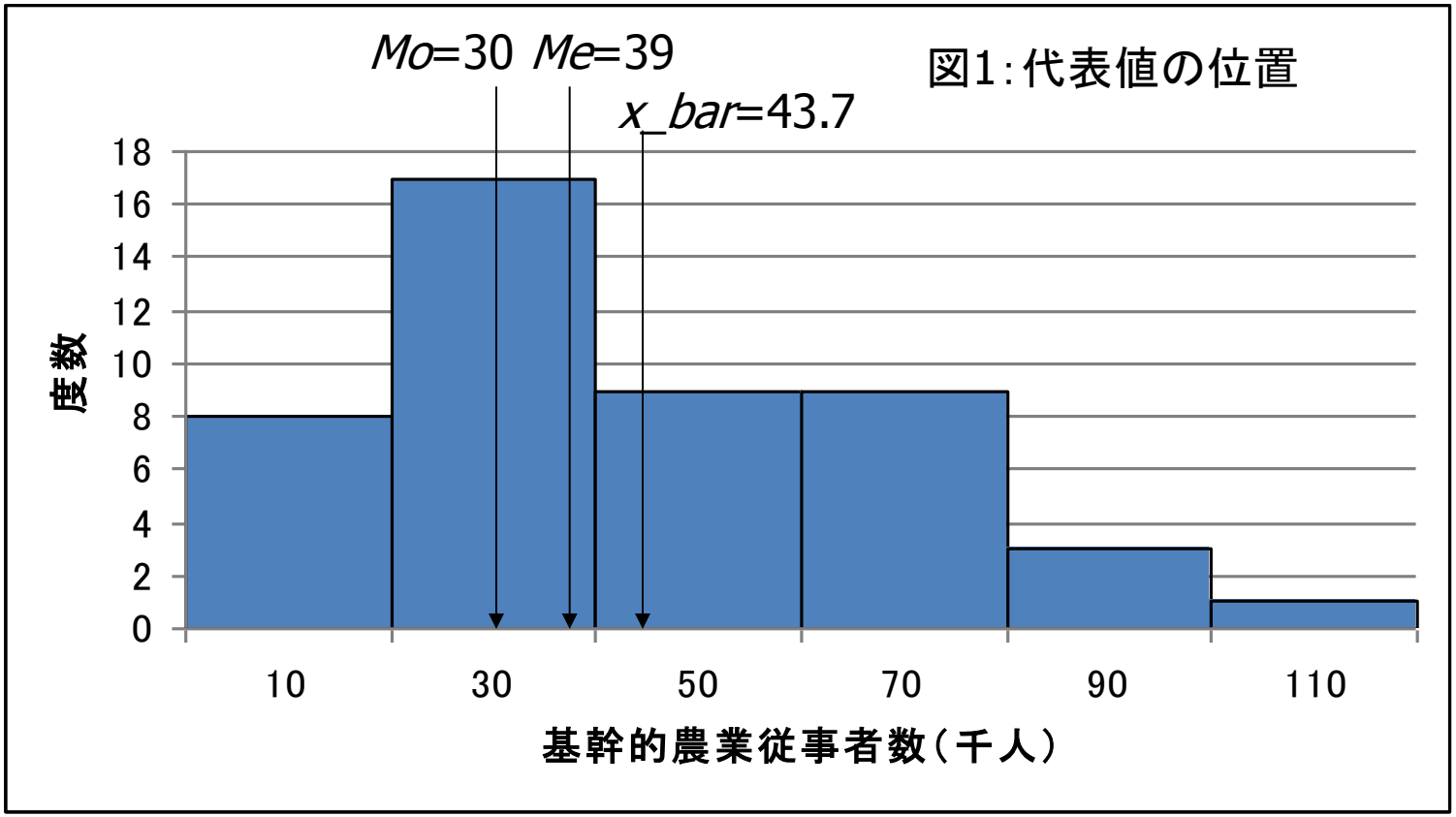
中央値(2)

- 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：都道府県別基幹的農業従事者数の *Me*
  - 昇順で24番目の値＝39千人

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：都道府県別基幹的農業従事者数の最頻値
    - 30千人（階級幅20千人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)



資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

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

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

- 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都道府県：26
  - 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 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)

