

2-1-4.データのばらつき (分散、標準偏差、偏差値)

2. データリテラシー 2-1. データを読む

ご利用・改変等ご自由です。必要でしたらPPTファイルも差し上げます。ただし国のご支援で作成したものですので、もし講義等でご利用でしたら、利用実績を把握するために、右記までご一報ください (uchida@ait.kyushu-u.ac.jp) 2020.4 内田誠一@九州大学

この項目で学ぶこと

- データの「ばらつき」
 - 3年10組の身長はみんな同じではない
- なぜ「ばらつき」を考える必要があるのか
- データの「ばらつき」の評価法
 - 分散や標準偏差
- 偏差値
 - 「平均からのずれ」を比較するときに便利

この項目を学ぶ意義

- 「ばらつき」は重要
 - 平均などの代表値($\rightarrow 2-3-1$)だけでは見えないデータの性質が見える
 - = 平均にだまされないようにする
- 分散や標準偏差を使えば「ばらつき」が適切に評価できる
 - 「平均からのずれ」を用いて, ばらつきを評価する
 - ただし, 意味や単位の違うものの分散を比べるのは無意味
- 偏差値の計算法がわかる
 - 平均からのずれを標準偏差で割っただけ

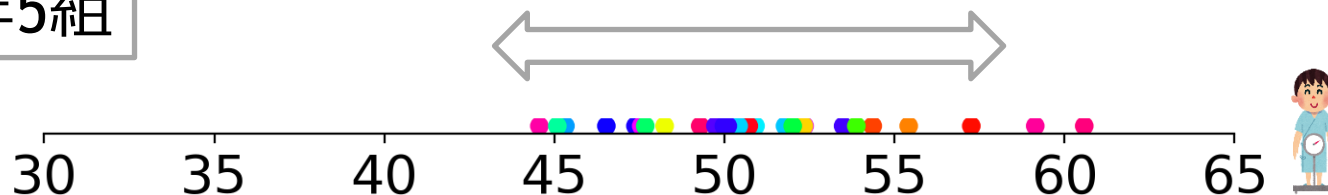
データの「ばらつき」とは何か？

ばらつき = データの広がり具合 = データの変動具合

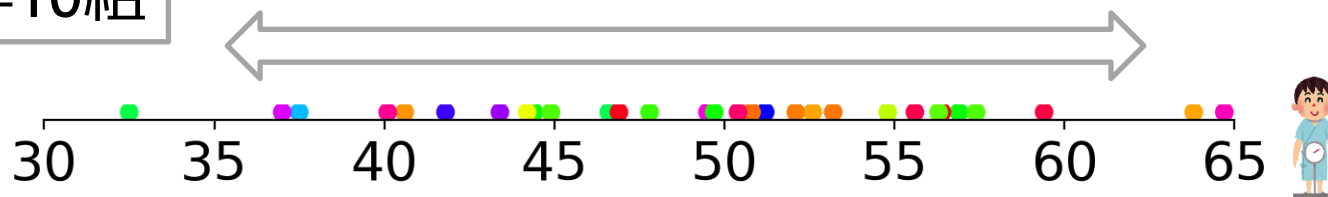
データのばらつき： 体重データの場合

- **ばらつき** = データの**広がり**具合 = データの**変動**具合

3年5組



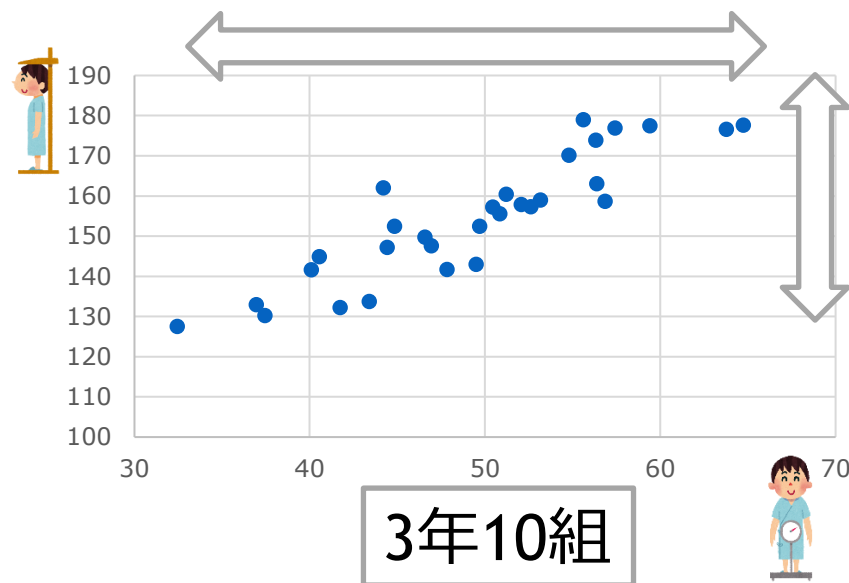
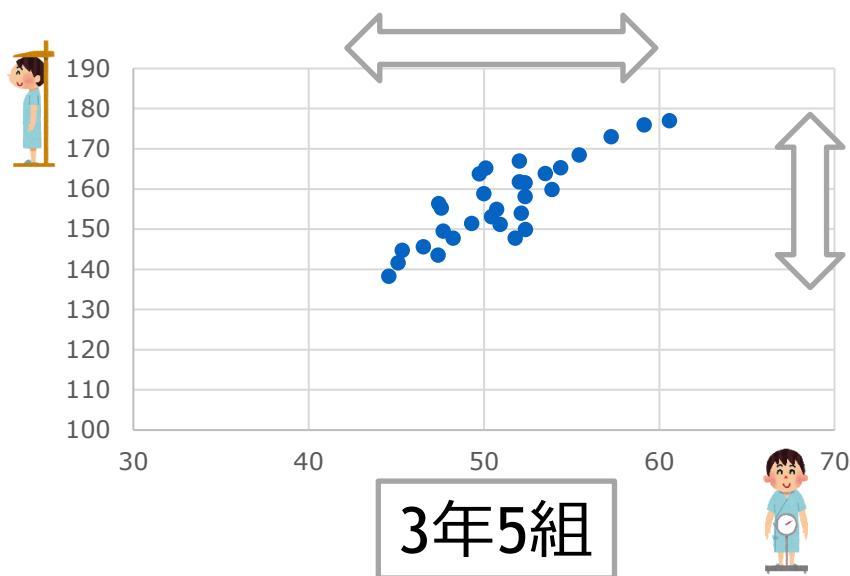
3年10組



