大学生のための統計学入門

―テーマ1. データの要約―

京都大学大学院医学研究科

臨床統計学/臨床統計家育成コース　田中司朗

テーマ1. データの要約

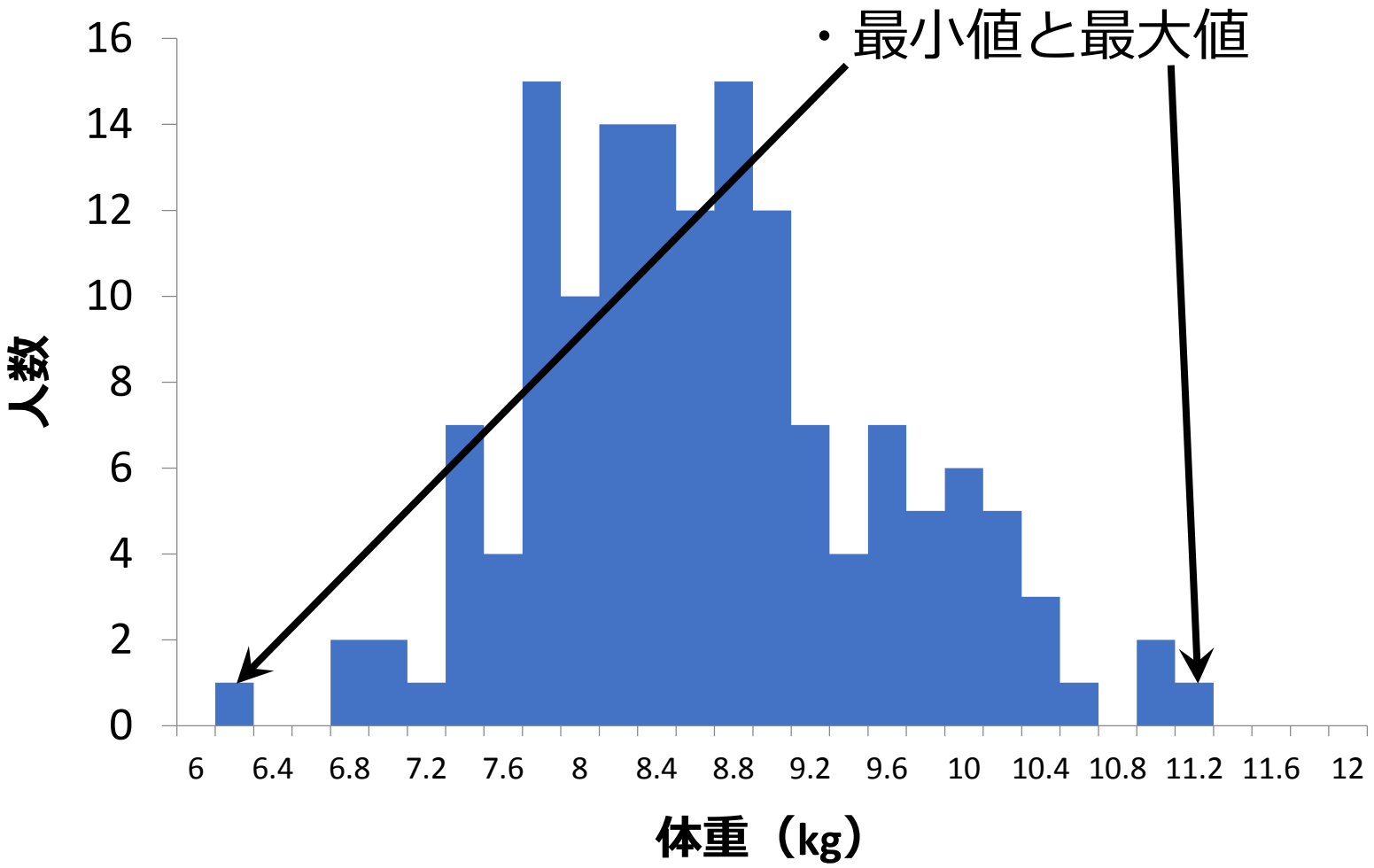
- 数値データの例
- 分布の位置の指標
- **バラツキの指標**
  - **分散**
  - **標準偏差**
  - **パーセント点**
  - **四分位偏差**
- 偏差値と標準化