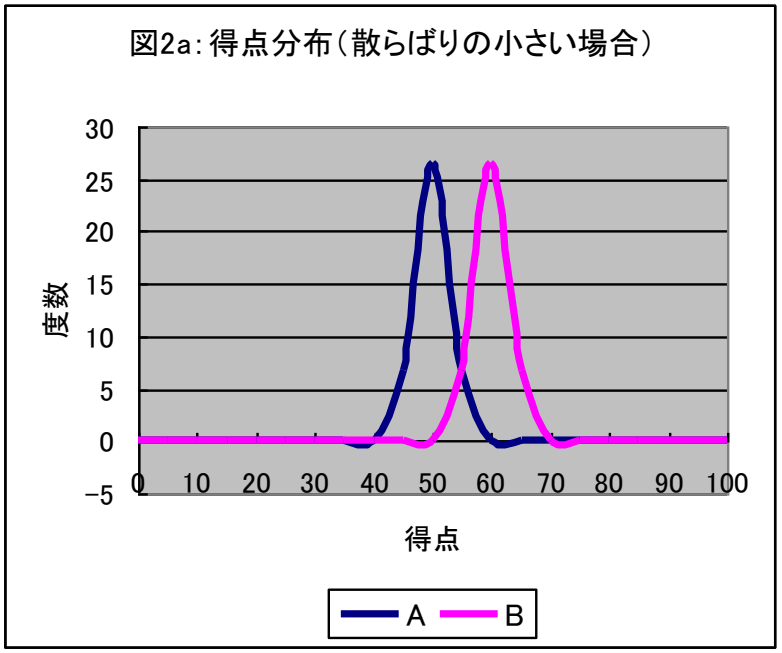


図2a：得点分布（散らばりの小さい場合）
30
25
20
15
10
5
0
−5
度数
0 10 20 30 40 50 60 70 80 90 100
得点
A
B

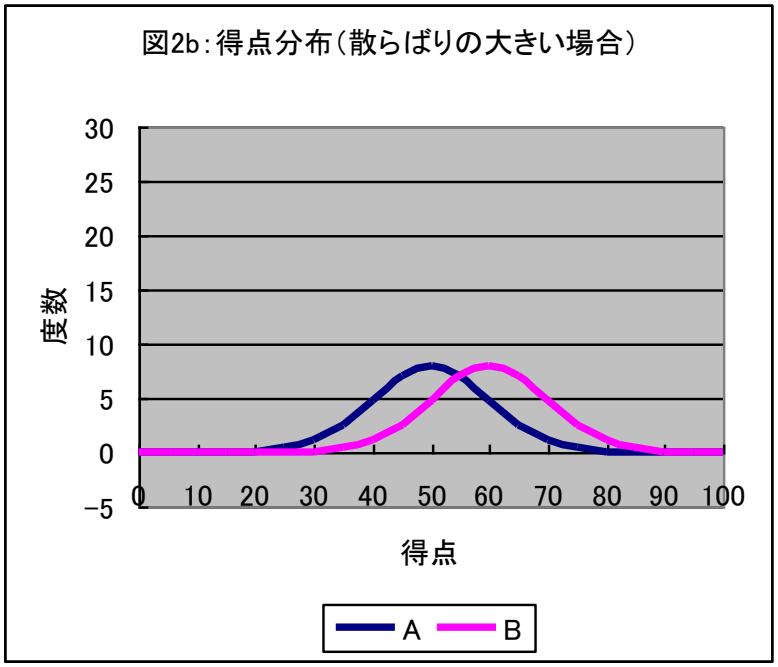


図2b：得点分布（散らばりの大きい場合）
30
25
20
15
10
5
0
−5
度数
0 10 20 30 40 50 60 70 80 90 100
得点
A
B

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：都道府県別基幹的農業従事者数
      - $R = 101 - 10 = 91$（千人）
  - 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
    - $x_{(1)} \leqq x_{(2)} \leqq \cdots\cdots\cdots\cdots\cdots\cdots \leqq x_{(n)}$

      Q_1　　Q_2　　Q_3
    - 第1四分位点
      - 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
      - 両者の平均