- 3年10組のほうが体重が「ばらついて」いる！
- 平均体重はどちらも同じぐらい(約50kg)

データのばらつき： データがペア（体重，身長）の場合

- 散布図(2-1-2)でみる



- 3年10組の方が，身長も体重もばらつきが大きい...
- 平均体重・平均身長はどちらも同じぐらい(約50kg, 155cm)

令和元年度 学校保健統計調査

都道府県別 身長・体重の平均値及び標準偏差 (14歳)

区 分	男	
	身長 (cm)	
	平均値	標準偏差
全 国	165.4	6.72
秋 田	166.9	6.37
青 森	166.7	6.22
神 奈 川	166.4	6.58
富 山	166.2	6.72
⋮		
徳 島	164.9	6.38
福 岡	164.9	6.95
佐 賀	164.9	6.27
⋮		
鹿 児 島	164.5	6.59
宮 崎	164.4	6.73
広 島	164.2	6.99
沖 縄	164.2	6.37

ばらつき
具合※

区 分	女	
	身長 (cm)	
	平均値	標準偏差
全 国	156.5	5.32
秋 田	157.2	5.36
神 奈 川	157.2	5.53
石 川	157.2	5.41
福 井	157.1	5.20
⋮		
山 梨	156.1	5.24
和 歌 山	156.1	5.17
福 岡	156.1	5.24
⋮		
愛 媛	155.6	5.43
広 島	155.6	5.27
島 根	155.5	5.58
沖 縄	155.1	5.40

ばらつき
具合



https://www.mext.go.jp/b_menu/toukei/chousa05/hoken/1268826.htm

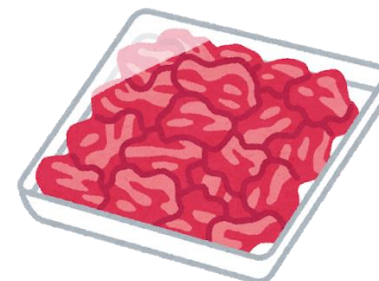
※標準偏差の意味については後述

全国物価統計調査 平成9年全国物価統計調査 大規模店舗編

● 福岡県の牛肉価格だけ取り出してみた

ばらつき
具合

	地域	調査数	平均価格	標準偏差
牛肉, 国産品	福岡県 町村	21	517.2	122.9
	福岡県 福岡	211	530.3	129.2
	福岡県 北九州	139	507.7	148.8
	福岡県 筑豊	35	514.9	181
	福岡県 筑後	52	488.3	143.2
牛肉, 輸入品	福岡県 町村	19	268.1	96.9
	福岡県 福岡	145	304.1	106
	福岡県 北九州	123	266.8	112.5
	福岡県 筑豊	27	251.7	86.6
	福岡県 筑後	42	309.3	122.9



筑豊のスーパー,
国産牛肉の価格の
店ごとの「ばらつき」が
他地域より大きい!
でも輸入牛肉については
他地域より小さい..
不思議...

↓全データはこちらから入手可能
<https://www.e-stat.go.jp/dbview?sid=0000100087>
 ↓最新の物価統計調査も手に入りますよ
<http://www.stat.go.jp/data/kouri/doukou/index.html>