バラツキの指標

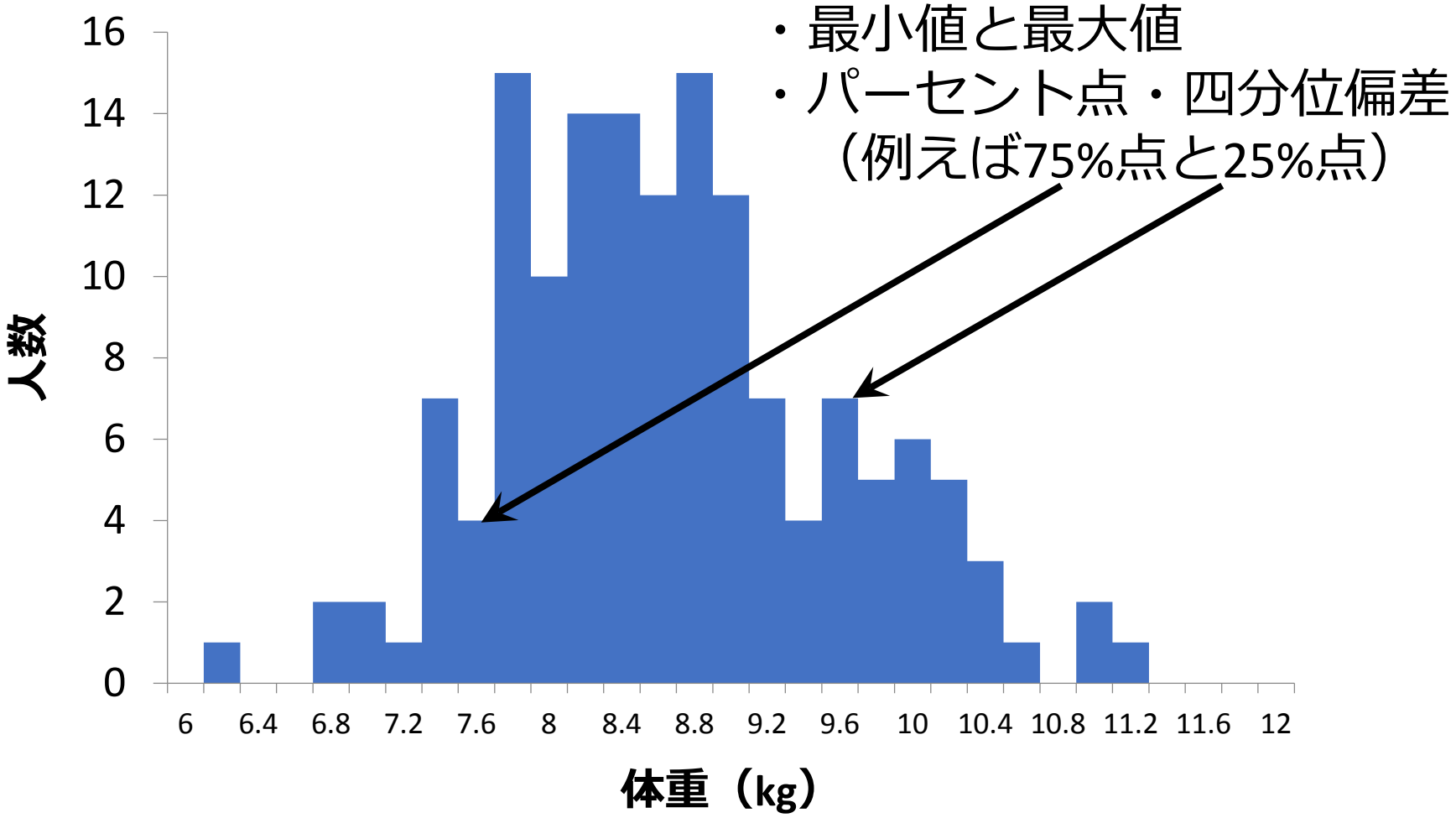
- 数値データには必ずバラツキがある
- 平均や中央値だけではなく，バラツキもきちんと考えておかなければならない