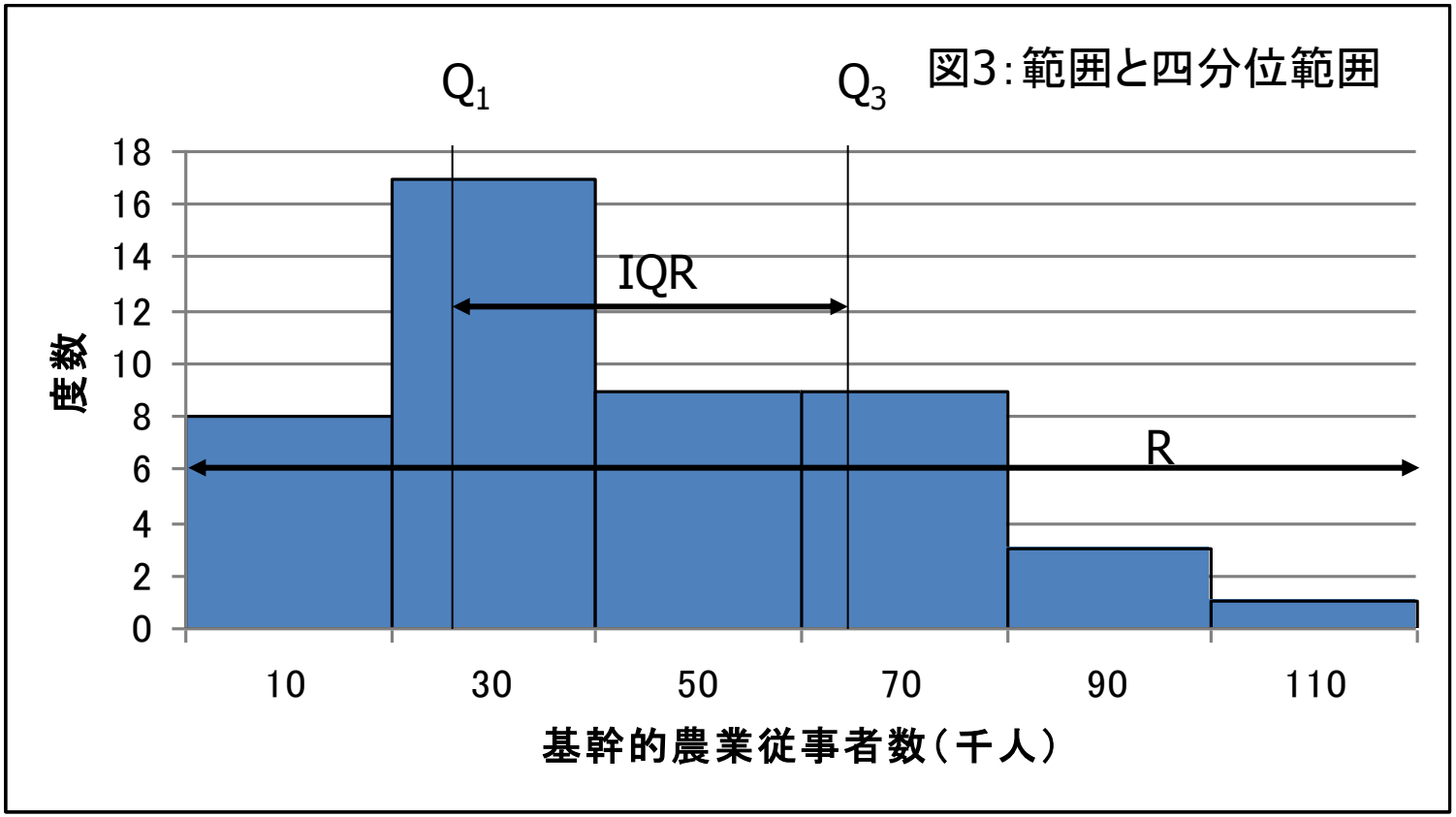
散らばりの尺度：四分位範囲(2)

- 四分位範囲 IQR $= Q_3 - Q_1$
  - 例：都道府県別基幹的農業従事者数
    - IQR = 64 - 26 = 38
      - $47 \times 1/4 = 11.75 \rightarrow Q_1 = x_{(12)} = 26$
      - $47 \times 3/4 = 35.25 \rightarrow Q_3 = x_{(36)} = 64$
- 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。