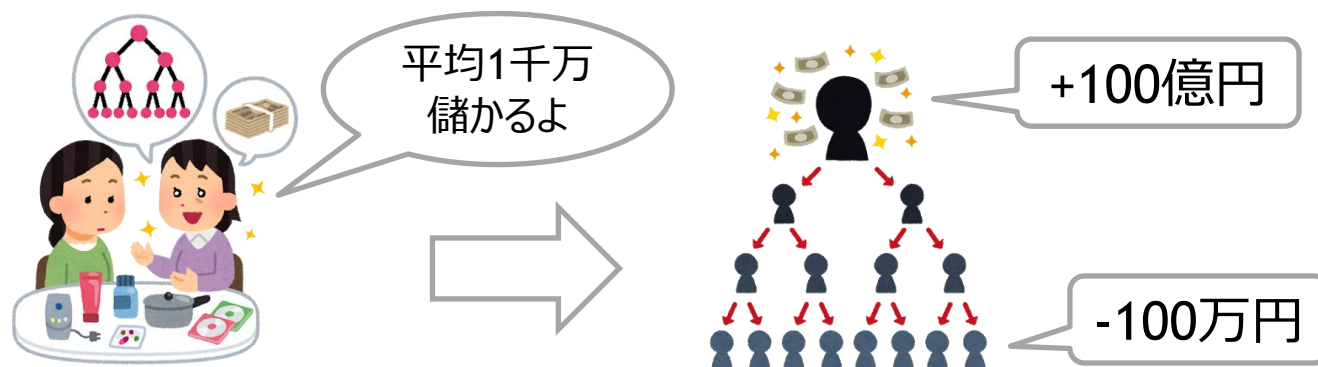
「ばらつき」を考える理由

- 代表値（平均・中央値・最頻値→2-1-2）に加えて、データの「ばらつき」を知りたいことは多い
- 例1: アンケート「この曲好きですか？ 1=大嫌い, 2, 3, 4, 5=大好き」
 - 「平均3で, ばらつき小」→ 大多数が3 → みんな普通 → あまり売れなそう
 - 「平均3で, 1から5までばらつく」→ 好みの別れる曲
 - ↑ 平均は同じでもばらつきが違えば, 状況は全然違う
- 例2: あるダイエット食品で減った体重
 - 「平均ゼロ, ばらつき小」→ 大多数がゼロ → 多くの人には効果なし
 - 「平均ゼロ, -10kgから+10kgまでばらつく」→ 非常に効果ある人もいるが, 逆効果大の人もいる（効果に個人差がありすぎて危険）
 - 「平均 -5kg, ばらつき小」→ 大多数が -5kg → 多くの人に効果あり



「ばらつき」を考える理由： 要するに「代表値だけではわからないことが多い」

- 思い出そう：「平均」を悪用(→ 2-1-3)



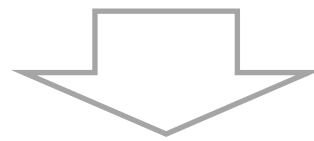
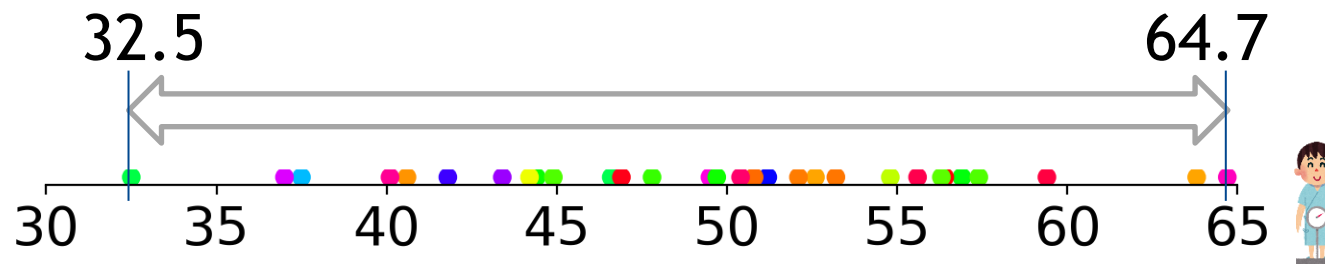
- ウソではないかもしれないが、ごく少数の人だけが莫大な利益を上げ、残り大多数は大損している可能性も

儲けの「ばらつき」を確認すれば「悪用」がわかる！
全員が1千万円儲けているわけではない！



「ばらつき」の測り方：「最大－最小」ではどうか？

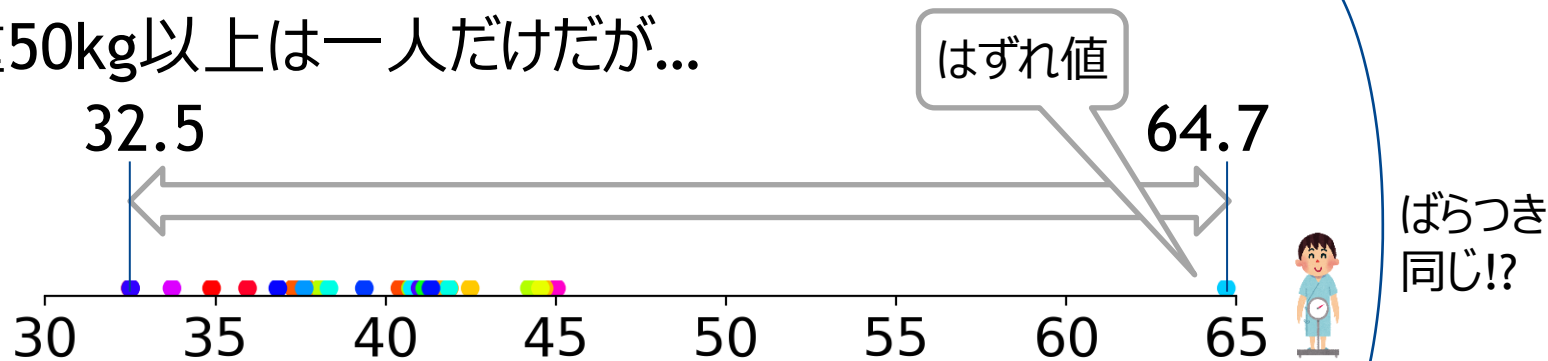
- 3年10組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$



