教養科目／B 自然の摂理の探求

2026 年度 統計学（集中 1 期）

第 2 回

前回の内容

- 度数分布表，ヒストグラム
  - 離散型
    - データ: 8, 5, 5, 9, 5, 8, 9, 9, 7, 9

階級値 (日数)	度数 (種の個数)
5	3
6	0
7	1
8	2
9	4
計	10

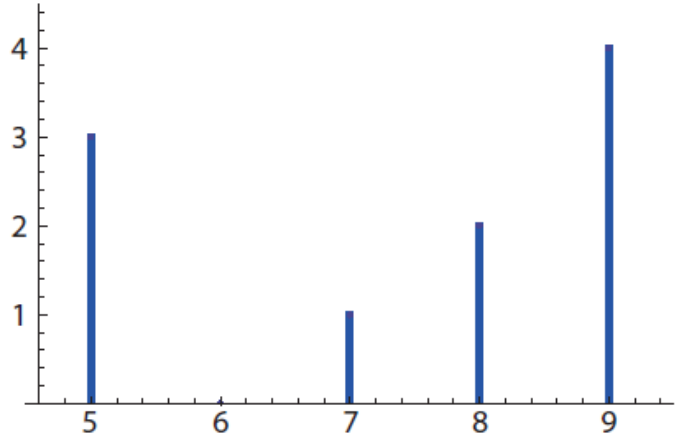


図 1.1: 例 1.1 のヒストグラム

前回の内容

- 度数分布表，ヒストグラム
  - 連続型
    - 階級の個数は少な過ぎても，多過ぎてもよくない．
    - スタージェスの方法により，階級の個数，階級の幅を決める．

表 1.7　例 1.5 の度数分布表

階級	級中央値	度数	相対度数
6.0 ～ 9.5	7.75	1	0.02
9.5 ～ 13.0	11.25	7	0.15
13.0 ～ 16.5	14.75	17	0.36
16.5 ～ 20.0	18.25	13	0.28
20.0 ～ 23.5	21.75	8	0.17
23.5 ～ 27.0	25.25	0	0.00
27.0 ～ 30.5	28.75	1	0.02
計	—	47	1.00

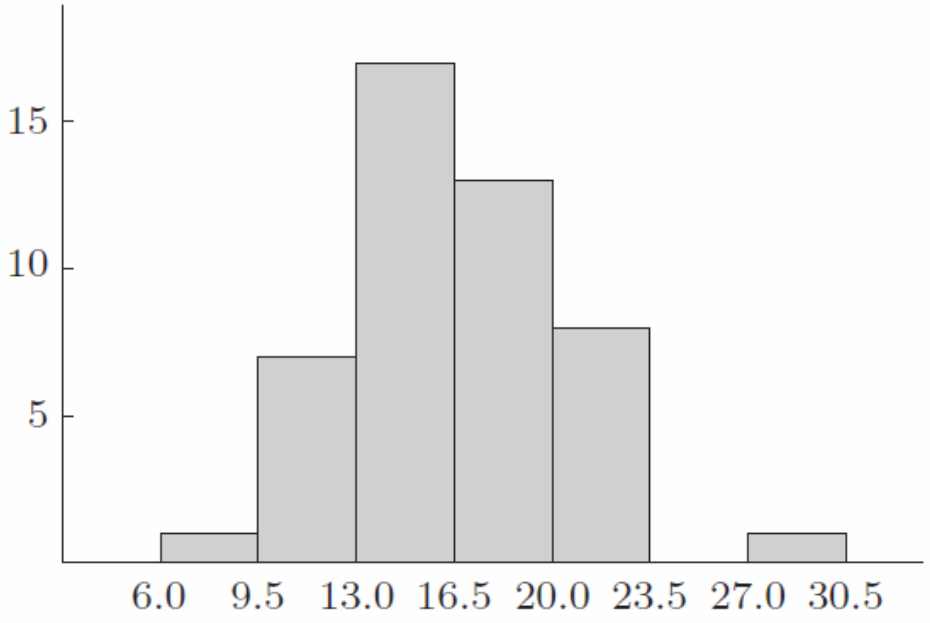


図 1.5　例 1.5 のヒストグラム

今回扱う内容

量的データ（当面，データと書けば量的データを意味するものとします）が得られたとき，それらがもっている特徴を**代表値**と呼ばれる１つの値を用いて表すことについて考えてみます.