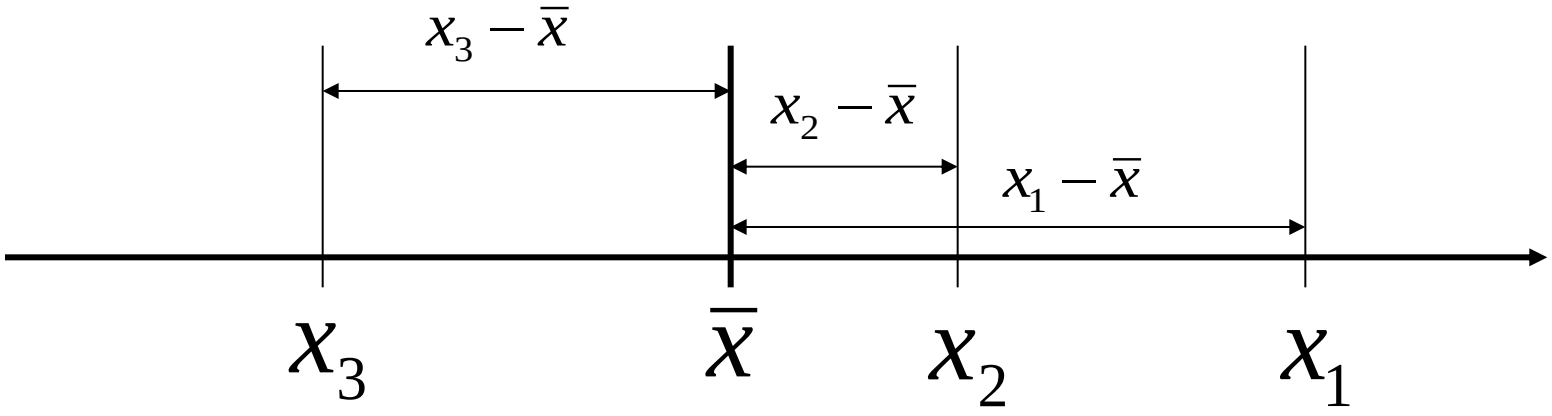
散らばりの尺度：範囲と四分位範囲



資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

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



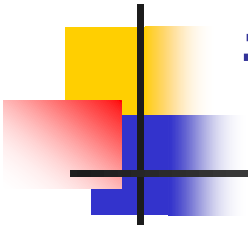
散らばりの尺度：平均からの偏差

- 散らばりとの関連　$x_i - \bar{x}$
  - 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：

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

散らばりの尺度：分散

- 分散 S^2
  - 分散： $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
    - 例：都道府県別基幹的農業従事者数
      - S^2 = 540.6（千人^2）
  - 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

- 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：都道府県別基幹的農業従事者数
      - S ＝ 23.3（千人）
  - 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。