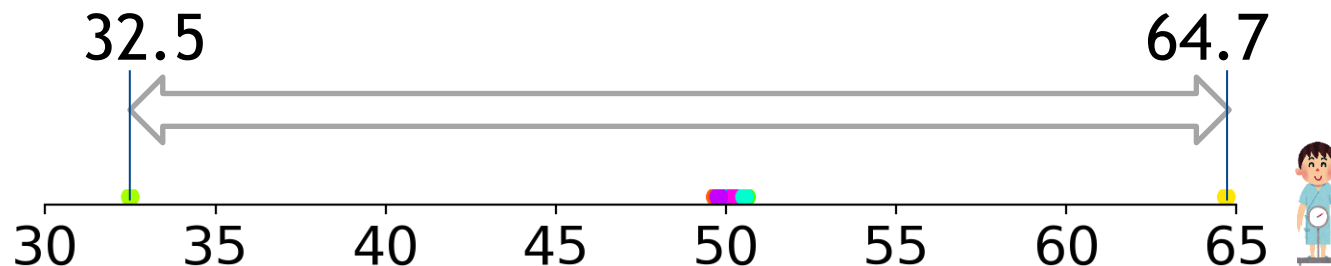
悪くなさそうだけど、
実は「ばらつき」を適切に表現できないケースも

「最大－最小」が「ばらつき」を適切に表現できない例

- 1年Q組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$
- 体重50kg以上は一人だけだが...



- 3年Z組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$
- 30人中28人が50kg付近=ほとんどばらついてないのに...



「最大－最小」ではなぜうまくいかないか、 そして解決へのアイデア

- 「最大－最小」がうまく行かないのは、**たった2つの値(最大値, 最小値)**だけで全データのばらつきを表そうとしているため
 - それら2つが例外的な値であれば、問題が発生
 - さらに悪いことに、最大値・最小値は「例外的」な値になりやすい
- そこで、**全データ**を使って「ばらつき」を計算したい
 - → 「各」データついて「ばらつき」の程度が計算できれば...
 - → そこで「分散」登場！

分散

「平均からのずれ」の平均

分散の考え方

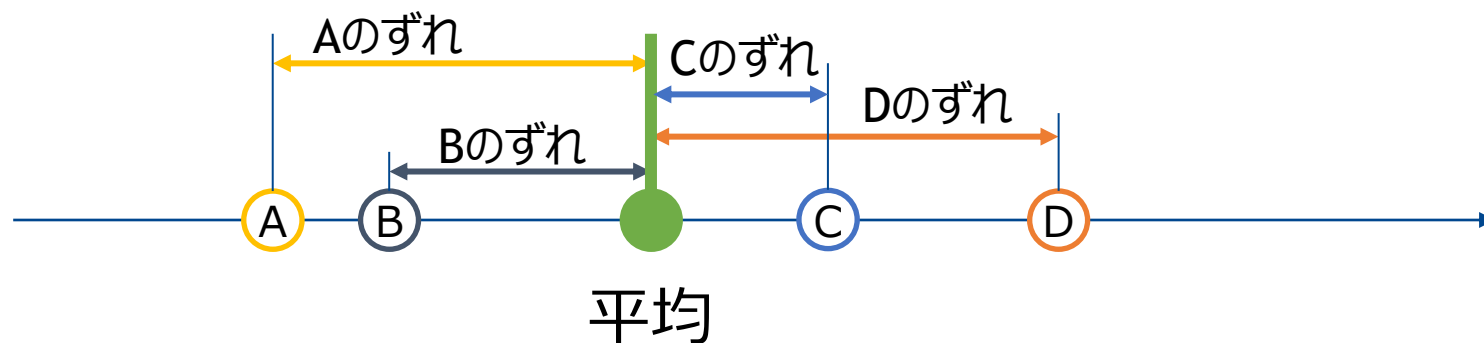
- 全データを使って「ばらつき」を計算したい
- そのために
 1. 全 N 個のデータの「平均」を，全データの「中心」と考える
 2. 各データについて，平均との「ずれ具合」を求める
 3. 全 N 個分の「ずれ具合」の平均を求める