・代表値

中心を表す代表値 ..... 標本平均, 中央値

ばらつきを表す代表値 ..... 標本分散, 標本標準偏差, 不偏分散

1．4　代表値

- 平均最低気温：各日ごとの最低気温の平均.
- ベルリンの 9 月の平均最低気温:10.6 ℃

表 1.8　東京の平均最低気温（℃）

1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1.1	2.4	5.1	10.5	15.1	18.9	22.5	24.2	20.7	15.0	9.5	4.6

平均最低気温のように最低気温についてのデータを１つの値で表すことがしばしばある．このような値を**代表値**という．

9 月の東京の平均最低気温よりベルリンの平均最低気温が低いことは必ずしも常にベルリンが東京より寒いことを意味しない．また，9 月にベルリンで 10.6 ℃を下回る日も当然あり得る．

1．4．1．中心的位置を表す代表値

0 と 2 の中心的位置を表す値は？

図 1.6　データ (1.1) のヒストグラム

通常，0 と 2 の真ん中の値である

$$\frac{0+2}{2}=1.$$

1．4．1．中心的位置を表す代表値

0, 2, 7 の中心的位置を表す値は？

図 1.7　データ (1.2) のヒストグラム

●1 つめの考え方

$$\frac{0+2+7}{3}=3.$$

■2 つめの考え方

3 つのデータ 0 と 2 と 7 の真ん中の値の 2.

