経営統計

02 1変数の記述統計量

分寺 杏介

神戸大学　経営学部

bunji@bear.kobe-u.ac.jp

※本スライドは，クリエイティブ・コモンズ 表示-非営利 4.0 国際 ライセンス（CC BY-NC 4.0）に従って利用が可能です。

（おさらい）統計学でやろうとしていること

主に統計学の知識が必要

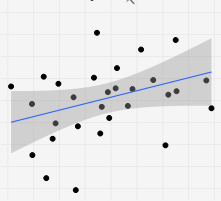
仮説をたてる → データを集める → 要約する → 統計的仮説検定 → 結論を出す → 一般化できる

領域の専門的な知識が必要

手元のデータで計算する → 全体にも当てはまるかを統計的に検証する

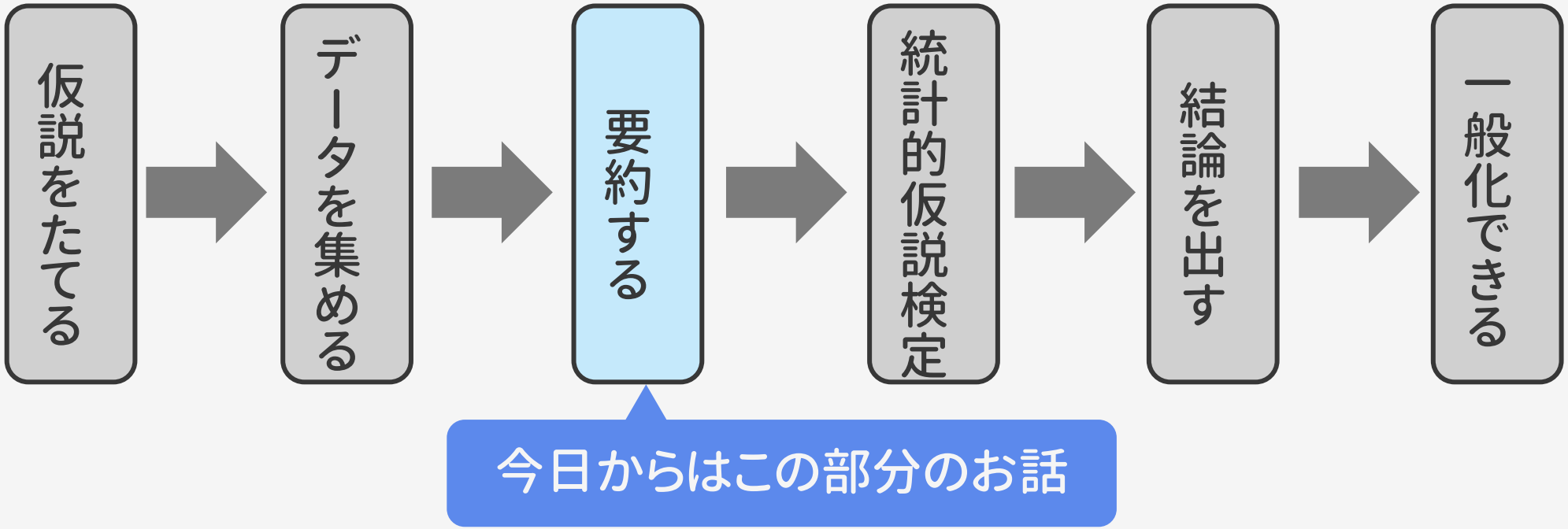


データの相関係数が0.32でした → 全体の相関係数もゼロではなさそうですね





（おさらい）統計学でやろうとしていること



データの特徴を簡潔に説明・把握するために

	A	B	C	D	E	F
1	ID	性別	身長	体重	勉強時間	テストの得点
2	1	男	172	70	3.0	91
3	2	女	158	47	0.8	31
4	3	女	160	54	4.0	34
5	4	男	178	69	3.8	51
6	5	女	153	42	1.9	87
			169	80	1.9	39
			165	70	2.1	48
			155	42	0.8	73
			159	63	2.4	62
			148	40	1.7	34
			147	42	1.6	84

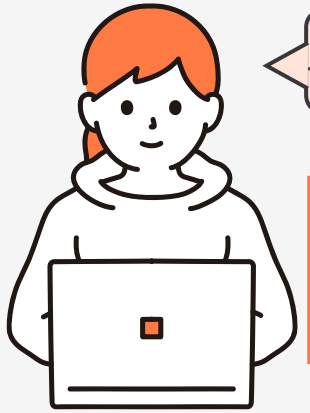
この前とったデータから
みんなの身長は
どんな感じだったか教えて





えっと…1人目が172cm，2人目が158cm，3人目が160cmで…（略）

データが多くなると，一つ一つを見ていてもきりがない上に，結局よくわからない



平均は169.5cm，標準偏差は5.6cmでした。

知りたいことは結構シンプルで，例えば
「平均はこれくらいで，これくらいのバラツキがある」
くらいがわかれば良かったりする

→ これを**記述統計量**と呼ぶ

代表値｜平均値 mean

すべてのデータを足して，個数nで割った値

これ以降，平均値は
変数の上に¯（バー）をつけて表します

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

「年齢」という変数をxと表す

	A	B	C
1	ID	性別	年齢
2	1	男	19 ← $x_1 = 19$
3	2	女	21 ← $x_2 = 21$
4	3	女	20
5	4	男	19
6	5	女	22 ← $x_n = 22$

（$n = 5$のデータ）

例｜5人の大学生の年齢がそれぞれ19, 21, 20, 19, 22だったら

$$\frac{19 + 21 + 20 + 19 + 22}{5} = \frac{101}{5} = 20.2$$

長所 すべてのデータを使っているので，いかにも「代表」らしい

無視されているデータが無く
全部を活用しているという意味です

短所 外れ値の影響を受けやすい

大人の学び直し

例｜5人の年齢がそれぞれ19, 21, 20, 19, 56だったら

一般的な大学生とはかけ離れてるなぁ

$$\frac{19 + 21 + 20 + 19 + 56}{5} = \frac{135}{5} = 27.0$$

【ちなみに】2013年花巻東高卒業生の
平均年収は約5000万円らしいです



外れ値があったらどうする?

単純なミスならば迷わず除外 or 修正する

例｜身長が1573cmだった ▶ 記録の元を確認したら**157.3**cmかもしれない

ありえる値だったら基本的には残しておく

もちろんミスではないことを確認した上で

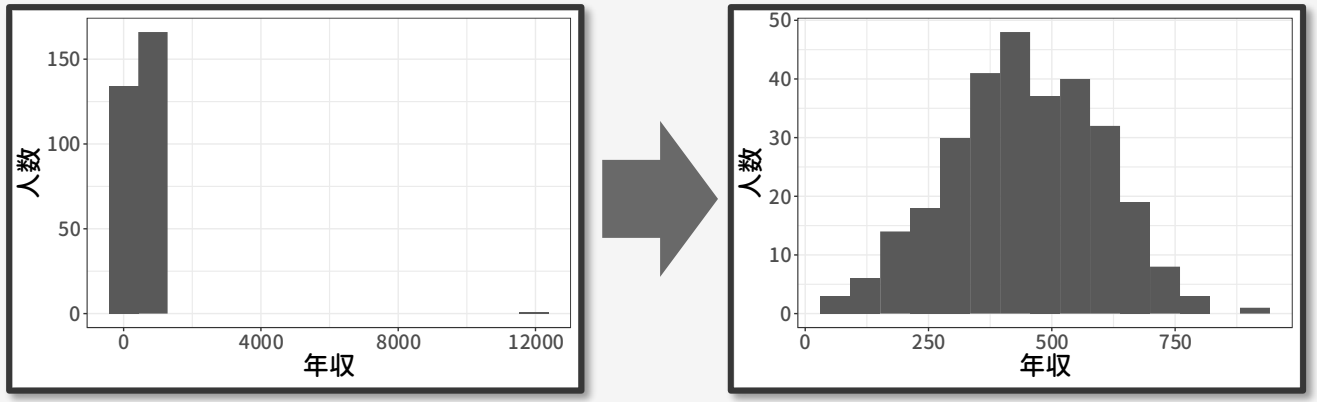
例｜年収が**12000**万円というのは「外れ値」だが現実にありえる値

▶ むしろ除外してしまうと**データ全体を反映した平均値にならない**のでダメ

ありえる値でも意図的に除外するケースもある

例｜ヒストグラムを作るとき

年収12000万円が入っている左図では，
データの分布がよくわからない。
このような場合**「外れ値は除外した」ことを明記して**
関心のある区間に絞ったヒストグラム（右図）を
作ったりする



代表値|中央値 median

データを小さい順に並べたときの真ん中の値

例|5人の年齢がそれぞれ19, 21, 20, 19, 22だったら

▶小さい順に並べると19, 19, 20, 21, 22なので，真ん中は**20**

データが偶数個のときは真ん中に近い2つの平均値をとる

長所 **外れ値につよい**

短所 **すべてのデータを使っているわけではない**

例|5人の年齢がそれぞれ19, 21, 20, 19, **56**だとしても

▶小さい順に並べると19, 19, 20, 21, 56なので，真ん中は**20**のまま

極端な話，5人の年齢がそれぞれ「20未満」，「20未満」，20，「20以上」，「20以上」ならば，
他の4人の年齢がいくつであっても中央値は20になる。
平均値がすべてのデータを反映した代表値であった一方で，**中央値は「小さい順での真ん中の値」以外は(順序以外)全く反映されていない**という点で，「代表」感は薄いと感じるかもしれない。

代表値|最頻値 mode

最も出現回数(度数)の多い値

例|5人の年齢がそれぞれ19, 21, 20, 19, 22だったら

▶最も出現回数が多いのは**19**(唯一2回出現)