犯行現場に残された二つの足跡の最大長	容疑者の家にあったブーツの足跡の最大長（10回測定）	
25.52 cm	24.84 cm	24.73 cm
25.33 cm	24.84 cm	24.64 cm
	23.89 cm	24.94 cm
	24.86 cm	24.92 cm
	24.27 cm	24.65 cm

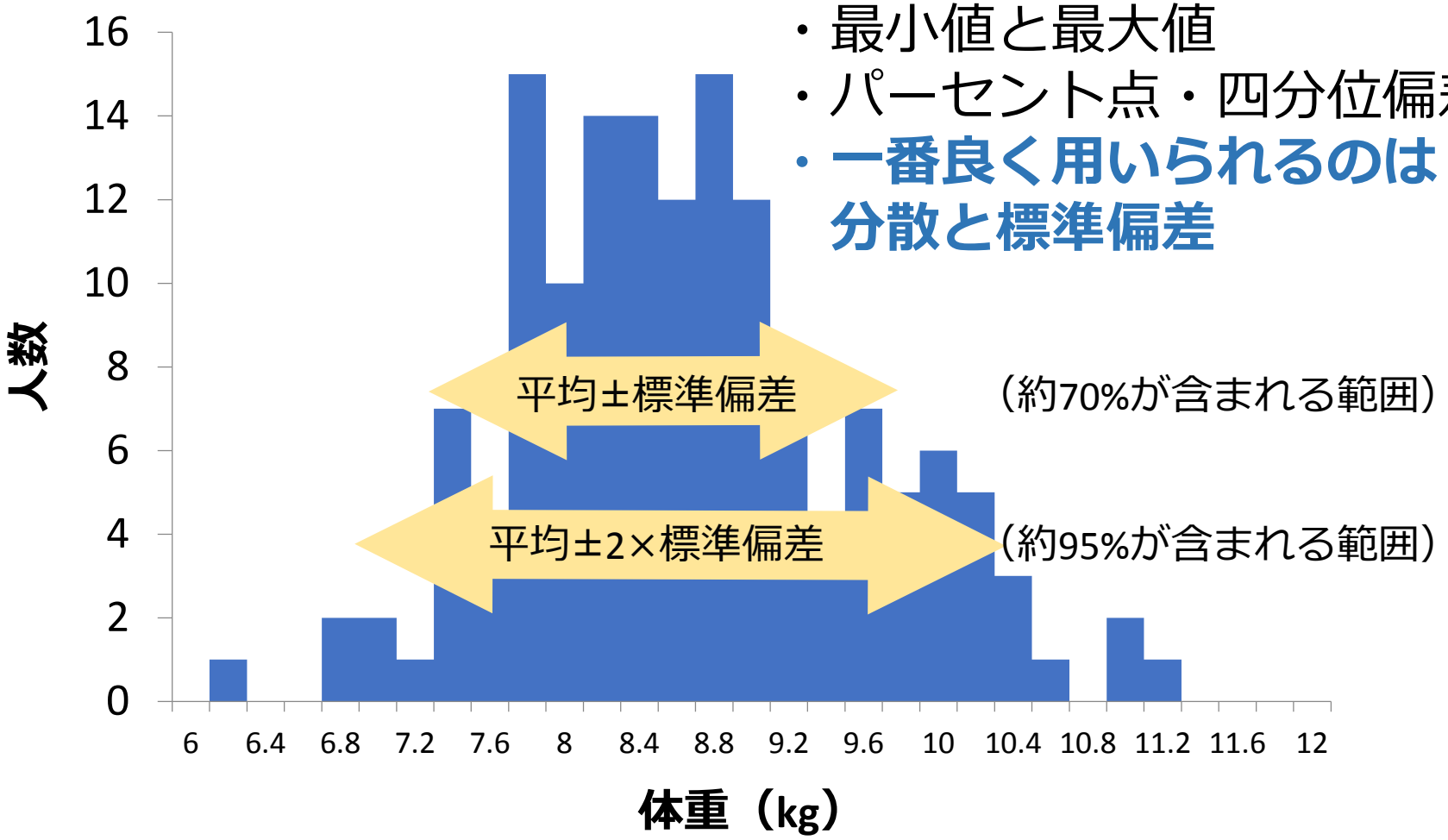
バラツキの指標


・最小値と最大値
16
14
12
10
8
6
4
2
0
人数
6
6.4
6.8
7.2
7.6
8
8.4
8.8
9.2
9.6
10
10.4
10.8
11.2
11.6
12
体重（kg）