全データが平均的に
「どれぐらい平均からずれているか」

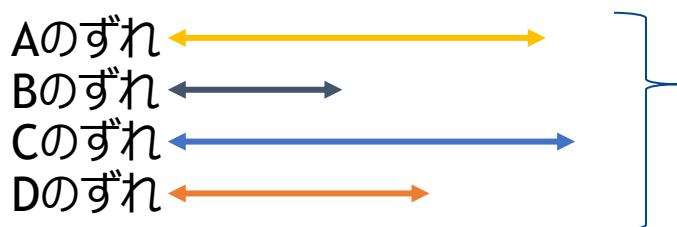


「分散」による「ばらつき」の数値化

- 各データの平均からの「ずれ」



- 分散 = これら全データ分の「ずれ」の平均



これらの「ずれ」の平均=分散

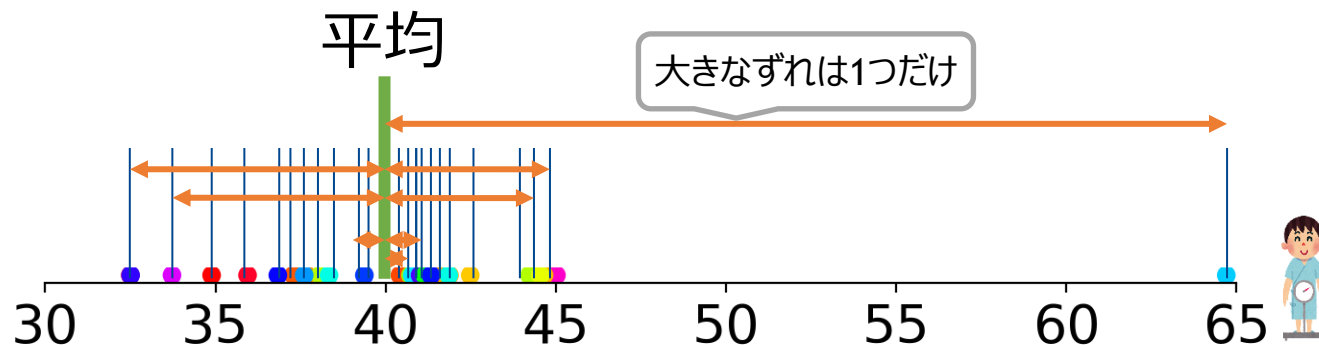
- 「ずれ」の測り方については後ほど

全データが平均的に
「どれぐらい平均からずれているか」

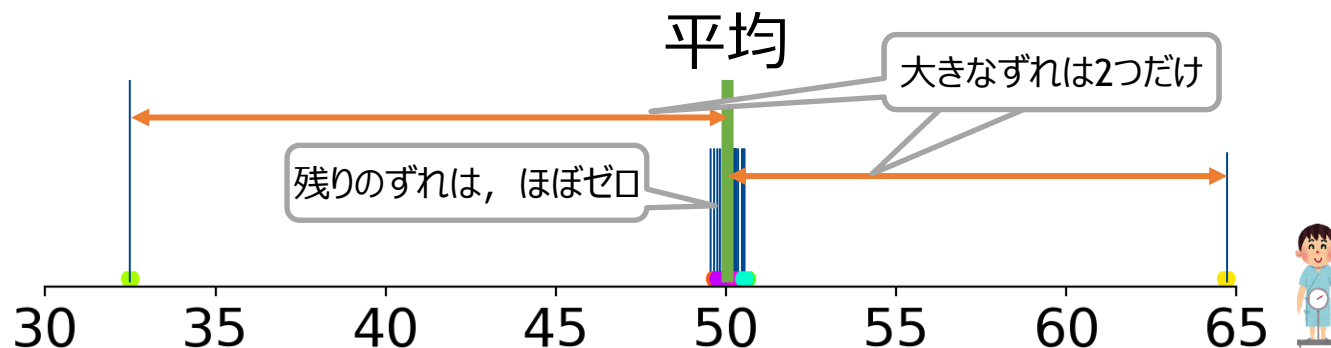


「最大－最小」でうまく行かなかったケースでも、分散ならば適切に「ばらつき」が求まる

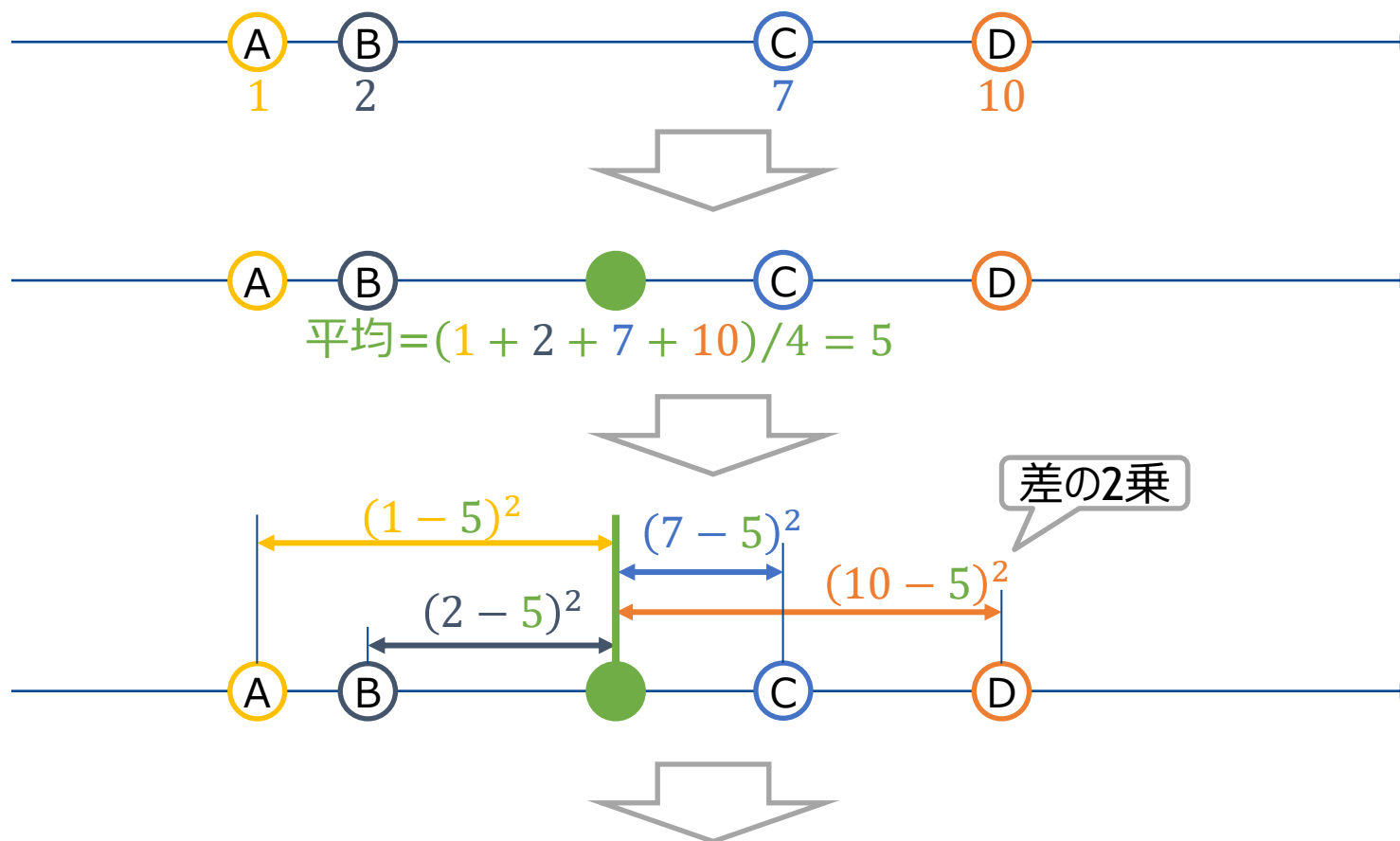