長所 外れ値につよい

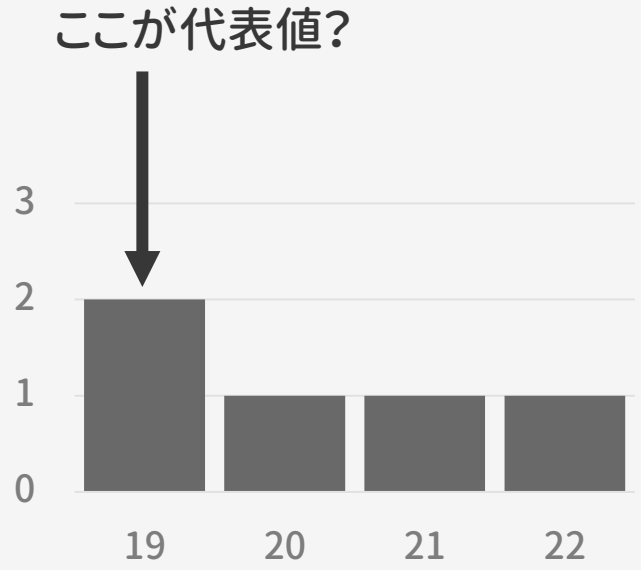
短所 すべてのデータを使っているわけではない

分布の端になってしまうかもしれない

> 上の例でも19はデータの中での最小値
> 「代表」と言っているくせに分布の端を採用するのは違和感がある

連続変数だと使いづらい

> 本来連続変数では小数を細かくしたら同じ値は無いはずなので,
> **すべての度数が1になり**最頻値が定義されなくなってしまう



で，結局どれを使えばいいの？

ルールも正解もないんだよ

❌「外れ値があるときは平均値はダメだ！」

❌「最頻値は代表値としては欠陥品だ！使わないほうがいい！」

「こういうときはこう」と決めつけてしまうと判断ミスを招く可能性があるので，状況に応じて柔軟に使い分けてください

みんな違って，みんないい

何をもって「代表」とするかの考え方が色々ある，というだけの話

▶**「なぜその代表値を使うのか」に責任と自覚を持ってください**

そのために，各代表値の定義と特徴をきちんと理解しておく必要があるのです

なんなら全部報告しても問題ない

例｜身長の平均値も中央値も最頻値も172cmでした

▶ 身長の分布が「だいたい左右対称」であることがわかる

後の回でもう少し出てきますが，とりあえず
「完全に左右対称の分布では平均値と中央値が同じになる」
ということは覚えておいて損はありません

代表値の使い所

とはいえ，まずは平均値がファーストチョイス

「データ全体の特徴」を表すものとして最もメジャーでもあるため

変数の性質・分布によっては別の代表値のほうが意味を持つことも

例 外れ値・歪んだ分布

▶ 平均値は実態に合わない可能性がある

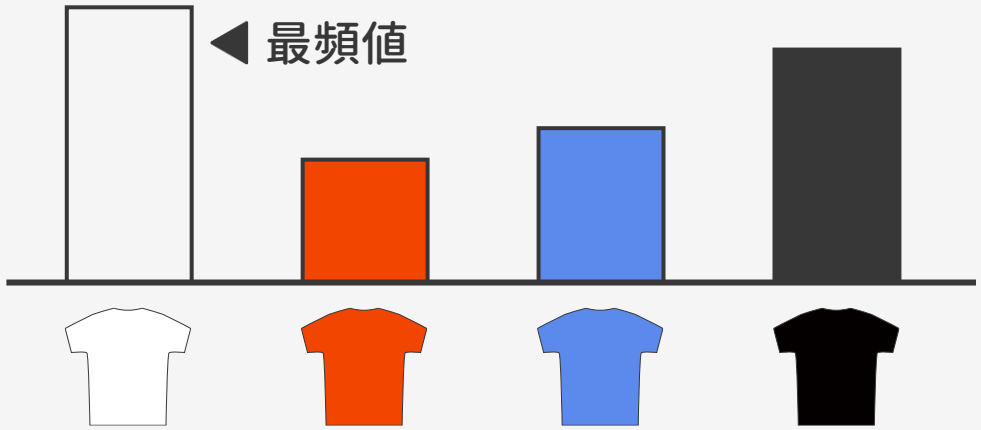
【マーケティング：コーヒー店の客単価】



例 質的変数

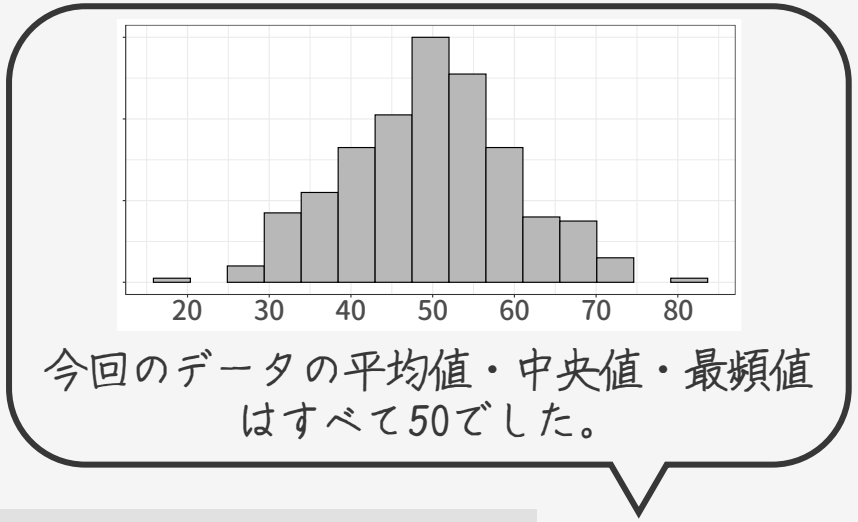
▶ そもそも平均値・中央値は求められない

【生産管理：Tシャツの販売枚数】



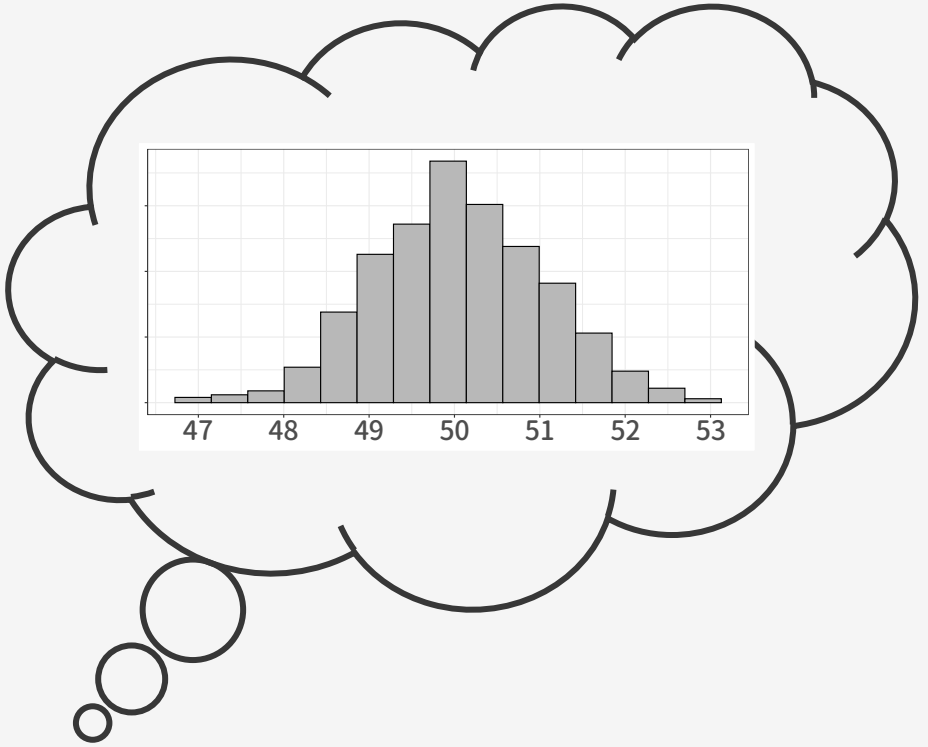
散布度

データの散らばり具合



代表値が同じでも，散布度が異なるとデータの分布のイメージは全く異なってしまう







実際には代表値も散布度も同じだからといって分布が完全一致するわけでもないので，
最終的には**ヒストグラムなど図を使って**
データを確認するのを忘れずに!

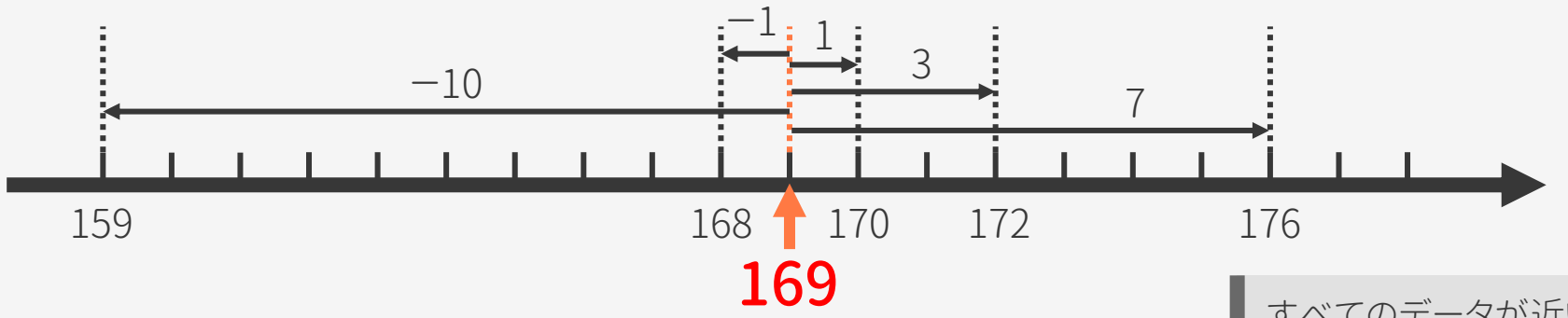
平均値からの差

例｜5人の身長がそれぞれ172, 168, 170, 159, 176だったら

身長	172	168	170	159	176
平均を引く	172−169	168−169	170−169	159−169	176−169
偏差	**3**	**−1**	**1**	**−10**	**7**

→ 平均値は**169**

数直線上で表すと…



すべてのデータが近い値だと
偏差もすべて小さくなるので

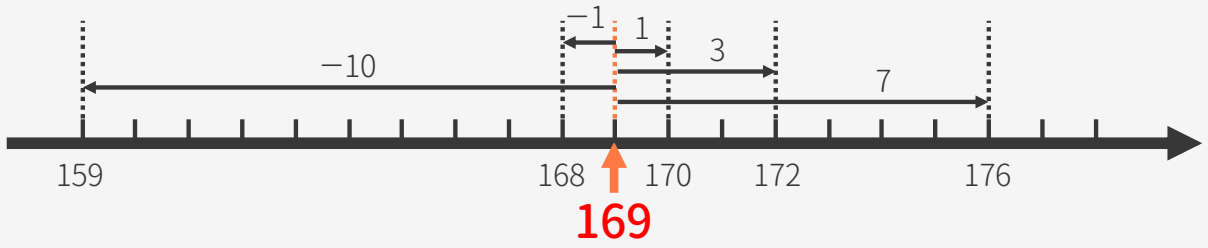
偏差の大きさ＝ばらつきの大きさ
▶散布度は「偏差の代表値」として表す

偏差の代表値＝偏差の平均値

実際のデータ分析で使うことはあまり無いですが、考え方の導入として紹介します

身長	172	168	170	159	176
平均を引く	172−169	168−169	170−169	159−169	176−169
偏差	3	−1	1	−10	7