1．4．1．中心的位置を表す代表値

0, 2, 7, 11 の中心的位置を表す値は？

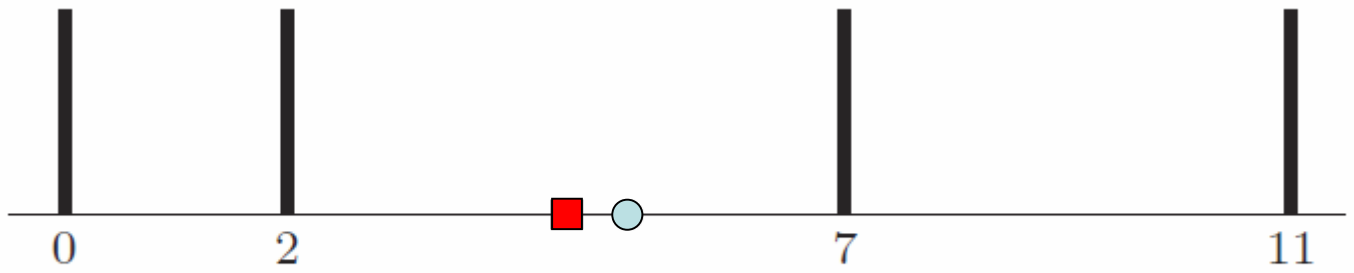


図 1.8　データ (1.3) のヒストグラム

○1 つめの考え方

$$\frac{0+2+7+11}{4}=5.$$

■2 つめの考え方: 候補としては 2 か 7 だが,
どちらもちょうど真ん中ではない.
そこで，2 と 7 の真ん中の値

$$\frac{2+7}{2}=4.5$$

・中心的位置を表す代表値として，２つの考え方がある

標本平均

前者を**標本平均**（または単に**平均**）という．

- **標本平均**

$$\text{標本平均} = \frac{\text{データの値の合計}}{\text{データの個数}}.$$

n 個のデータ

$$x_1, \quad x_2, \quad \cdots, \quad x_n$$

に対して，標本平均 $\bar{x}$（エックスバーと呼びます）は

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

で与えられる．

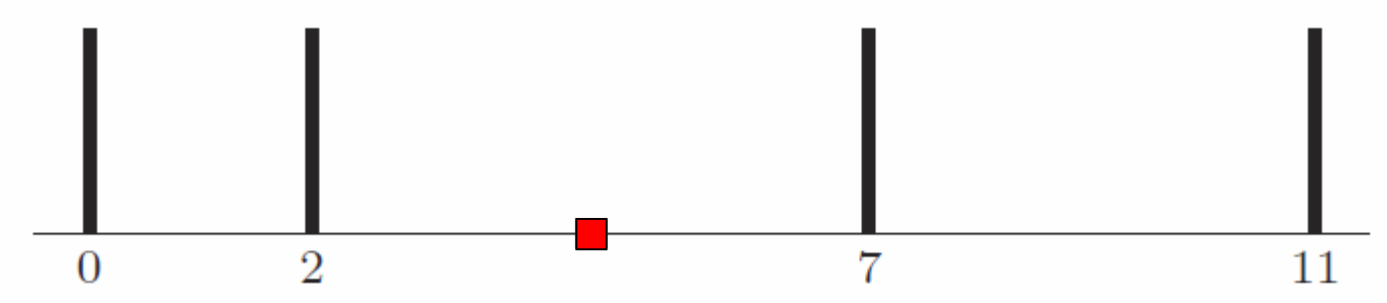
中央値

後者を**中央値またはメジアン**という.

■**中央値**

データの真ん中の位置にある値 (データの個数が奇数の場合)

$$\frac{\text{データの真ん中の値の２つの候補の和}}{2}$$ (データの個数が偶数の場合)



例 1.1の標本平均および中央値

種 10 粒を蒔いて発芽するまでの日数:

$$8,\quad 5,\quad 5,\quad 9,\quad 5,\quad 8,\quad 9,\quad 9,\quad 7,\quad 9$$

○ 標本平均

$$\text{標本平均} = \frac{8+5+\cdots+9}{10} = 7.4$$

■ 中央値

データを大きさの順に並べ替える：

$$5,\quad 5,\quad 5,\quad 7,\quad \textcircled{8},\quad \textcircled{8},\quad 9,\quad 9,\quad 9,\quad 9.$$

データの個数は 10 であり，これは偶数であるのでデータの真ん中の位置にある値の候補は 2 つある.

5 番目に小さい値 8 と 6 番目に小さい値 8.

中央値は, 中央値 $= \frac{8+8}{2} = 8.$

例 1.5の平均と中央値

例 1.5 各都道府県の 2010 年の人口 10 万人あたりの結核罹患者数

9 11 11 11 12 12 12 12 13 13 13 14 14 14 14 14 15 15 15 15
16 16 16 16 16 17 17 17 17 17 18 18 18 19 19 19 19 19 20 21
21 21 21 23 23 23 30

- **標本平均**

$$\text{標本平均} = \frac{12 + 14 + \cdots + 19}{47} = \frac{776}{47} \fallingdotseq 16.5.$$

- **中央値**

データの個数は $47 (= 24 \times 2 - 1)$ であり，これは奇数である．
つまり，小さい方から 24 番目の値 16 が中央値である．

9 11 11 11 12 12 12 12 13 13 13 14 14 14 14
14 15 15 15 15 16 16 16 16 16 17 17 17 17 17
18 18 18 19 19 19 19 19 20 21 21 21 21 23 23
23 30

中央値と平均値

・標本平均よりむしろ中央値の方が中心的位置を如実に表している例.