バラツキの指標


・最小値と最大値
・パーセント点・四分位偏差
（例えば75%点と25%点）
人数
0
2
4
6
8
10
12
14
16
6
6.4
6.8
7.2
7.6
8
8.4
8.8
9.2
9.6
10
10.4
10.8
11.2
11.6
12
体重（kg）

バラツキの指標

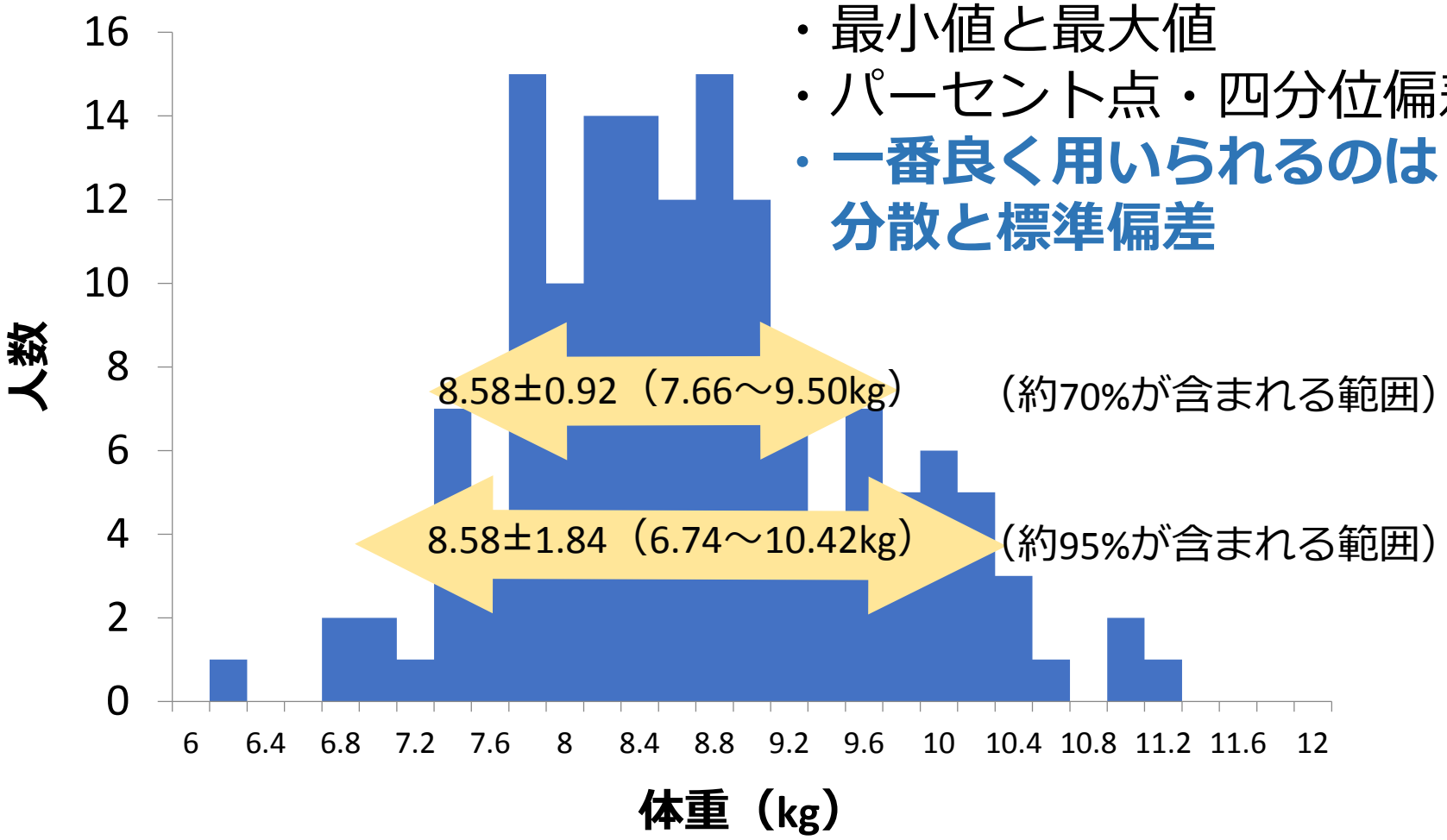


・最小値と最大値
・パーセント点・四分位偏差
・**一番良く用いられるのは**
分散と標準偏差

（約70%が含まれる範囲）

（約95%が含まれる範囲）

バラツキの指標



- 最小値と最大値
- パーセント点・四分位偏差
- **一番良く用いられるのは分散と標準偏差**

平均, 分散, 標準偏差の公式

- 平均 $= \dfrac{\text{一つ一つのデータの合計}}{\text{データの個数}}$
- 分散 $= \dfrac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{データの個数}-1}$