平均値は0



偏差	3	−1	1	−10	7
絶対値	3	1	1	10	7

$$\frac{3+1+1+10+7}{5}=\frac{22}{5}=4.4$$ 平均偏差

平均値169cmから平均して4.4cmずれているという解釈になる

偏差の平均値をとると必ず0になる

「ばらつき」という意味では偏差がプラスかマイナスかは関係ない

偏差の絶対値の平均値を計算してあげる

→ これが**平均偏差**

平均値からのズレの大きさは「−10」も「10」も同じ

偏差の平均値その2

テキストによっては $n-1$ で割っているものもありますが
今は完全無視してください。

身長	172	168	170	159	176
平均を引く	172−169	168−169	170−169	159−169	176−169
偏差	3	−1	1	−10	7

平均値は0

偏差	3	−1	1	−10	7
二乗	9	1	1	100	49

$$\frac{9+1+1+100+49}{5}=\frac{160}{5}=32$$ 分散

本当は $\sqrt{32}\fallingdotseq\pm5.656$ ですが，散布度は負の値にならないので+5.656になります

$$\sqrt{32}\fallingdotseq 5.656$$ 標準偏差

これも平均値169cmから平均して5.656cmずれているという感じで解釈できる

偏差の二乗の平均値を計算してあげる
→ これが**分散**

$$s_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

分散は二乗されているため
平均偏差のように解釈できない

ルートをとることで
単位をもとに戻してあげる
→ これが**標準偏差**

$$s_x=\sqrt{s_x^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

分散の単位は二乗されているため
「平均値169cmから32cm
ずれている」という解釈はできない

変数変換と代表値・散布度

① 変数を a 倍したとき

テスト得点	80	56	76	63	70
平均を引く	80−69	56−69	76−69	63−69	70−69
偏差	11	−13	7	−6	1
偏差の二乗	121	169	49	36	1

→ 平均値は**69**

→ 分散は**75.2**

標準偏差はおよそ**8.67**

例｜2倍すると

テスト得点	2×80	2×56	2×76	2×63	2×70
平均を引く	2×(80−69)	2×(56−69)	2×(76−69)	2×(63−69)	2×(70−69)
偏差	2×11	2×−13	2×7	2×−6	2×1
偏差の二乗	2^2×121	2^2×169	2^2×49	2^2×36	2^2×1

→ 平均値は**2×69**

→ 分散は**2^2×75.2**

標準偏差はおよそ**2×8.67**

$$\sqrt{2^2 \times 75.2} = \sqrt{2^2} \times \sqrt{75.2} = 2 \times \sqrt{75.2} \fallingdotseq 2 \times 8.67$$

② 変数に b を足したとき

テスト得点	80	56	76	63	70
平均を引く	80−69	56−69	76−69	63−69	70−69
偏差	11	−13	7	−6	1
偏差の二乗	121	169	49	36	1

→ 平均値は**69**

→ 分散は**75.2**

標準偏差はおよそ**8.67**

例｜5を足すと

テスト得点	5+80	5+56	5+76	5+63	5+70
平均を引く	(5+80)−(5+69)	(5+56)−(5+69)	(5+76)−(5+69)	(5+63)−(5+69)	(5+70)−(5+69)
偏差	11	−13	7	−6	1
偏差の二乗	121	169	49	36	1

→ 平均値は**5+69**

→ 分散は**75.2**

標準偏差はおよそ**8.67**

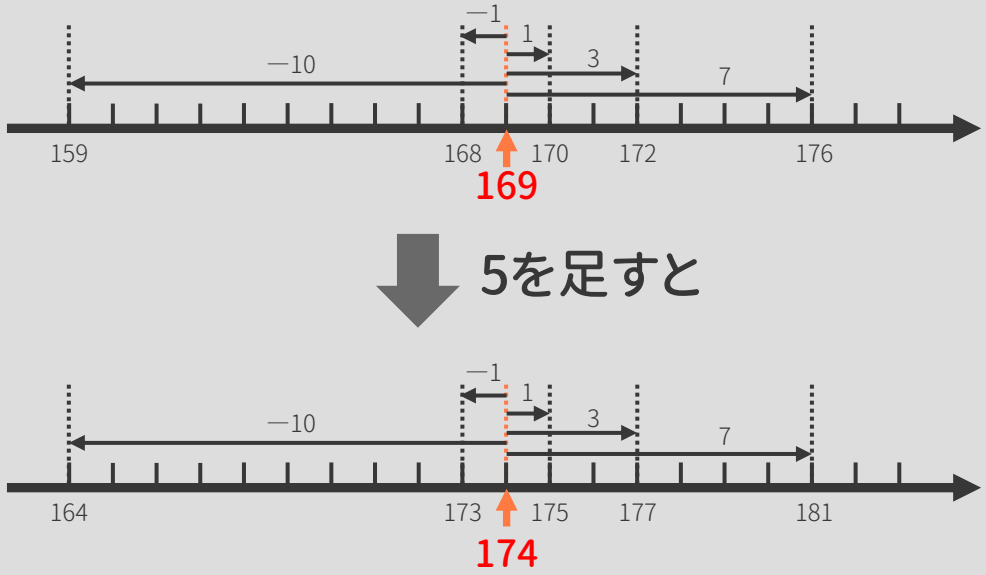
つまり**散布度は変わらない**

変数変換と代表値・散布度

まとめると

$|a|$ は a の絶対値を表す記号
例：$|-3| = 3, |3| = 3$

	a倍する	bを足す	a倍してbを足す
平均値	a倍になる	b大きくなる	a倍＋bになる
中央値	a倍になる	b大きくなる	a倍＋bになる
最頻値	a倍になる	b大きくなる	a倍＋bになる
平均偏差	$\|a\|$倍になる	変わらない	$\|a\|$倍になる
分散	a^2倍になる	変わらない	a^2倍になる
標準偏差	$\|a\|$倍になる	変わらない	$\|a\|$倍になる



この表をそのまま覚えるのではなく，
それぞれの統計量の定義をきちんと理解しましょう

全員同じ値を足しても数直線上では平行移動するだけ
▶ばらつきが変わることはない

記述統計量は他にもいろいろ

用途に応じて柔軟に考えましょう

変動係数 coefficient of variation

一般的に、値が大きいものほどばらつきは大きくなる

- CEOの年収の標準偏差 … ウン千万?
- フリーターの年収の標準偏差 … せいぜい100-200万くらい?