- 1年Q組 → 分散はそう大きくならない



- 3年Z組 → 分散はゼロに近い! → ばらつきは非常に小さい



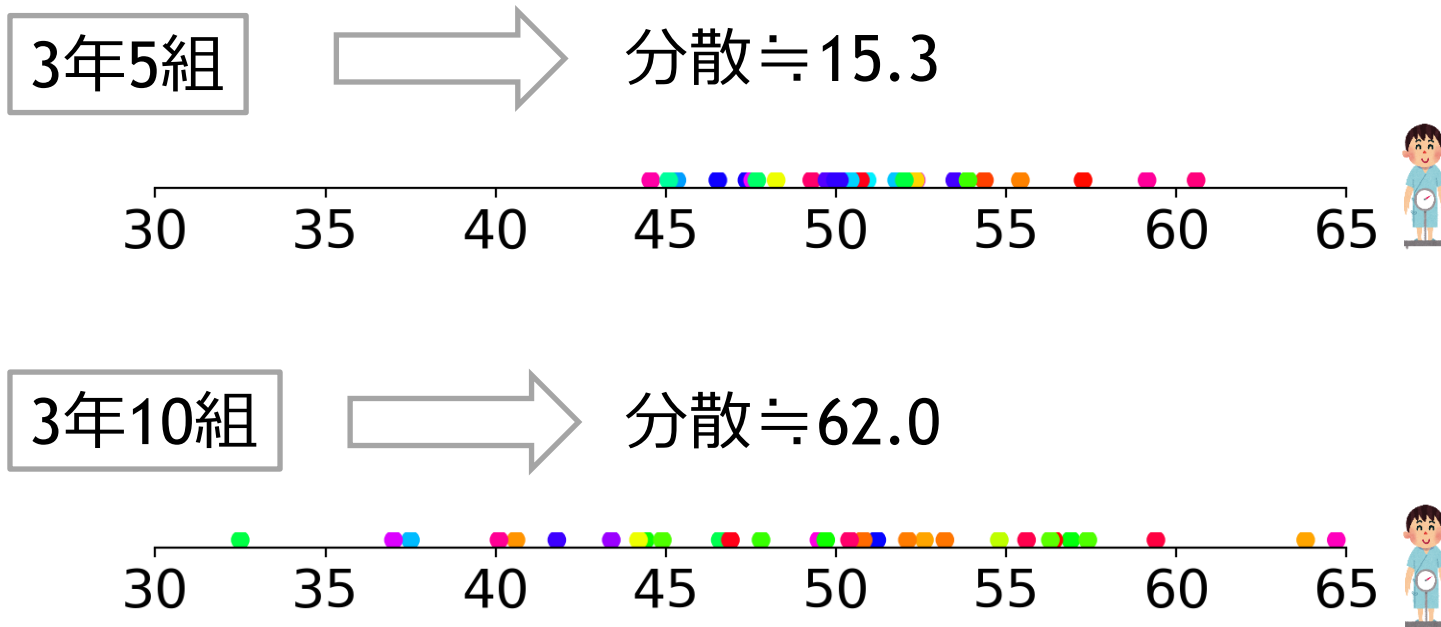
分散における「ずれ」の測り方： 平均との差の「2乗」で測る



$$\begin{aligned} \text{分散} &= ((1 - 5)^2 + (2 - 5)^2 + (7 - 5)^2 + (10 - 5)^2) / 4 \\ &= (16 + 9 + 4 + 25) / 4 = 13.5 \end{aligned}$$

分散を求めてみる

- 平均はどちらも50kgぐらいだが...



- やはりばらつきの大い3年10組のほうが分散大!