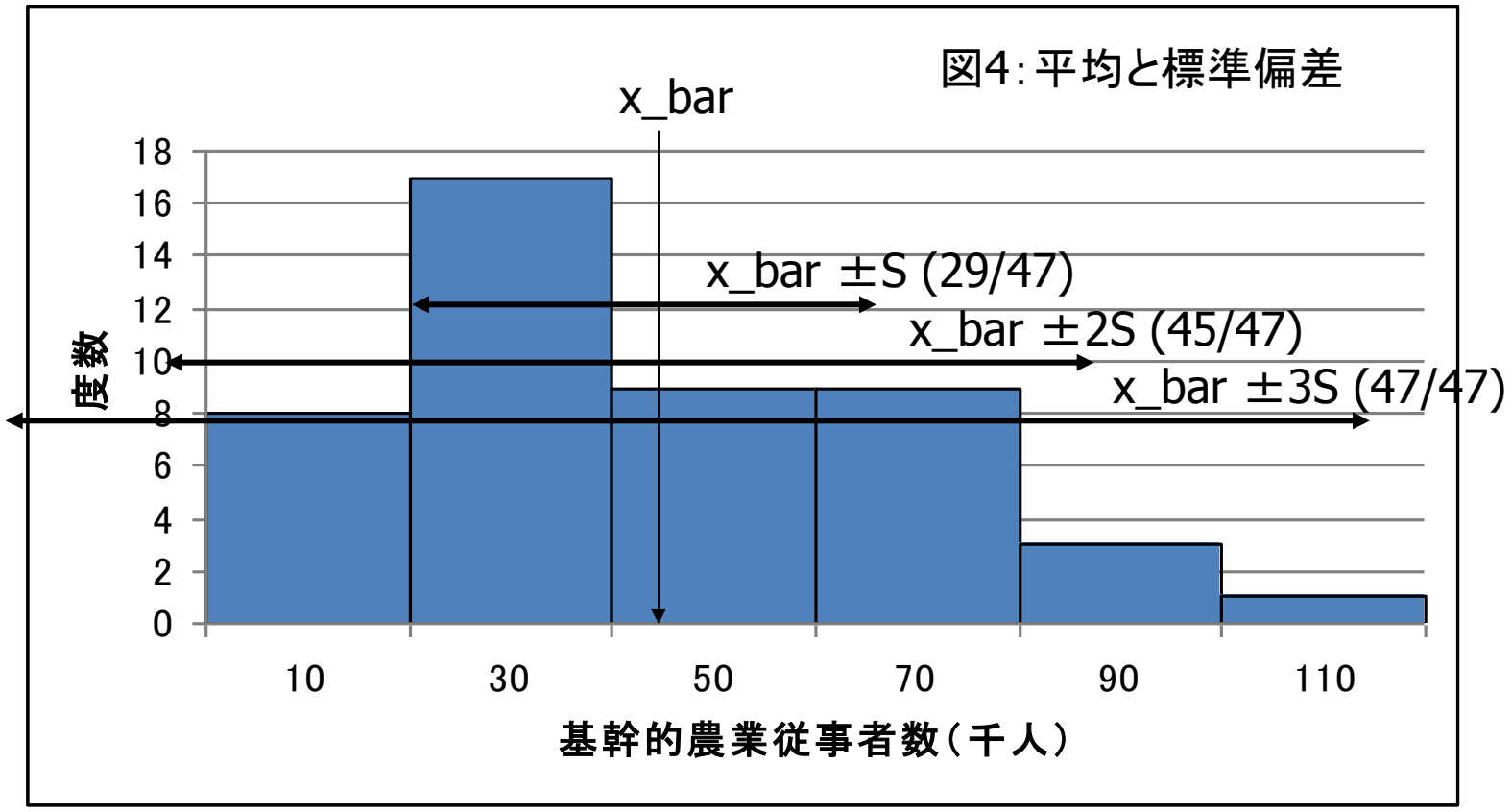
散らばりの尺度：標準偏差(2)

- 長短
  - 長所：
    - 元と同じ測定単位
    - 便利な性質

	分布の型	
	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)



資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

散らばりの尺度：変動係数(1)

- 変動係数　$CV = S/\bar{x}$
  - 例：都道府県別基幹的農業従事者数
    - CV = 0.53
  - 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:

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

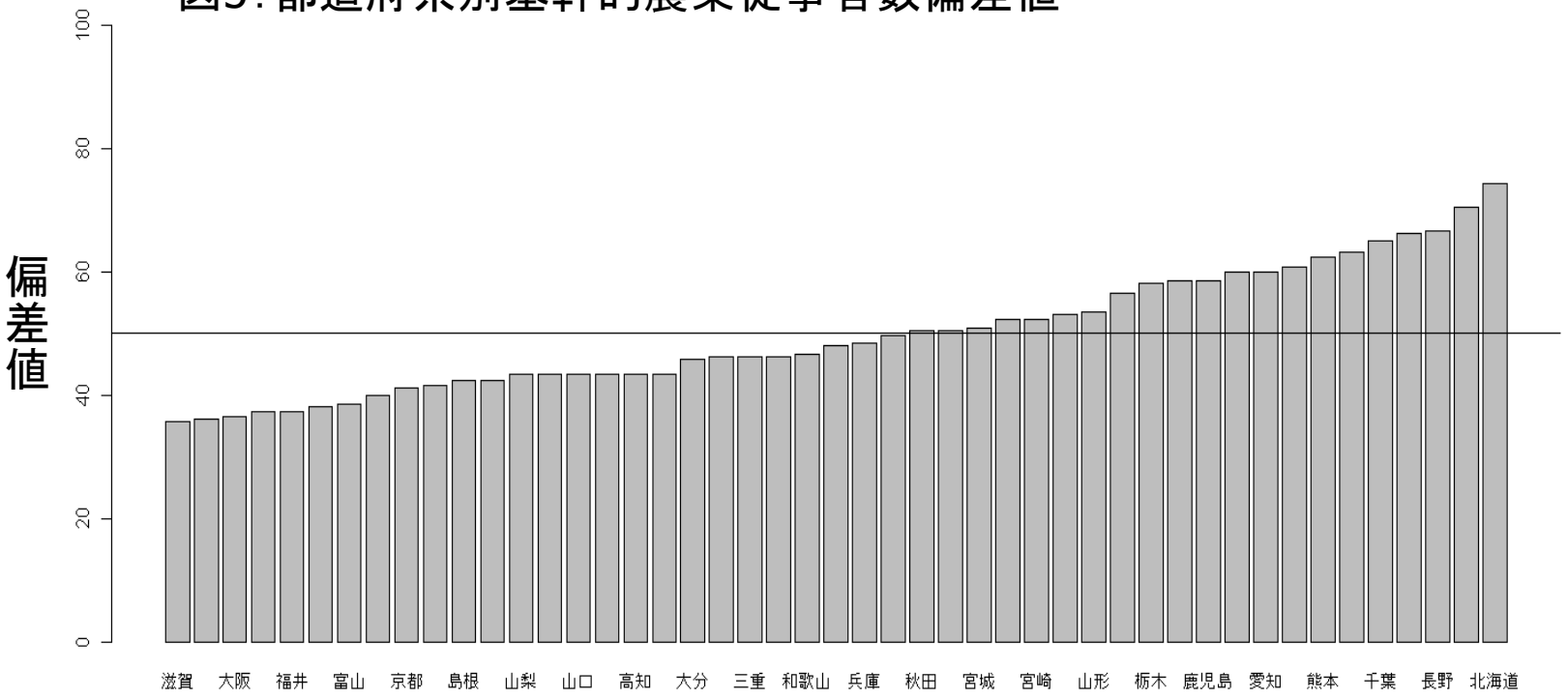
  - t の
    - 算術平均=a;
    - 分散 b^2;
    - 標準偏差=$|b|$

偏差値(1)

- とくに、$a = 50$, $b = 10$ ：偏差値
  - 元の点の平均点 → 偏差値 50点
  - 元の点の平均点＋S → 偏差値 60点
  - 元の点の平均点＋$2S$ → 偏差値 70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)