⟶ **「値の大きさの割にばらついている」度**を表現したい

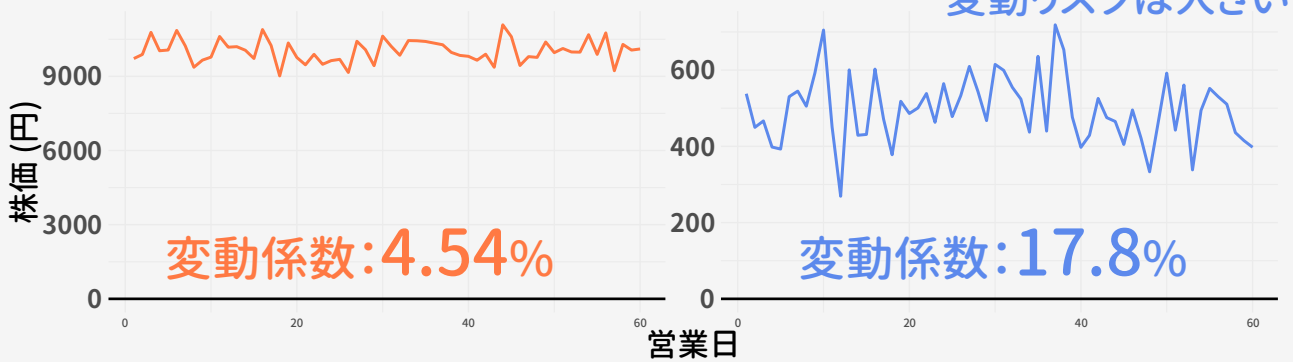
定義上，変動係数は比率尺度に対してのみ意味のある統計量です

$$\text{変動係数} = \frac{\text{標準偏差}}{\text{平均値}} = \frac{s_x}{\bar{x}}$$

例 2つの銘柄の株価の推移



▲ 赤の銘柄のほうが変動が大きいが基準額が高いので当たり前



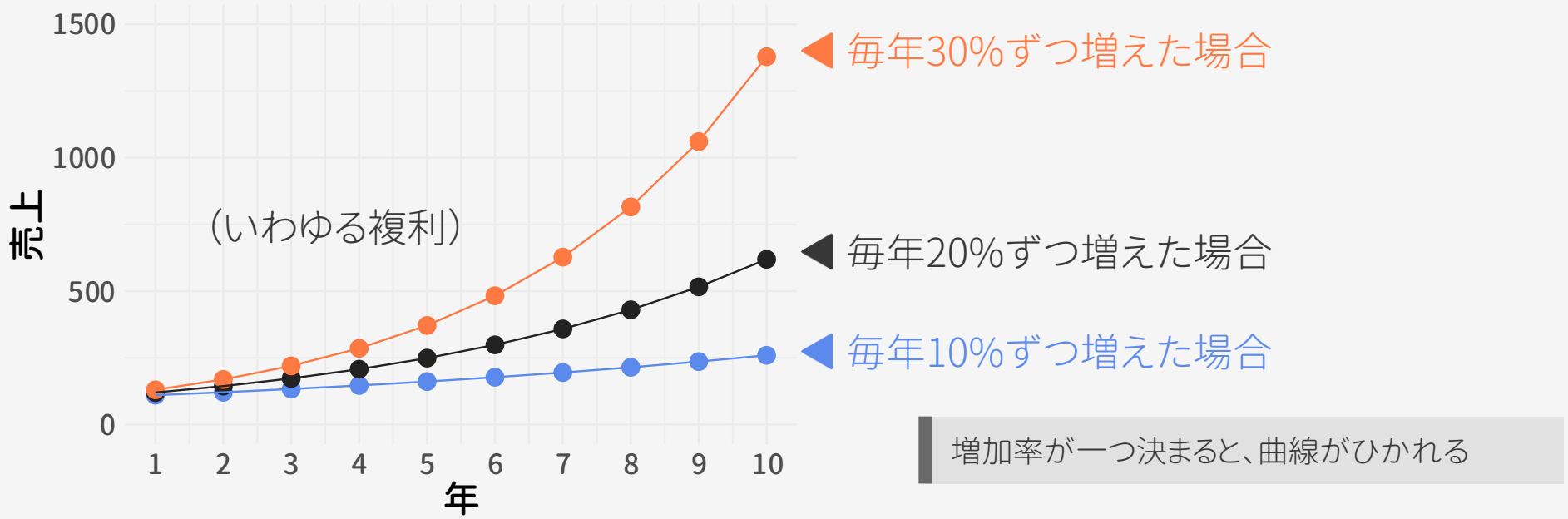
▼ 青の銘柄のほうが変動リスクは大きい

用途に応じて柔軟に考えましょう

幾何平均 geometric mean

徐々に増えるものの「率」に関心がある場合に使われる

- 10年間で、売り上げは年平均何%増加したか？



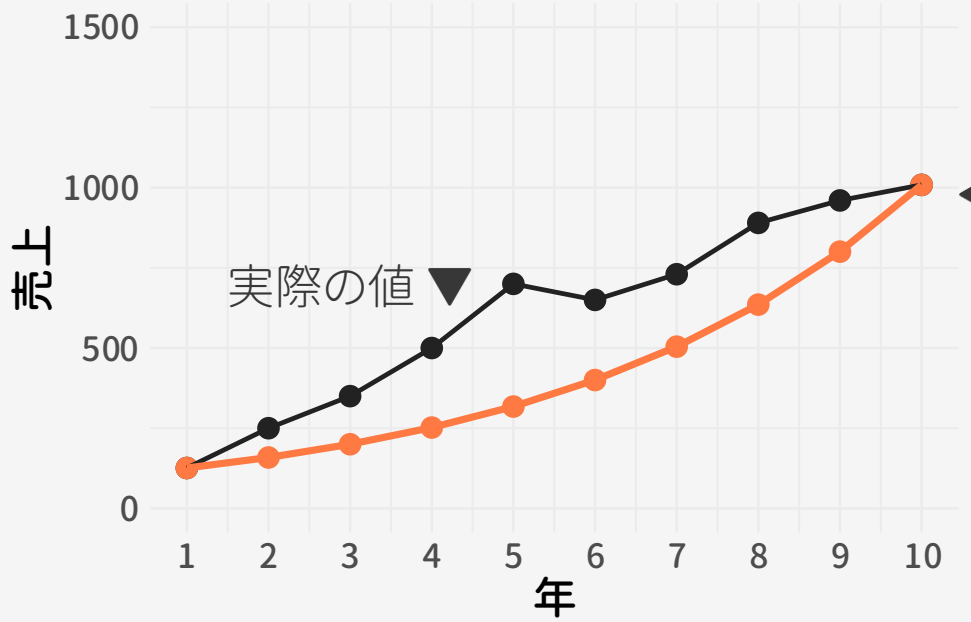
増加率が一つ決まると、曲線がひかれる

用途に応じて柔軟に考えましょう

幾何平均 geometric mean

徐々に増えるものの「率」に関心がある場合に使われる

- 10年間で、売り上げは年平均何%増加したか？



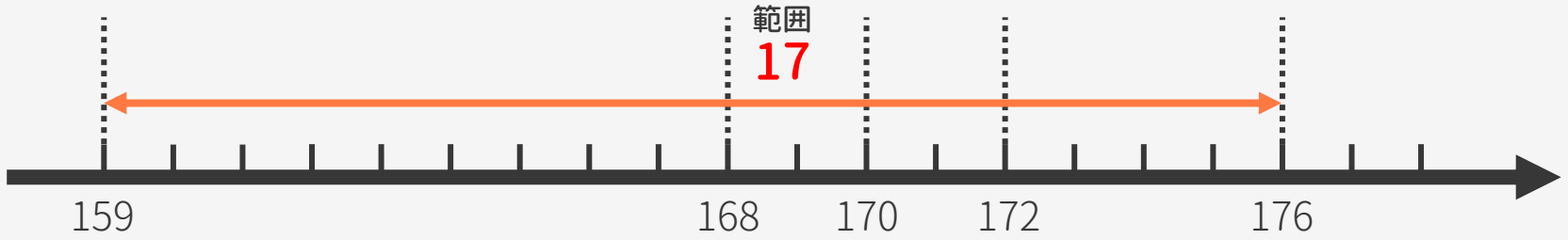
◀ 始点と終点を結ぶ曲線を作る増加率が**幾何平均**

$$\text{幾何平均} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 a_3 \cdots a_n}$$

※ ここでの a_i は i 時点目での増加率 $a_i = \frac{x_i}{x_{i-1}}$

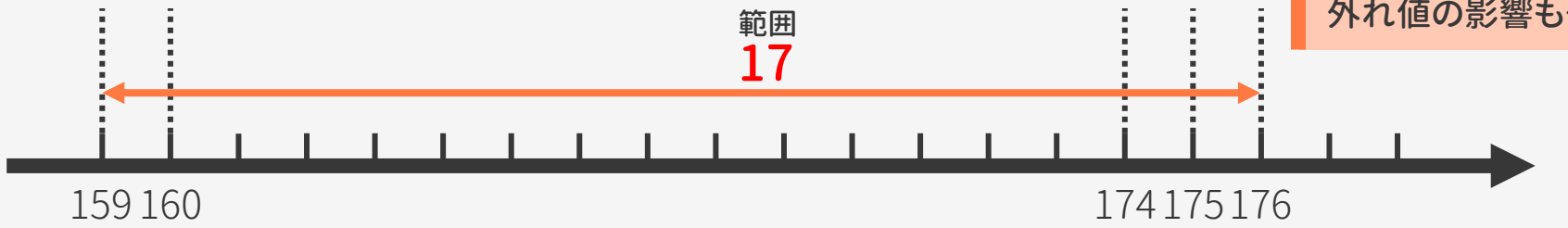
最大値と最小値の差

身長	172	168	170	159	176



範囲内の散らばり具合は全くわからない

身長	175	174	160	159	176



最大値と最小値のみを用いて計算するので
外れ値の影響もモロに受ける

四分位数と四分位範囲 quartile and interquartile range

データを4つに分けた範囲

例｜100人の身長を小さい順に並べたもの

【データの数が奇数の場合】
1. 中央値にあたるデータを除外してから
2. 上位群の中央値が第3四分位数
3. 下位群の中央値が第1四分位数

	第1四分位群				第2四分位群			第3四分位群			第4四分位群			
順位	1	2	…	25	26	…	50	51	…	75	76	…	99	100
身長	148.0	149.2	…	160.2	160.6	…	169.2	169.4	…	175.4	175.8	…	189.9	192.1

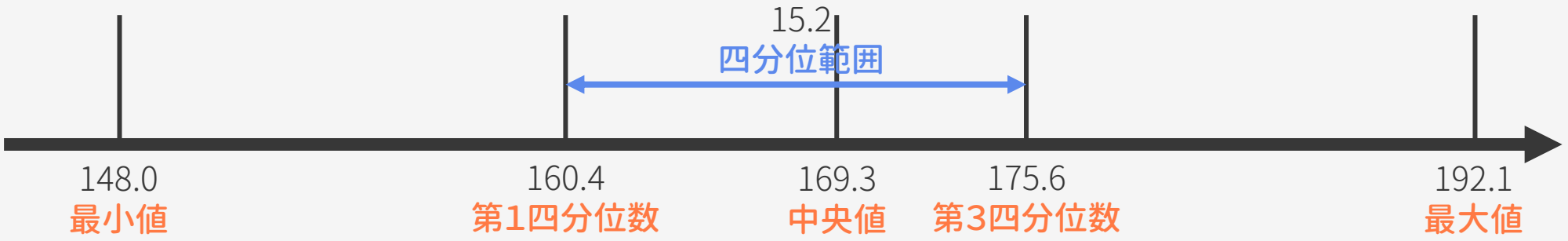
下位群の人数をちょうど
半分ずつに分ける値（中央値）

160.4
第1四分位数

169.3
中央値
（第2四分位数）

175.6
第3四分位数

上位群の人数をちょうど
半分ずつに分ける値（中央値）



（おまけ）平均偏差よりも標準偏差が使われる理由①

平均偏差では場合分けが必要になるため

平均値が a のときのデータ x の偏差 y を式で表してみる

平均偏差の場合

$$y = |x - a|$$

なのだが，実際には

$$y = \begin{cases} x - a & (x \geq a\text{のとき}) \\ -x + a & (x < a\text{のとき}) \end{cases}$$