練習

- 1, 1, 1, 1, 1 の分散は？

- 1, 5, 4, 2, 8 の分散は？

- 3, 7, 6, 4, 10 の分散は？



全データに+2

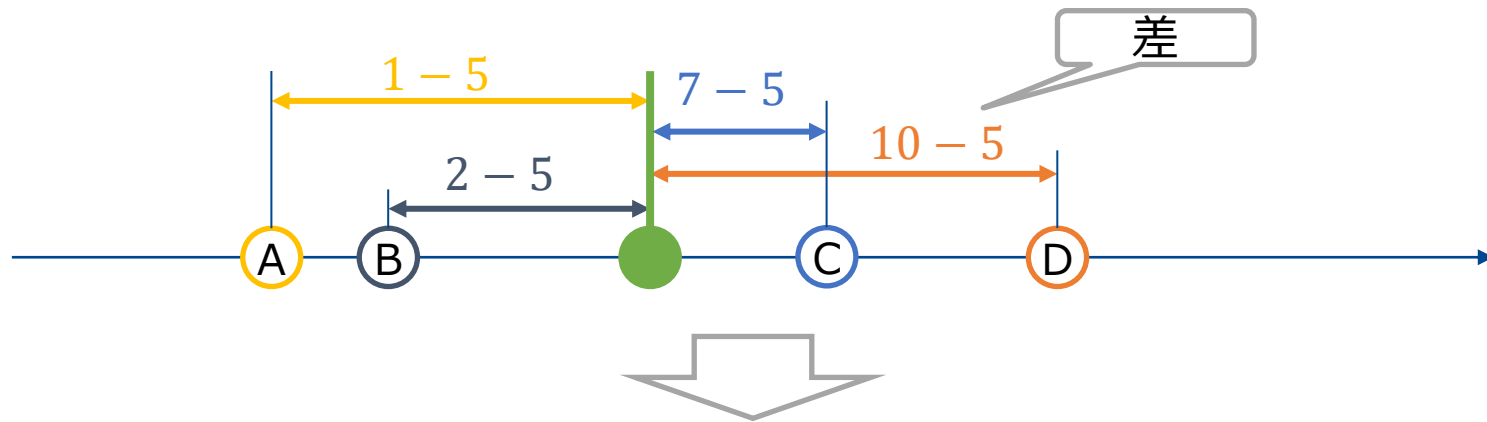
(データ全体が平行移動したということ)

参考：なぜ二乗で「ずれ」を測るのか？ (1/2)

「平均との差」ではだめか？

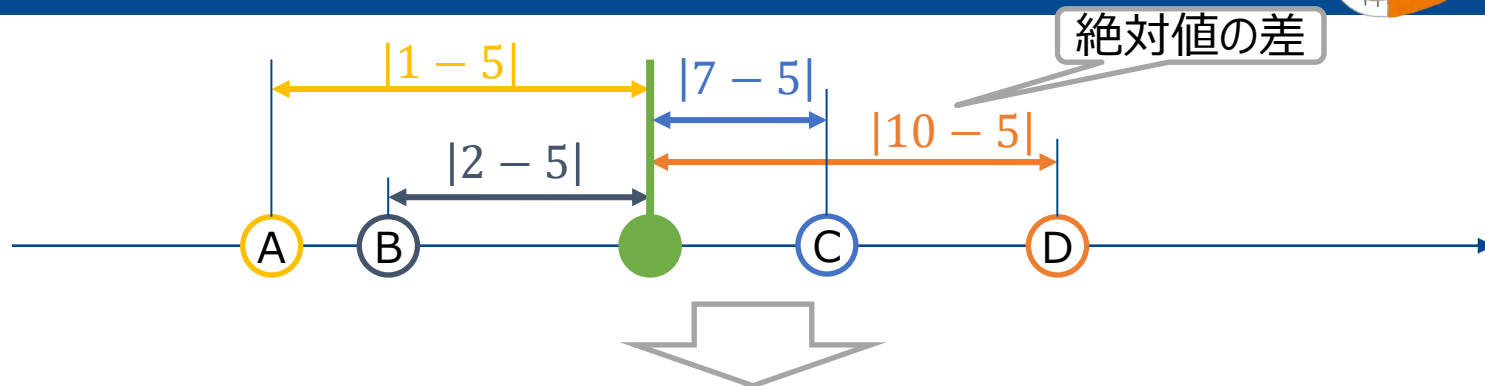


- 「平均との差」をそのまま平均すると，必ずゼロになる



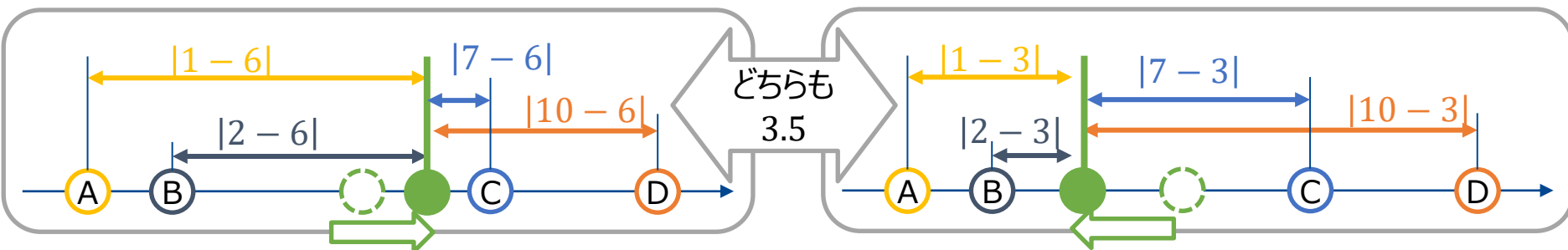
$$\begin{aligned}
 & (1 - 5 + 2 - 5 + 7 - 5 + 10 - 5) / 4 \\
 & = (-4 - 3 + 2 + 5) / 4 \\
 & = 0
 \end{aligned}$$