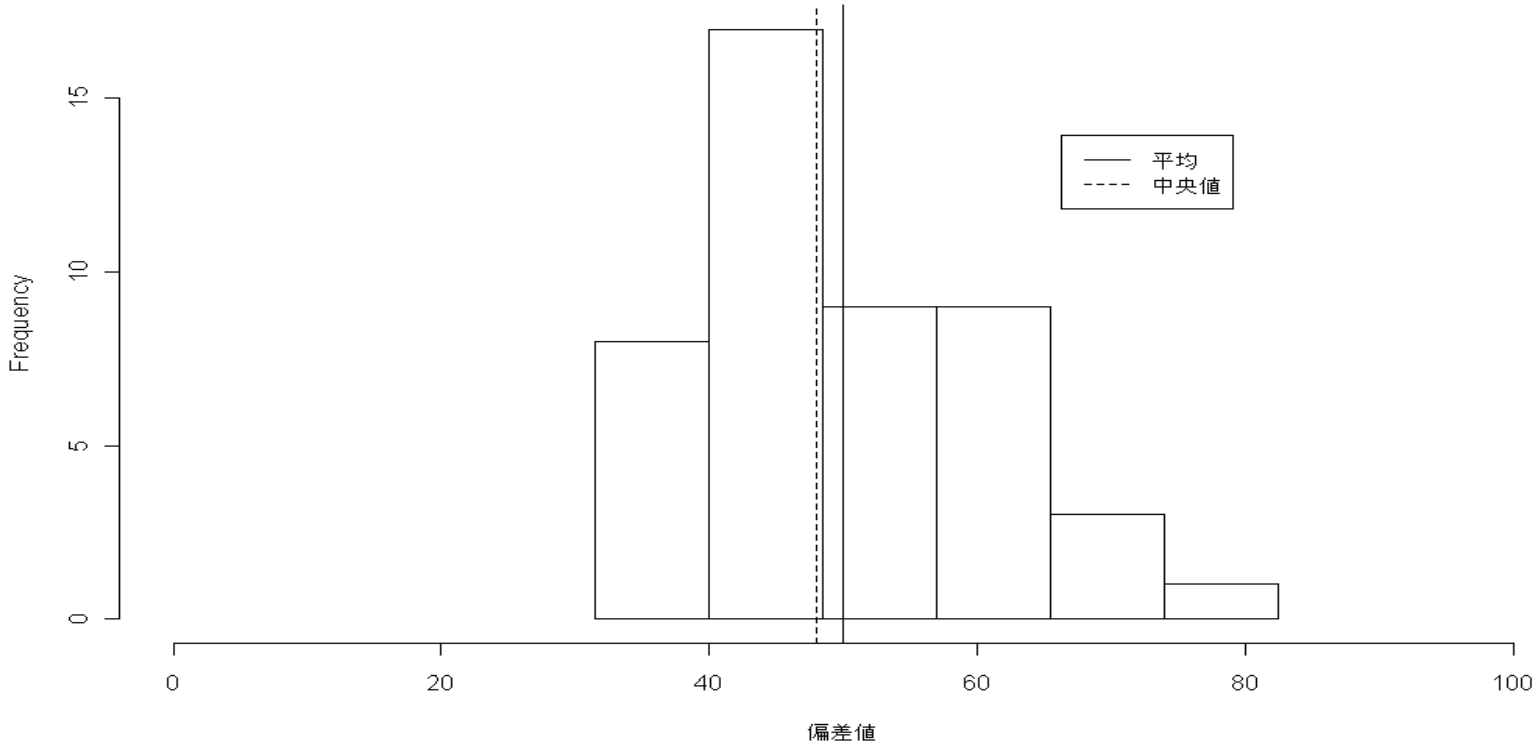
図5：都道府県別基幹的農業従事者数偏差値



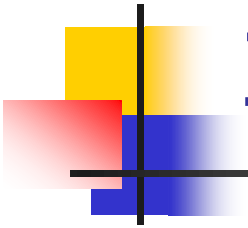
資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

偏差値(3)

図6：偏差値のヒストグラム



資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3



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

幹葉表示(2)

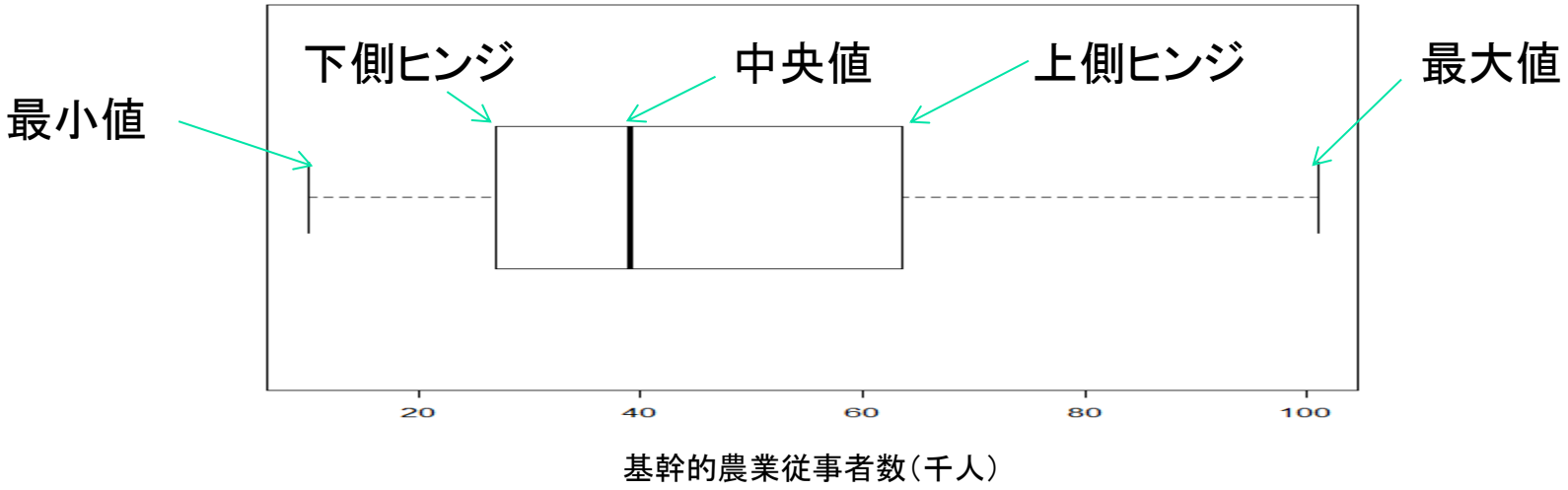
図7：都道府県別基幹的農業従事者数の幹葉表示

```
 1 | 0124467
 2 | 03466888888
 3 | 455569
 4 | 0355699
 5 | 129
 6 | 344779
 7 | 359
 8 | 23
 9 | 2
10 | 1
```