絶対値を使うと
いちいち場合分けして
2つの式を考える必要が生じる

標準偏差の場合

$$y = (x - a)^2$$

これはそのまま扱える

二乗を使うと
場合分けする必要がなく
1つの式を考えるだけで済む

加えて，絶対値がある式は微分ができないという
弱点があるため，平均偏差は何かと使いづらいのです

（おまけ）平均偏差よりも標準偏差が使われる理由②

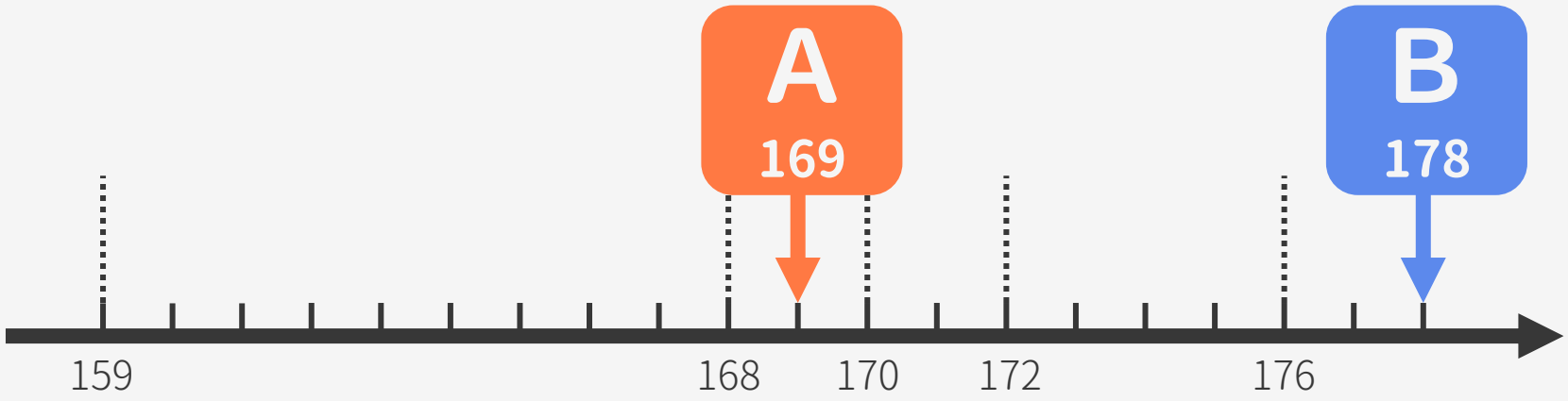
分散のほうが平均値と密接に関係しているため

ここでは，代表値を代表値たらしめる理由について考えます。

Q.どちらのほうが代表値っぽいですか？

「平均値」だ何だという知識は一旦おいて
直感的に「代表」っぽいと感じる方を選ぶとすると…？

身長	172	168	170	159	176



**たぶん平均値を知らない人に聞いても
大多数はAを選ぶのではないでしょうか**

リーダーってクラスの中心人物が
選ばれることが多いですよね

（おまけ）平均偏差よりも標準偏差が使われる理由②

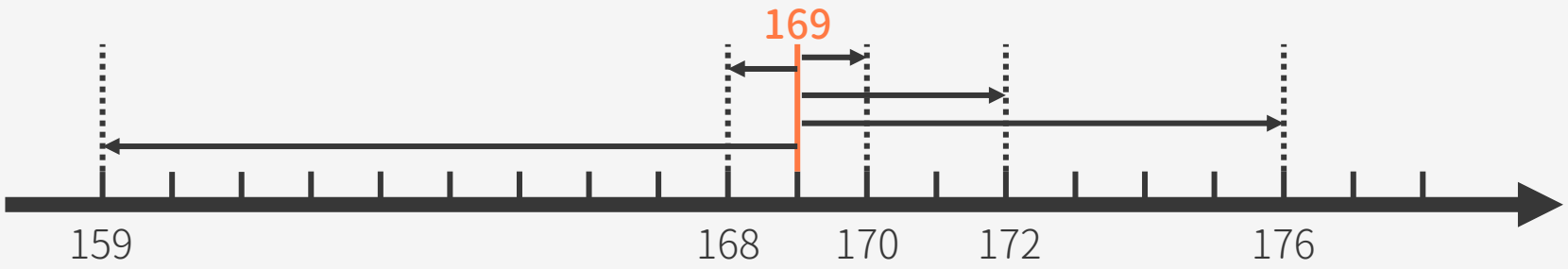
なんでAのほうが「代表」っぽいのか？

→ BよりもAのほうがデータに近いから

人によっては別の理由を思いつくかもしれませんが
ここではあくまで「一般論」として話を進めます

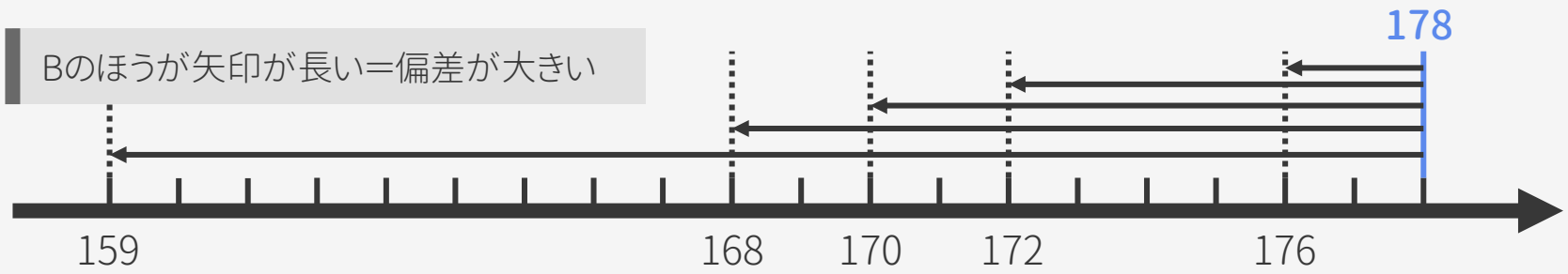
身長	172	168	170	159	176

A
169



B
178

Bのほうが矢印が長い＝偏差が大きい



言い方を変えると

代表値とは
全体的にその値からの偏差が小さくなるような値のことである

（おまけ）平均偏差よりも標準偏差が使われる理由②

具体的に「全体的に偏差が小さくなる値」って？

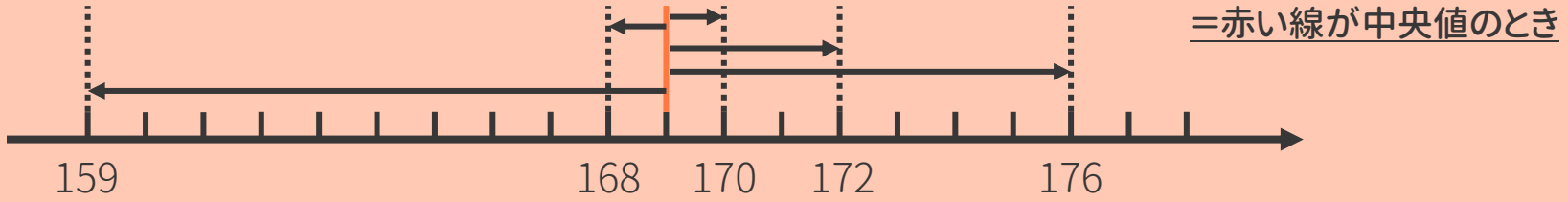
「全体的に」なので「偏差の○○の合計値」が小さくなる値を考えます

→ **偏差の扱い方によって変わります**

偏差の絶対値の場合

偏差の絶対値の合計が
最小になる値は
データの**中央値**

赤い線（代表値）を右に動かすと「偏差の絶対値の合計」は1小さくなる
▶ 赤い線より大きいデータの数のほうが小さいデータの数よりも多いため
▶ 赤い線より大きいデータの数と小さいデータの数が同じになると「偏差の絶対値の合計」は最小



偏差の二乗の場合

偏差の二乗の合計が
最小になる値は
データの**平均値**

代表値として中央値よりも
平均値＝「偏差の二乗の合計が最小になる値」がよく使われる

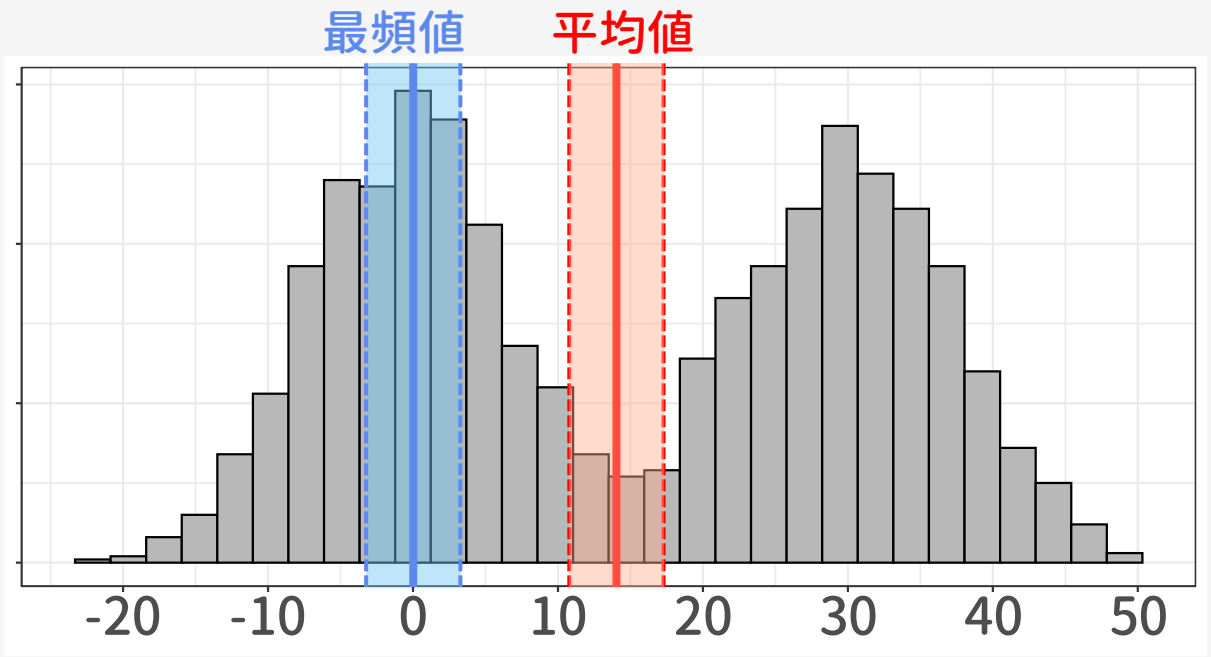
散布度も，平均値によって最小化された値である
分散（および標準偏差）がよく使われる

（おまけついでに）最頻値は偏差の何を最小化する代表値なのか？

答．偏差の絶対値が一定の値以上のデータの割合を最小化する代表値

→データのボリュームゾーンを捉えるのに適した代表値と言える

例｜ある変数のヒストグラムが2つの山になっていたら



平均値の場合

平均値は2つの山の
ちょうど中間くらいになる

例えば**平均値**±3の範囲のデータは
結構少ない

最頻値の場合

最頻値は2つの山の
どちらか一方の頂点になる

最頻値±3の範囲のデータは
平均値±3の範囲のデータよりも多い

標準化とは

異なる変数のスケールを揃えるために

前回の模試では70点だったのに，
今回は50点に下がってるじゃない！
成績が落ちてるから睡眠禁止です!



前回より難しかっただけなのになぁ・・・

異なる変数をそのまま比較する場合
ただ値を比較すると良くない

変数のスケールを揃えて
比較できる形にしよう

テストで大事なことは
他の人よりもできているか

言い換えると

平均値よりどれくらい高得点か

言い換えると

平均値からの偏差はいくつか

偏差に変換してみます

標準化とは

偏差に変換した結果

前回の模試では平均値+10点だったのに，
今回は平均値+5点に下がってるじゃない！
やっぱり成績が落ちてるから自由民禁止です!