計算手順

1. 一つ一つのデータと平均との差をとる
2. 「平均との差の二乗」を求める
3. 「平均との差の二乗」を合計する
4. 合計を「データの個数-1」で割る

- 標準偏差 $= \sqrt{\text{分散}}$

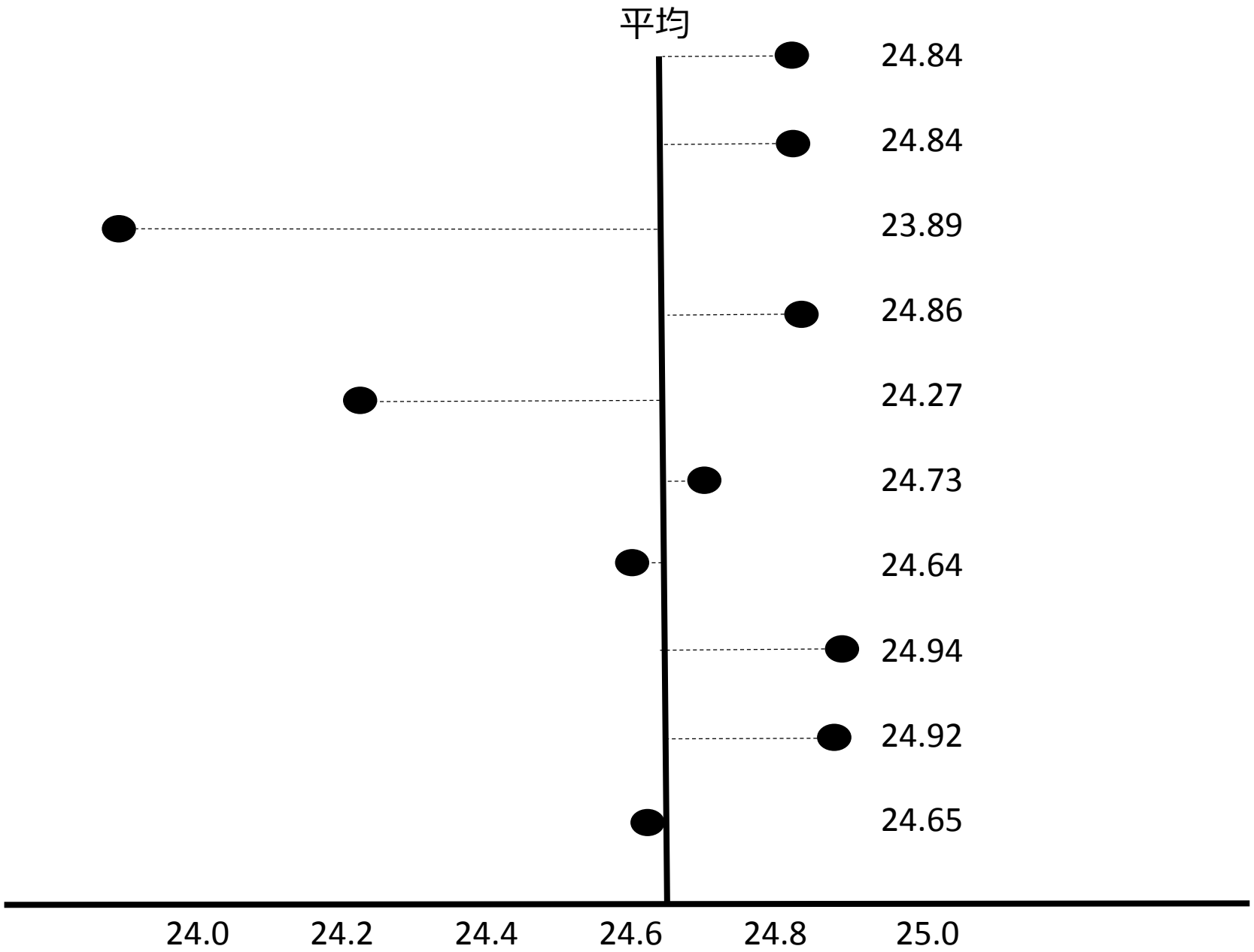
平均±2×標準偏差の例

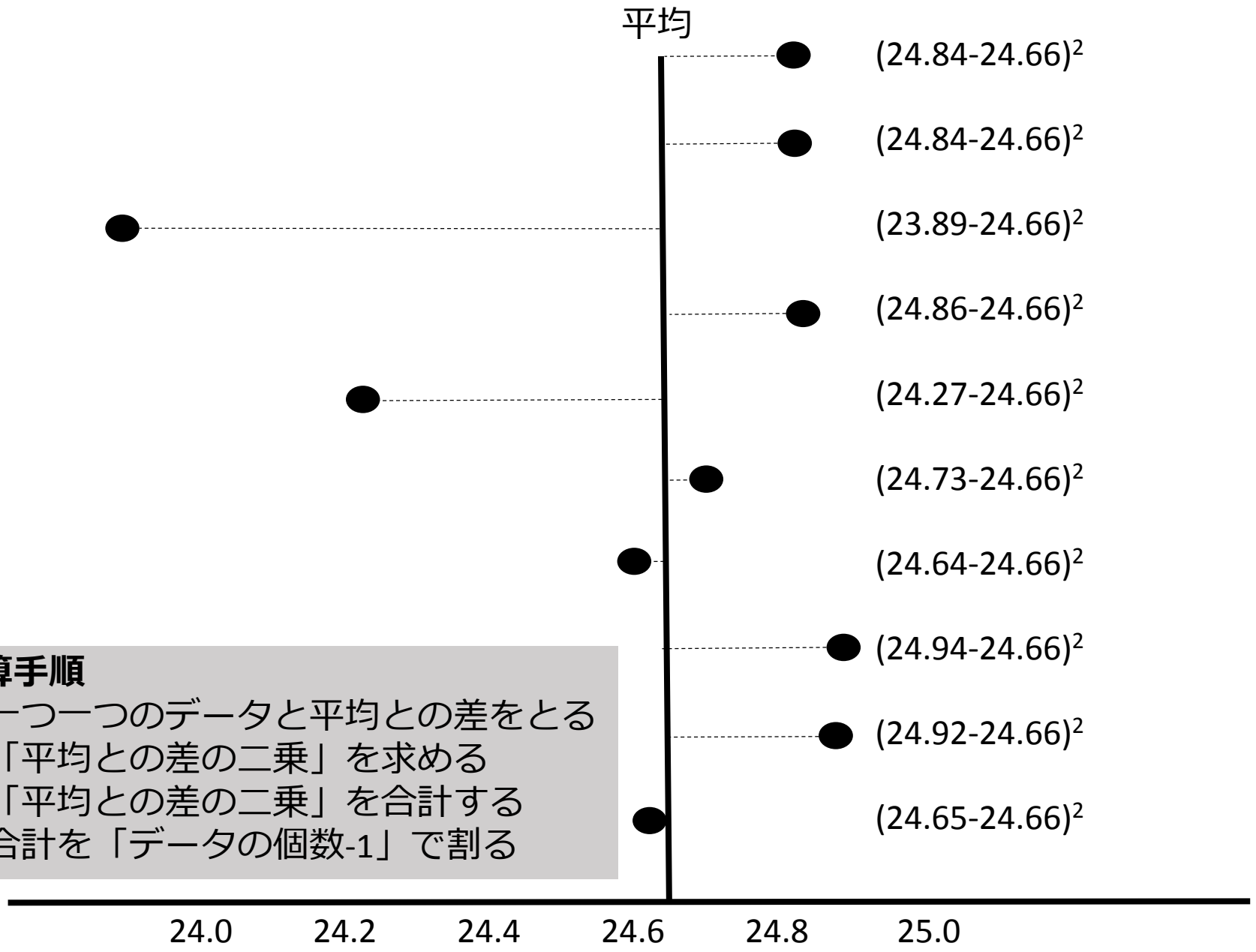
- 血糖値などの臨床検査が正常なバラツキの範囲内かどうかを判断するとき，**集団基準値**が用いられる
- 集団基準値の求め方
  - （健常人の集団における）**平均±2×標準偏差**
- 補足
  - 病気を診断するために用いられる診断基準値というのもある
    - 収縮期血圧だと140mmHg以上で高血圧と診断
  - 健常人のデータだけでなく，患者のデータや医療の必要性などによって設定される