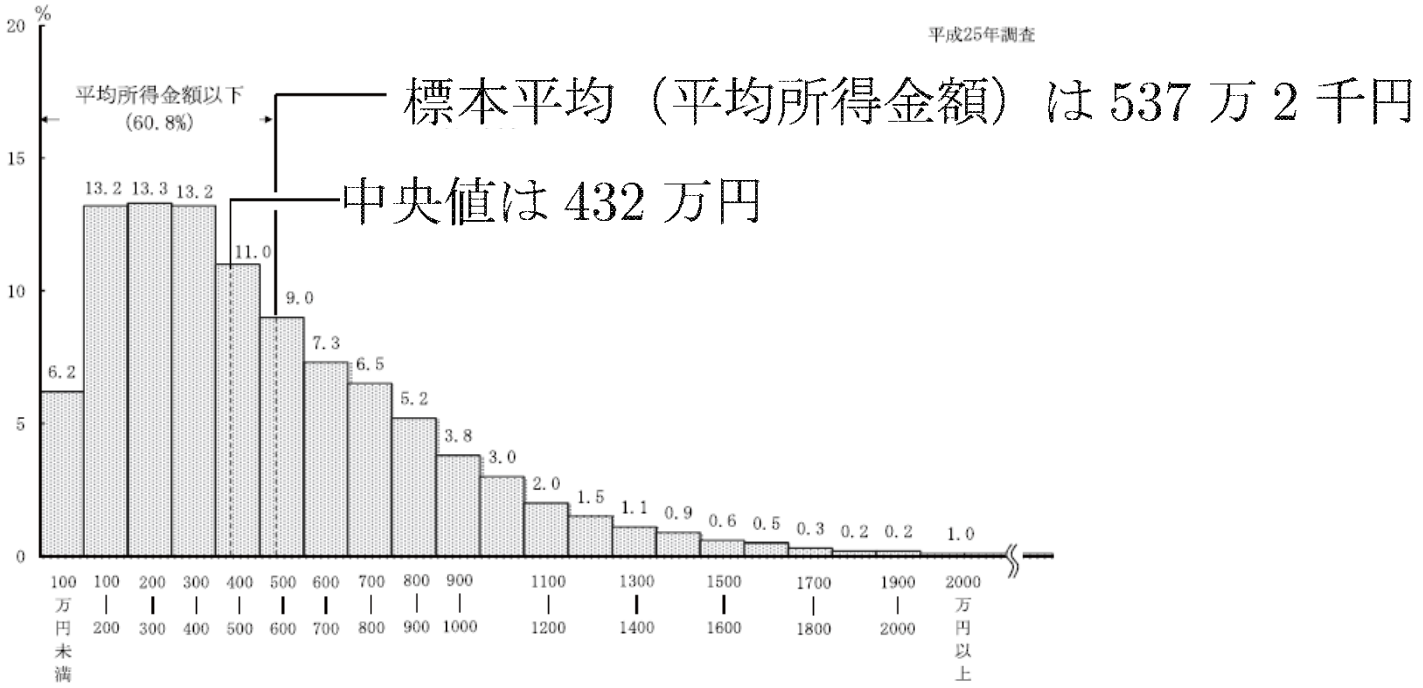


図 1.9: 所得金額の相対度数のヒストグラム

・平均所得以下は, 60%.

・所得金額のような左に歪んでいるデータについては，標本平均以下の割合は 50% より多くなることに注意.

例 1.8 シンプソンのパラドックス

ある会社の面接は 5 人ずつ 2 組に分かれて行われた.

- 1 組目：男性 4 人，女性 1 人
- 2 組目：男性 1 人，女性 4 人

面接官の付けた点数を男女の平均でまとめたもの (表 1.10).

面接官 K

差別はしていません

男女間で差別をしていないか？

別の社員 A

表 1.9

	1 組目	2 組目
男性	4	1
女性	1	4

表 1.10

	1 組目	2 組目
男性	80	90
女性	70	85

シンプソンのパラドックス

面接官の主張は, 本当か？

・男性全体の平均

$$\text{男性全体の平均} = \frac{80\times4+90\times1}{5} = \frac{410}{5} = 82$$

・女性全体の平均

$$\text{女性全体の平均} = \frac{70\times1+85\times4}{5} = \frac{410}{5} = 82$$

つまり全体での男女間で平均に違いはない.