期末はみんな似たような点数だったんだよなぁ…

平均値からの偏差に変換しただけでは
まだ比較は十分ではない

【極端な例】

前回の模試 平均60点

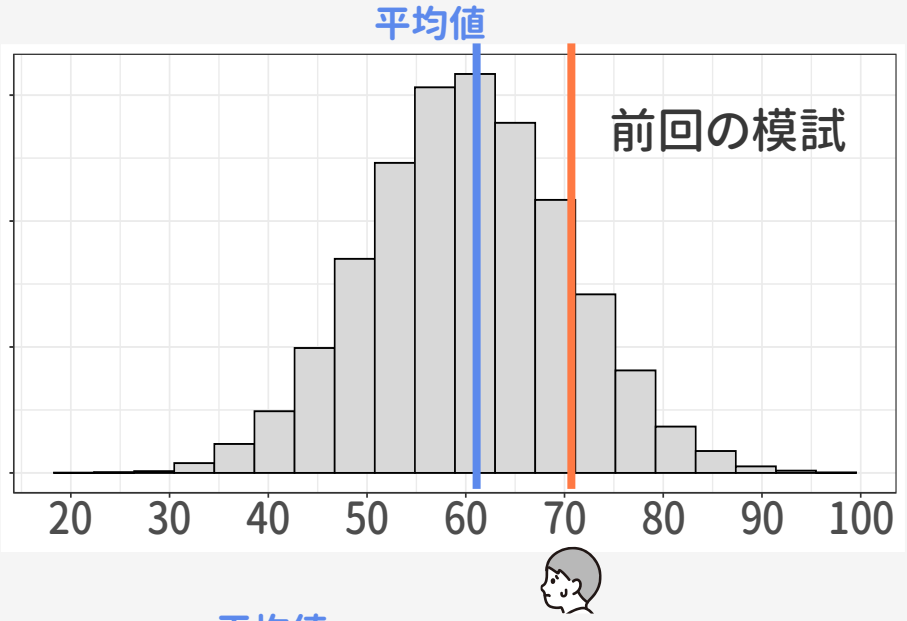
順位	点数
1	100
2	99
3	99
⋮	⋮
70	70
⋮	⋮
最下位	32

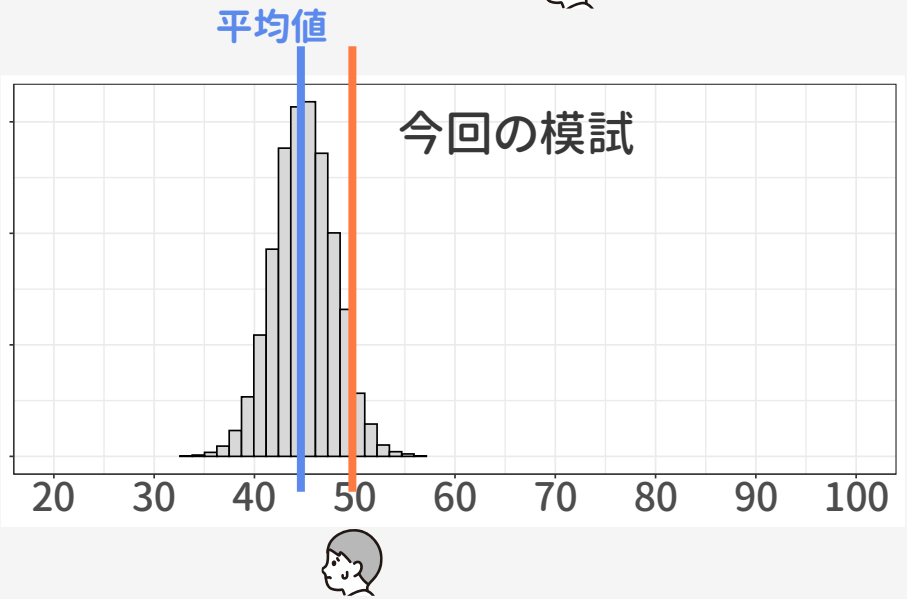
今回の模試 平均45点

順位	点数
1	55
2	55
3	54
⋮	⋮
50	50
⋮	⋮
最下位	36

どうしたら比較できる?