容疑者のブーツの足跡の最大長データから計算せよ

- 分散
- 標準偏差

犯行現場に残された二つの足跡の最大長	容疑者の家にあったブーツの足跡の最大長（10回測定）	
25.52 cm	24.84 cm	24.73 cm
25.33 cm	24.84 cm	24.64 cm
	23.89 cm	24.94 cm
	24.86 cm	24.92 cm
	24.27 cm	24.65 cm


平均
24.84
24.84
23.89
24.86
24.27
24.73
24.64
24.94
24.92
24.65
24.0
24.2
24.4
24.6
24.8
25.0



計算手順

1. 一つ一つのデータと平均との差をとる
2. 「平均との差の二乗」を求める
3. 「平均との差の二乗」を合計する
4. 合計を「データの個数-1」で割る

答え

- 平均 $= \frac{24.84+24.84+\cdots+24.65}{10}$
  $= 24.66$
- 分散 $= \frac{(24.84-24.66)^2+(24.84-24.66)^2+\cdots+(24.65-24.66)^2}{10-1}$
  $= 0.11124$
- 標準偏差 $= \sqrt{0.11124}$
  $= 0.333526611$
- ここからわかること
  - 容疑者の平均±標準偏差=24.66±0.33=24.33～24.99cm
  - 犯行現場の足跡は, 25.52と25.33cmで, この範囲より大きい

どこまで表示すべきか（計算精度と見やすさ）

- すべての桁を示すと見づらい
  - 標準偏差 $= \sqrt{0.11124}$
    $= 0.333526611$
- 有効数字3桁までは標準偏差の計算精度が保たれる
  - 標準偏差=0.334
  - 日本工業規格（JIS Z9041-1）の考え方
  - 計算途中では丸めを行わない
- 測定値の表示桁数に合わせた方が見やすい
  - 足跡の最大長は小数点2桁まで測定（24.84cmなど）
  - それに合わせて表示
    - 平均=24.66
    - 標準偏差=0.33

数学記号によるデータの表現

- □番目の数値は,右下に小さく$x_{\square}$で表す
- この二つは同じ意味
  - 1番目のデータは24.84cm
  - x_1=24.84

犯行現場に残された二つの足跡の最大長	容疑者の家にあったブーツの足跡の最大長（10回測定）	
25.52 cm	x_1	x_6
25.33 cm	x_2	x_7
	x_3	x_8
	x_4	x_9
	x_5	x_{10}

数学記号によるデータの表現

- 数式では, □の代わりに i を使う
- データの個数には, n を使う
- 合計を表す記号$\sum_{i=1}^{n} x_i$
  - 最初のデータx_1から最後のデータx_nまでの和
  - $\sum_{i=1}^{10} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$

平均, 分散, 標準偏差の公式

- 平均 $= \dfrac{\text{一つ一つのデータの合計}}{\text{データの個数}}$

  $= \dfrac{\sum_{i=1}^{n} x_i}{n}$

- 分散 $= \dfrac{(\text{データ}-\text{平均})^2\text{の合計}}{\text{データの個数}-1}$

  $= \dfrac{\sum_{i=1}^{n}(x_i-m)^2}{n-1}$

  - 見やすくするため平均をmで表した

- 標準偏差 $= \sqrt{\text{分散}}$