表 1.9

	1 組目	2 組目
男性	4	1
女性	1	4

表 1.10

	1 組目	2 組目
男性	80	90
女性	70	85

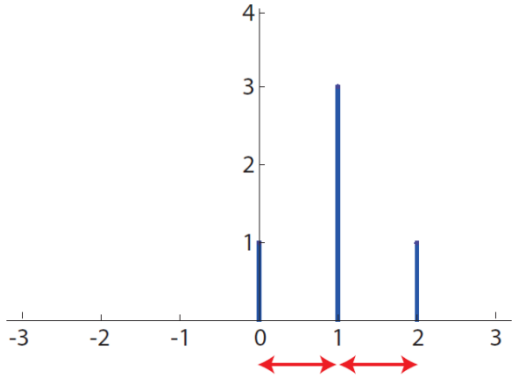
1．4．2．ばらつきを表す代表値

2 組のデータ

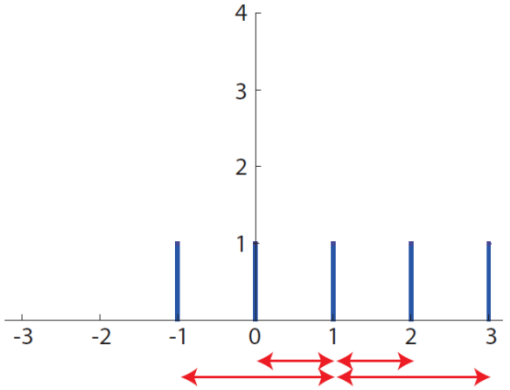
$$\mathrm{A}\ :\ 0,1,1,1,2 \qquad \mathrm{B}\ :\ -1,0,1,2,3$$

・**中央値・標本平均で比較**：A, B いずれも標本平均，中央値ともに 1.

・**ヒストグラムの比較:**



Aのヒストグラム



B のヒストグラム

・A は 1 に集中し, B は平らである.

・B の 1 の周りのばらつきの度合いが A より大きい.

1. 4. 2. ばらつきを表す代表値

ばらつきの違いを数値を用いて表すことを考える.

・**偏差**

A ： $0, 1, 1, 1, 2$　　B ： $-1, 0, 1, 2, 3$

の各値から, 標本平均 1 を引いた

Aの偏差 ： $-1, 0, 0, 0, 1$　　B の偏差 ： $-2, -1, 0, 1, 2$

を考える. このような値を**偏差**という. 一般には

$$\text{偏差} = \text{データの値} - \text{標本平均}$$

で与えられる.

・**Note**：A や B における偏差の総和は 0 になっている.

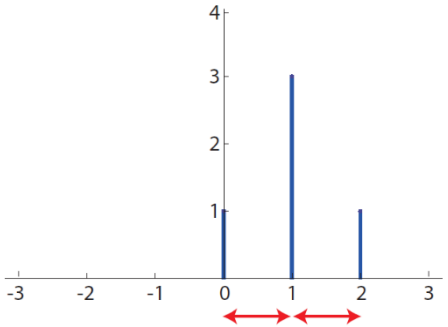
公式 1.1

偏差について，次が成り立ちます.