ヒストグラムを書いてみると





今回の模試のほうが

平均値からの偏差は小さいが

分布の中では上位にいる

なぜ？

今回の模試のほうが

散布度が小さいため

平均値付近に多くの人がいる

つまり？

偏差を散布度で調整したら

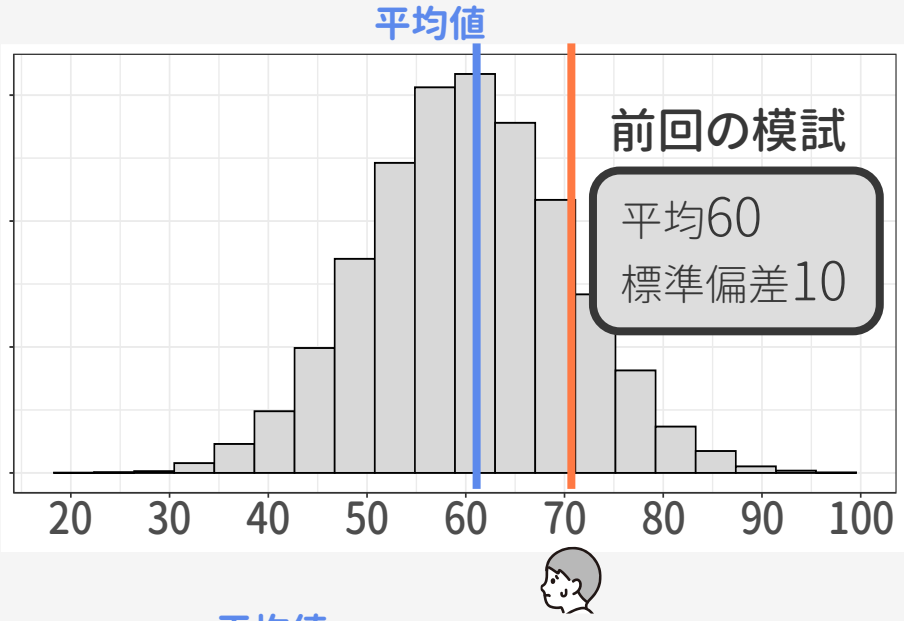
異なる変数を比較できる！

→ この手続きを**標準化**と呼ぶ

標準化のイメージ

散布度は「**偏差の代表値**」でした

実際には標準偏差は「偏差の平均値」ではないですが，ここではイメージを掴むために「偏差の平均値」として話を進めます



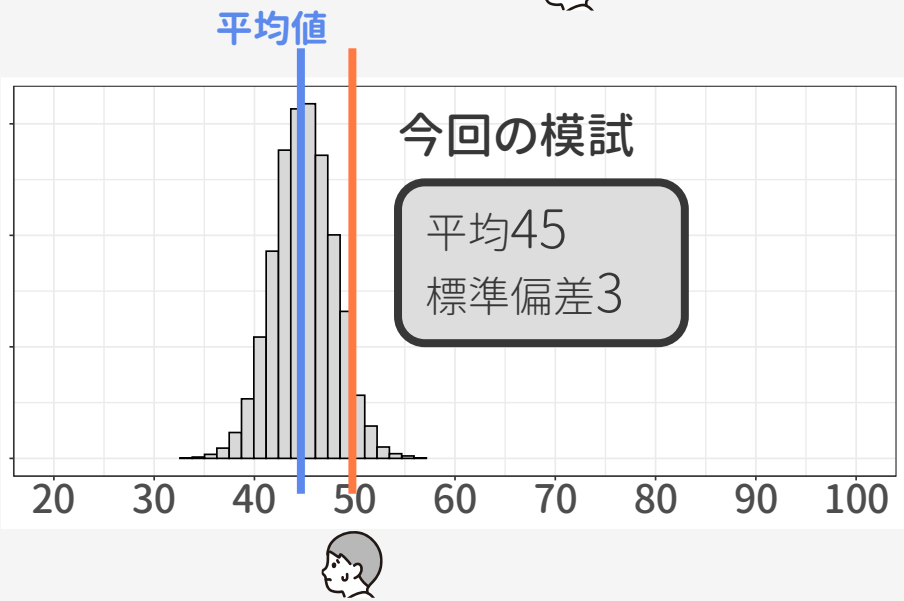
標準偏差10ということは**平均値からの偏差の平均が10**

→ の平均値+10点は**ある意味「平均的」**なズレ
あるいは偏差10

標準化とは

偏差が「標準偏差」いくつ分なのか

に変換すること



標準偏差3ということは**平均値からの偏差の平均が3**

→ の平均値+5点は**「平均的」よりもちょっと大きな**ズレ
あるいは偏差5

偏差が標準的（＝偏差が3）な人と比べるとの偏差は5/3=1.67倍もある，ということ

標準化の手順

① 変数を平均値からの偏差に変換する＝平均値を引く

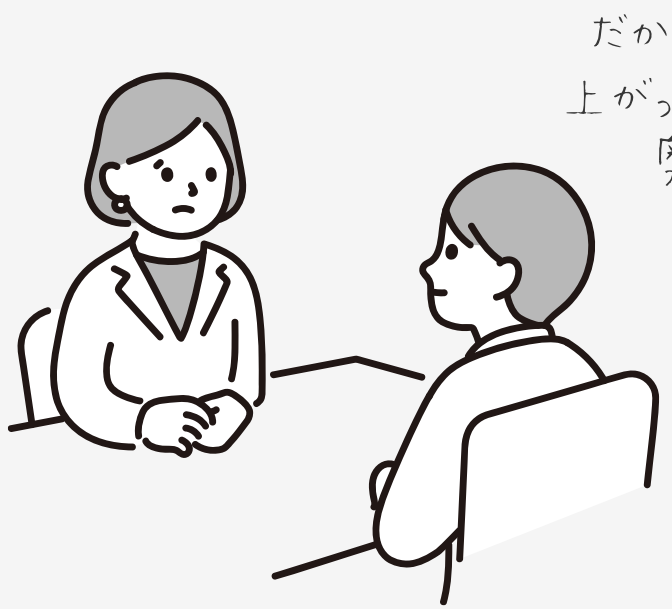
② 偏差を標準偏差で割る

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

①

②

これだけ。



前回の模試

平均点60点，標準偏差10点のテストで70点をとったので

$$\frac{70-60}{10} = \frac{10}{10} = 1$$

今回の模試

平均点45点，標準偏差3点のテストで50点をとったので

$$\frac{50-45}{3} = \frac{5}{3} \fallingdotseq 1.67$$

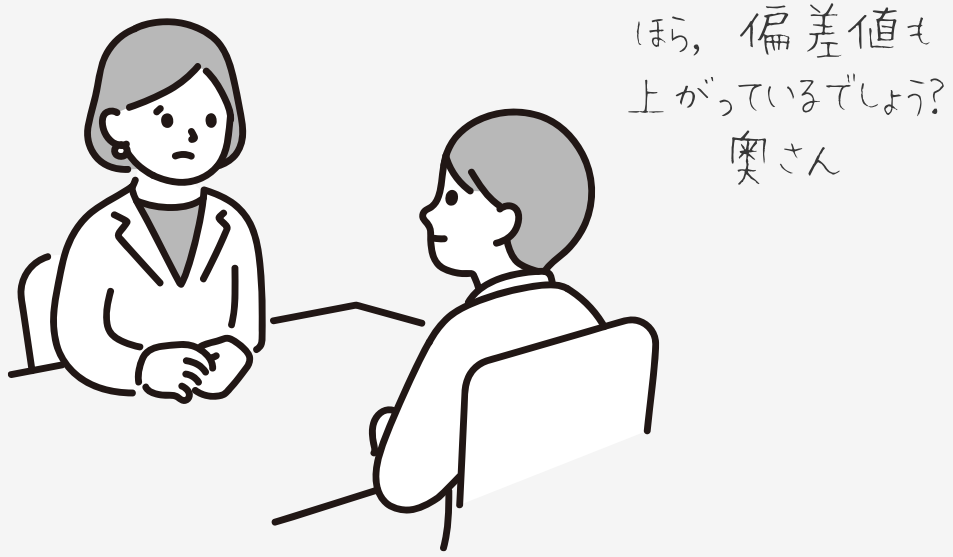
標準化得点

実はおなじみの標準化

日本の高校生は標準化得点に人生を左右されている…

偏差値 $= 50 + (10 \times$ 標準化得点$)$

あるいはテスト得点を平均値50，標準偏差10に標準化した値とも言える



前回の模試

平均点60点，標準偏差10点のテストで70点をとったので標準化得点は1

→ 偏差値は $50 + (10 \times 1) = 60$

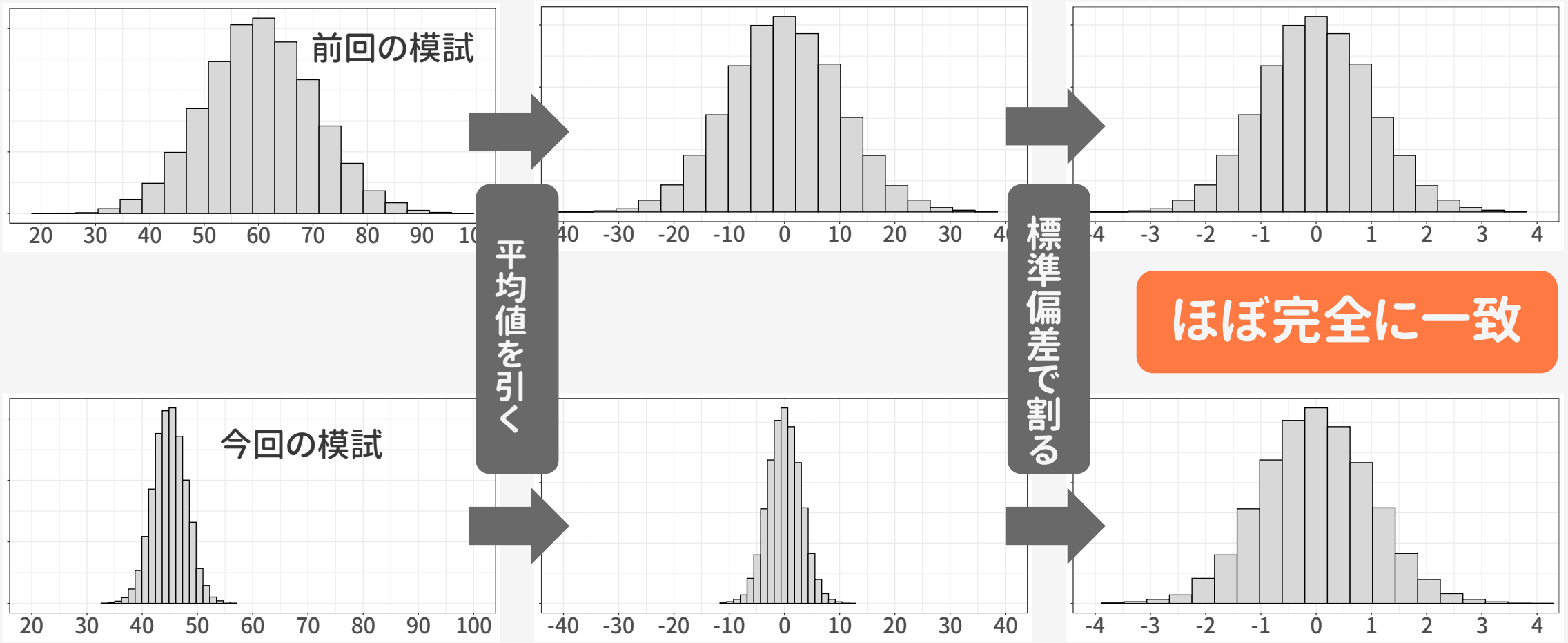
今回の模試

平均点45点，標準偏差3点のテストで50点をとったので標準化得点は1.67

→ 偏差値は $50 + (10 \times 1.67) = 66.7$

標準化のメリット

変数の分布を揃えることができる



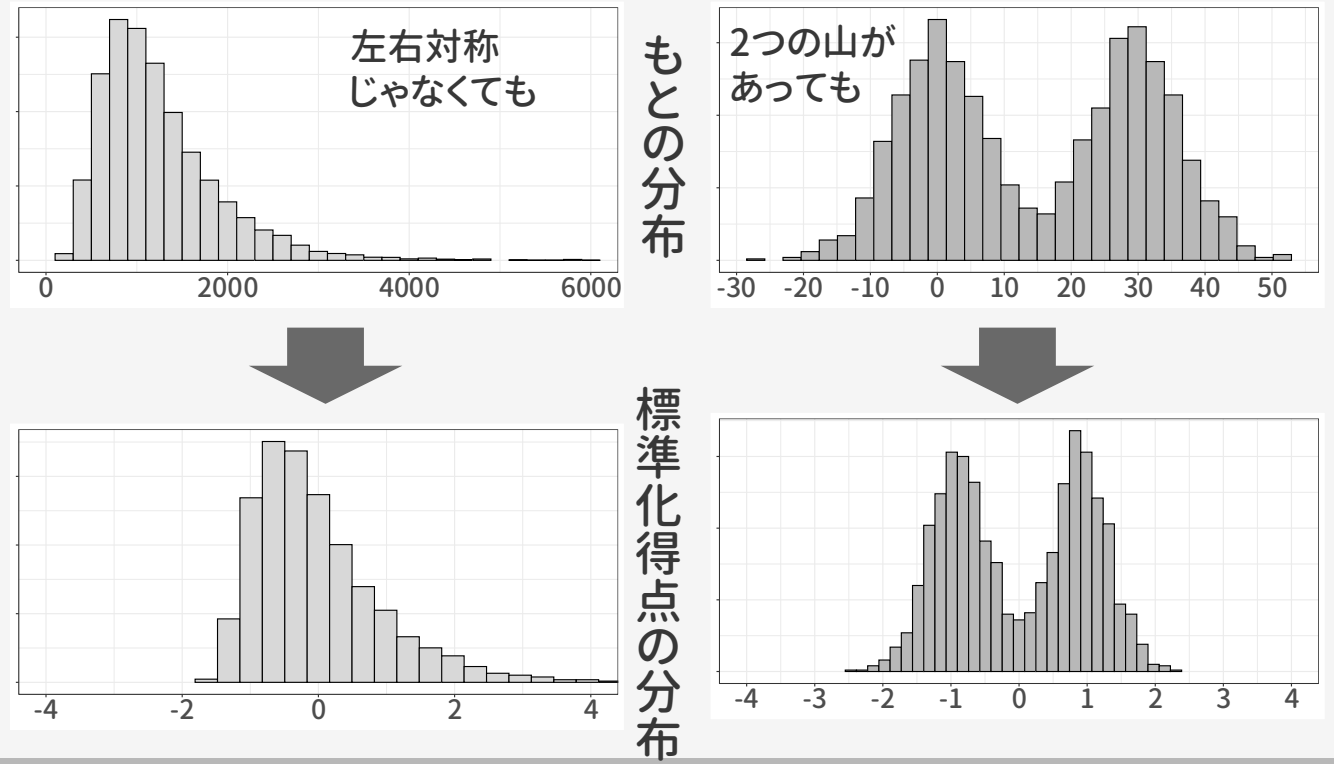
標準化の特徴

標準化得点は必ず平均0，標準偏差1になる pp. 15-17

当然もとの平均値・標準偏差がいくつであっても成り立つ

厳密には「もとの標準偏差が0じゃなければ」

標準化前後で分布の大まかな形は変わらない



前スライドの内容と合わせると

もとの分布の形状が同じであれば
平均値・標準偏差が何であれ
標準化得点の分布は同じ形になる

といえる

「分布の形状が同じ」の意味に関しては
どこかで詳しく説明しますが，例えば
前スライドの「前回の模試」と「今回の模試」は
同じ形状の分布（正規分布）として扱われます。

「標準化」standardization

本当は平均値と標準偏差が特定の値になるような操作全般を指す

→ 平均値0，標準偏差1に限った話ではない

例｜IQは平均値100，標準偏差10〜15程度に標準化された指標

IQテストは何種類かあって
それぞれ異なる標準偏差を採用している

ただ一般的には平均値0，標準偏差1に合わせることを指す

こうして得られた標準化得点は，特に**Z得点**と呼ばれる

一般的には「標準化得点」といえばZ得点のことを指す

平均をzeroに合わせるから，という説

以後の資料においても，Z得点のことを「標準化得点」と呼びます

なんで偏差値を使うの？

【前提】偏差値を教育場面で使用するのはほぼ日本だけです。

偏差値を使う理由は標準化得点のメリットから説明できる？

模試の得点から入試の合格可能性を判定するときに

- 素点について「何点以上なら合格可能性80%か」を算出したいが，
  模試の難易度も入試の難易度と同じではないので難しい
- 一方標準化得点は難易度に関わらず「集団内での位置」を表してくれる