資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

箱ひげ図(1)

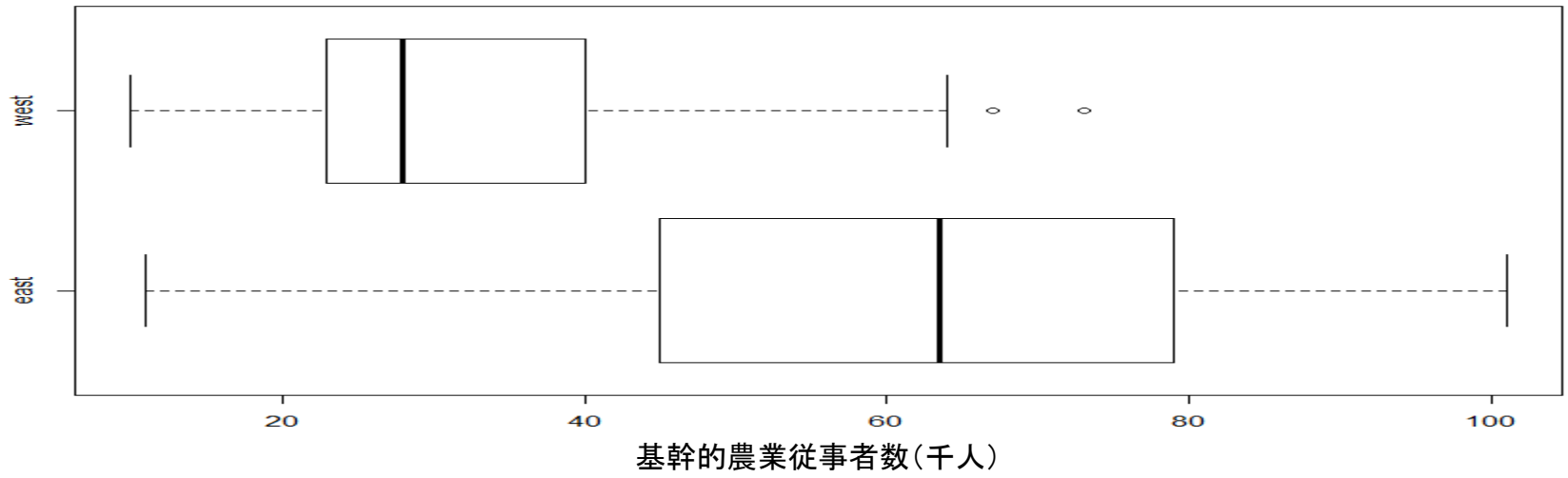
図8：都道府県別基幹的農業従事者数の箱ひげ図



資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3

箱ひげ図(2)

図9：都道府県別基幹的農業従事者数の箱ひげ図（東日本・西日本）



注：新潟・長野・静岡までを東日本とした。

資料：総務省統計研修所(2012)『第61回日本統計年鑑』表7-3