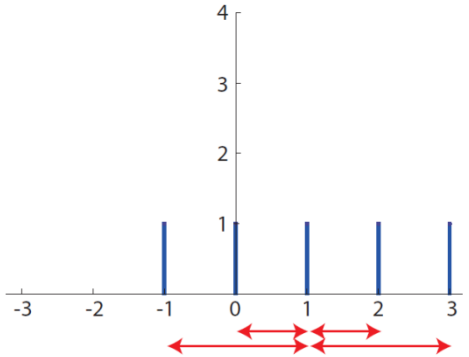
$$\text{偏差の総和} = 0.$$

・**偏差の平均はばらつきの度合いを示す代表値には使えない.**

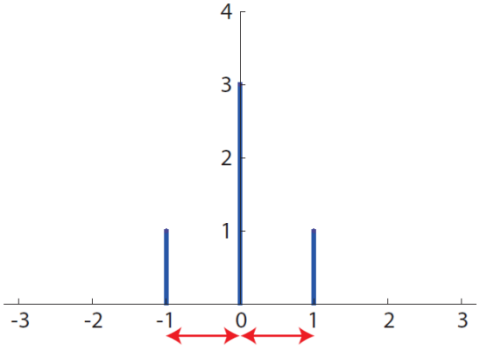
1．4．2．ばらつきを表す代表値



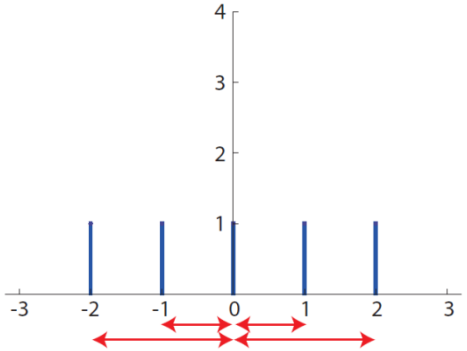
A のヒストグラム



B のヒストグラム



A の偏差のヒストグラム



B の偏差のヒストグラム

・元のデータとその偏差はばらつきの度合いは同じ．

Aの偏差 ： $-1, 0, 0, 0, 1$

Bの偏差 ： $-2, -1, 0, 1, 2$

偏差の平均 → **偏差の 2 乗の平均**

そこで, 偏差の 2 乗和を計算する.

Aの偏差の 2 乗 ： $1, 0, 0, 0, 1$

Bの偏差の 2 乗 ： $4, 1, 0, 1, 4$

さらに, 偏差の 2 乗和の平均を計算する.

Aの偏差の 2 乗の平均 ： $\frac{1+0+0+0+1}{5} = 0.4$

Bの偏差の 2 乗の平均 ： $\frac{4+1+0+1+4}{5} = 2$

このような値を**標本分散**といい, ばらつきを表す代表値として知られている.

1．4．2．ばらつきを表す代表値

・標本分散

$$\text{標本分散} = \frac{\text{偏差の 2 乗の総和}}{\text{データの個数}}$$

・標本分散 (数式版)

n 個のデータ ： $x_1,\quad x_2,\quad \cdots,\quad x_n$ の標本分散 s_x^2 は

$$s_x^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

となる.

・標本分散が大きければ大きいほど標本平均の周りのばらつきの度合いが大きい.

1．4．2．ばらつきを表す代表値

- **標本標準偏差**

単位をデータの単位に揃えるために標本分散の正の平方根をとったものを**標本標準偏差**という．

標本標準偏差 s_x は

$$\text{標本標準偏差} = \sqrt{\text{標本分散}}$$

で与えられる．

- **標本標準偏差 (数式版)**

n 個のデータ ： $x_1,\quad x_2,\quad \cdots,\quad x_n$ の標本分散 s_x^2 を用いると $s_x = \sqrt{s_x^2}$ と表せる．

1．4．2．ばらつきを表す代表値

・不偏分散

不偏分散は

$$\text{不偏分散} = \frac{\text{偏差の２乗の総和}}{(\text{データの個数}) - 1}$$

で与えられる．

・不偏分散 (数式版)

n 個のデータ ： $x_1,\quad x_2,\quad \cdots,\quad x_n$ の不偏分散 u_x^2 は

$$u_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

となる．

1. 4. 2. ばらつきを表す代表値

例 1.5 各都道府県の 2010 年の人口 10 万人あたりの結核罹患者数

9 11 11 11 12 12 12 12 13 13 13 14 14 14 14 14 15 15 15 15
16 16 16 16 16 17 17 17 17 17 18 18 18 19 19 19 19 19 20 21
21 21 21 23 23 23 30

- **標本平均**

$$\text{標本平均} = \frac{12 + 14 + \cdots + 19}{47} = \frac{776}{47} \fallingdotseq 16.5$$

- **標本分散**

$$\begin{aligned} \text{標本分散} &= \frac{1}{47}\left\{\left(12 - \frac{776}{47}\right)^2 + \left(14 - \frac{776}{47}\right)^2 + \cdots + \left(19 - \frac{776}{47}\right)^2\right\} \\ &= \frac{35144}{2209} \fallingdotseq 15.9 \end{aligned}$$

- **標本標準偏差**

$$\text{標本標準偏差} = \sqrt{\frac{35144}{2209}} \fallingdotseq 3.99$$

まとめ

・代表値

中心を表す代表値 ..... 標本平均, 中央値

ばらつきを表す代表値 ..... 標本分散, 標本標準偏差, 不偏分散