  例｜理論上は，標準化得点が1の人は全受験者の上位16%になる

→ 難易度によらず「偏差値いくつ以上なら80%か」を考えることはできそう

入試で「全受験者の上位何%が合格するか」がわかれば，対応する標準化得点＝偏差値を計算できる

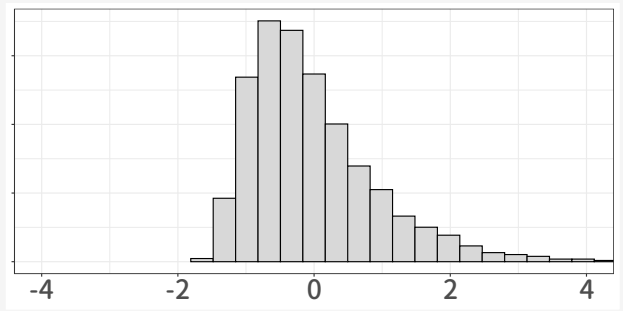
平均値50，標準偏差10にした理由はよくわかりません

理論的には標準化得点のままでも何の問題もないけど…

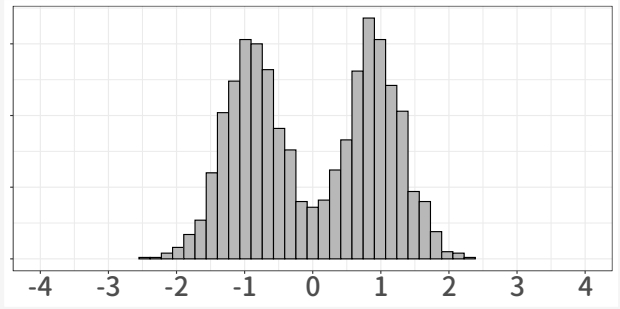
（たぶん）テストはふつう100点満点なので50点が真ん中になるのがしっくり来たから？

もうちょっと記述統計量

平均値と分散（標準偏差）だけでは分布の形はわからない



標準化得点の分布



どちらも
平均0、分散1

→ もともと n 個あったデータなので2つ（平均・分散）だけでは表せない

→ **同じような統計量が n 個あればデータを完全再現できる** (しないけど)

モーメントを使った記述統計量

モーメントとは？

モーメントとはある点を中心とし、物体を回転させる運動を起こす能力の大きさのことです。

物体を回転させようとする力の働きを「モーメント」や「トルク」と表現します。モーメントは機械技術では日常的に使用される用語で、定点から任意の点までの位置ベクトルとベクトル量の積によって表されます。

https://www.oreyume.com/magazine/learn/19702/

平均値は「原点　から値がどれだけ動くか」＝「原点　からの偏差」　　の平均値
分散　は「平均値から値がどれだけ動くか」＝「平均値からの偏差」の二乗の平均値
▶ 同じようにして三乗、四乗と考えてみましょう

もうちょっと記述統計量

平均値は「原点　から値がどれだけ動くか」=「原点　からの偏差」　　の平均値
分散　は「平均値から値がどれだけ動くか」=「平均値からの偏差」の二乗の平均値

歪度 skewness

「平均値からの偏差」の**三乗**の平均値

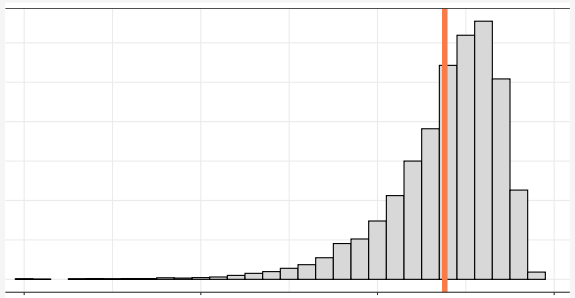
身長	172	168	170	159	176
平均を引く	172−169	168−169	170−169	159−169	176−169
偏差	3	−1	1	−10	7
三乗	27	−1	1	−1000	343

→ 三乗の場合、**もとの偏差と符号がおなじになる**

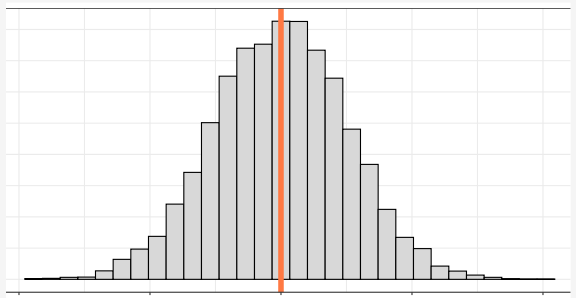
→ **平均値から大きく離れた値があると、歪度はその符号に合わせて動く**

$$\text{歪度} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3$$

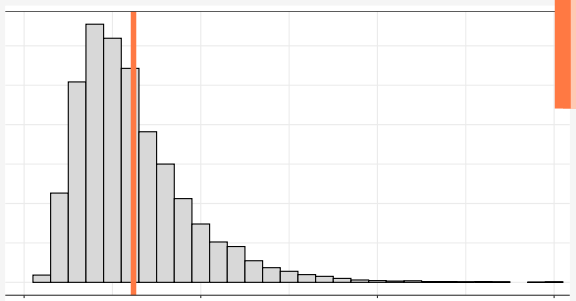
標準化はしてもしなくても良いが、一般的な定義では標準化している



歪度マイナス



歪度ゼロ



歪度プラス

もうちょっと記述統計量

平均値は「原点　から値がどれだけ動くか」=「原点　からの偏差」　　の平均値
分散　は「平均値から値がどれだけ動くか」=「平均値からの偏差」の二乗の平均値
歪度　は「平均値から値がどれだけ動くか」=「平均値からの偏差」の三乗の平均値

尖度 kurtosis

「平均値からの偏差」の**四乗**の平均値

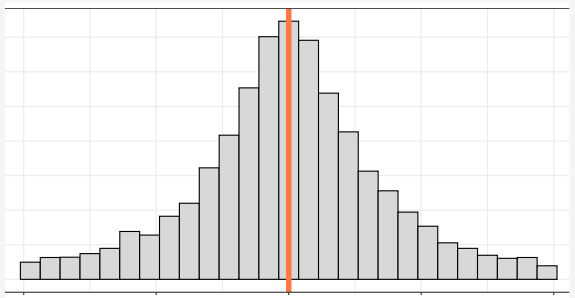
身長	172	168	170	159	176
平均を引く	172−169	168−169	170−169	159−169	176−169
偏差	3	−1	1	−10	7
四乗	81	1	1	10000	2401

→ 四乗の場合、**偏差の影響が分散よりも相当拡大される**

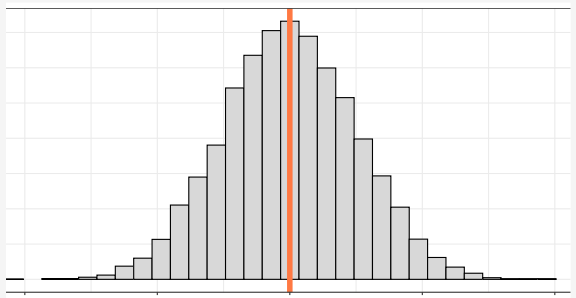
→ **尖度は分散よりも「平均値周辺の集約度」を表しているといえる**

$$\text{尖度} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4$$

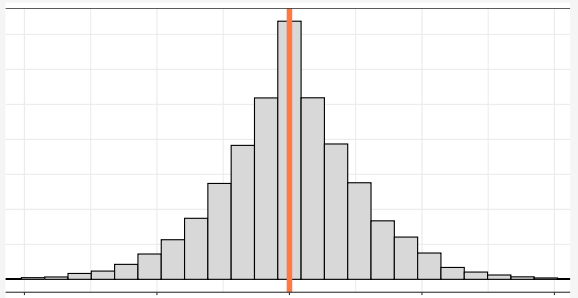
標準化はしてもしなくても良いが、一般的な定義では標準化している



尖度小さめ



尖度ふつう



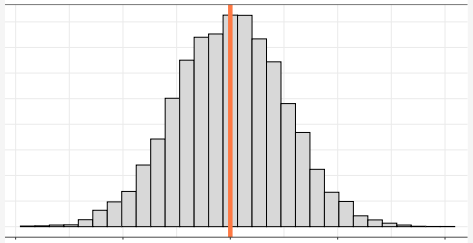
尖度大きめ

そんな統計量どこで使うのさ?

例 品質管理

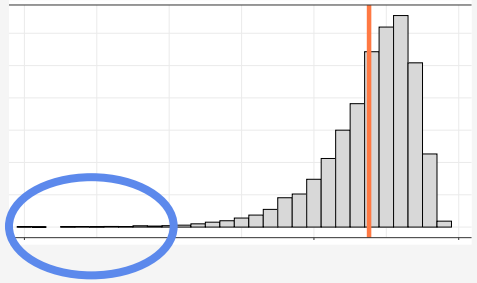
ある製品の出来上がり寸法が…

歪度ゼロ



ならば問題はなさそう
(多少のばらつきは許容される)

歪度マイナス



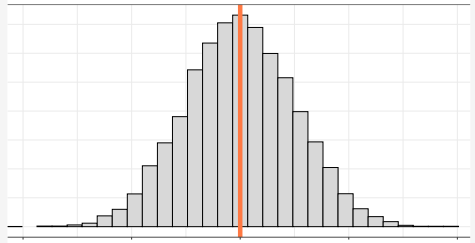
ならば**問題ありかも**
たまに異常に小さいものができる

▶ 製造工程の見直しが必要かも?

例 為替レートの変動

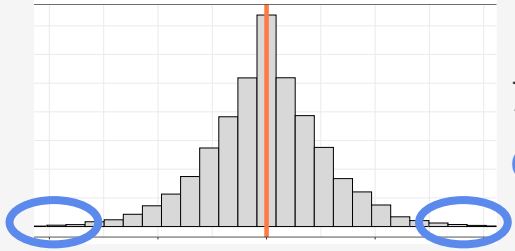
ある通貨の為替レートは…

尖度ふつう



と仮定されていることがあるが

もし実際には **尖度大きめ**



国際テロ,大規模災害,パンデミック…

だとすると,**突発的な暴騰・暴落のリスクを抱えている**と言える

▶ ポートフォリオ設計に役立つかも?