参考：なぜ二乗で「ずれ」を測るのか？ (2/2)
 ならば、「差の絶対値」では？



$$(|1 - 5| + |2 - 5| + |7 - 5| + |10 - 5|) / 4 = (4 + 3 + 2 + 5) / 4 = 3.5$$

- 別にこれでもOK (「平均偏差」という名前がついている)
- ただ「平均が移動しても変わらない」ので、やや解釈がむずかしい



分散に関する注意

分散は便利な「ばらつき」の指標ですが、
その見かたには注意も必要

分散に関する注意： 身長と体重の分散は比べられるか？ (1/4)

● 3年10組の身長と体重

(49.5, 143.0)	(63.8, 176.6)	(56.4, 163.1)	(64.7, 177.7)	(44.9, 152.5)	(40.1, 141.6)
(46.6, 149.7)	(50.8, 155.6)	(52.1, 157.9)	(56.3, 173.9)	(41.8, 132.3)	(55.6, 179.0)
(56.9, 158.7)	(40.6, 144.9)	(57.4, 176.9)	(54.8, 170.2)	(53.2, 159.0)	(59.4, 177.5)
(47.8, 141.7)	(43.4, 133.8)	(37.5, 130.3)	(44.4, 147.2)	(49.7, 152.5)	(44.2, 162.0)
(51.2, 160.5)	(52.6, 157.3)	(32.5, 127.5)	(37.0, 133.0)	(46.9, 147.7)	(50.4, 157.3)

体重の分散 $\div 62.0$

身長の分散 $\div 237.7$

よって、身長分散のほうが大きい！
...**と言って大丈夫？**

確かに
数字そのものは
大きいけど...

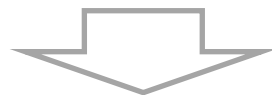


分散に関する注意： 身長と体重の分散は比べられるか？ (2/4)

- 体重をkgではなくg（グラム）で表してみる

(kg,cm)

(49.5, 143.0)	(63.8, 176.6)	(56.4, 163.1)	(64.7, 177.7)	(44.9, 152.5)	(40.1, 141.6)
(46.6, 149.7)	(50.8, 155.6)	(52.1, 157.9)	(56.3, 173.9)	(41.8, 132.3)	(55.6, 179.0)
(56.9, 158.7)	(40.6, 144.9)	(57.4, 176.9)	(54.8, 170.2)	(53.2, 159.0)	(59.4, 177.5)
(47.8, 141.7)	(43.4, 133.8)	(37.5, 130.3)	(44.4, 147.2)	(49.7, 152.5)	(44.2, 162.0)
(51.2, 160.5)	(52.6, 157.3)	(32.5, 127.5)	(37.0, 133.0)	(46.9, 147.7)	(50.4, 157.3)



(g,cm)

(49500, 143.0)	(63800, 176.6)	(56400, 163.1)	(64700, 177.7)	(44900, 152.5)	(40100, 141.6)
(46600, 149.7)	(50800, 155.6)	(52100, 157.9)	(56300, 173.9)	(41800, 132.3)	(55600, 179.0)
(56900, 158.7)	(40600, 144.9)	(57400, 176.9)	(54800, 170.2)	(53200, 159.0)	(59400, 177.5)
(47800, 141.7)	(43400, 133.8)	(37500, 130.3)	(44400, 147.2)	(49700, 152.5)	(44200, 162.0)
(51200, 160.5)	(52600, 157.3)	(32500, 127.5)	(37000, 133.0)	(46900, 147.7)	(50400, 157.3)

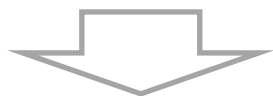
「単位が変わっただけで**体重そのものは変わってない**」点に注意



分散に関する注意： 身長と体重の分散は比べられるか？ (3/4)

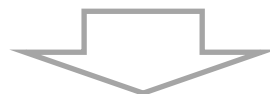
● 3年10組の身長と体重

(49500, 143.0)	(63800, 176.6)	(56400, 163.1)	(64700, 177.7)	(44900, 152.5)	(40100, 141.6)
(46600, 149.7)	(50800, 155.6)	(52100, 157.9)	(56300, 173.9)	(41800, 132.3)	(55600, 179.0)
(56900, 158.7)	(40600, 144.9)	(57400, 176.9)	(54800, 170.2)	(53200, 159.0)	(59400, 177.5)
(47800, 141.7)	(43400, 133.8)	(37500, 130.3)	(44400, 147.2)	(49700, 152.5)	(44200, 162.0)
(51200, 160.5)	(52600, 157.3)	(32500, 127.5)	(37000, 133.0)	(46900, 147.7)	(50400, 157.3)



体重の分散 $\div 62000000$

身長の分散 $\div 237.7$



単位を変えただけで，体重の分散のほうが非常に大きくなった!?

分散に関する注意： 身長と体重の分散は比べられるか？ (4/4)

- 何が問題だったのか？
 - 「同じ数字」であっても，身長・体重のように意味が違うものや $\text{kg} \cdot \text{g}$ のように単位が違うものは，直接比べても意味がない
 - 「10gと10cmは同じ10」や「10kgと10gは同じ10」と言ってる感じ
- 平均についても，「体重の平均値50」と「身長の平均値155」を比べて「身長のほうが大きい！」と言っても意味がない
- 同様に，**体重の分散と身長の分散を比べても意味がない**
 - 3年5組と10組の体重の分散を同じkg単位で比べるのはOK



分散に限らず、「性質が違うものを 数字の上で比べる」トリックには要注意 (1/2)



- 似た者親子の場合

2000万円も出して家を買えるんだから、
10万円の自転車ぐらい買ってよ～!



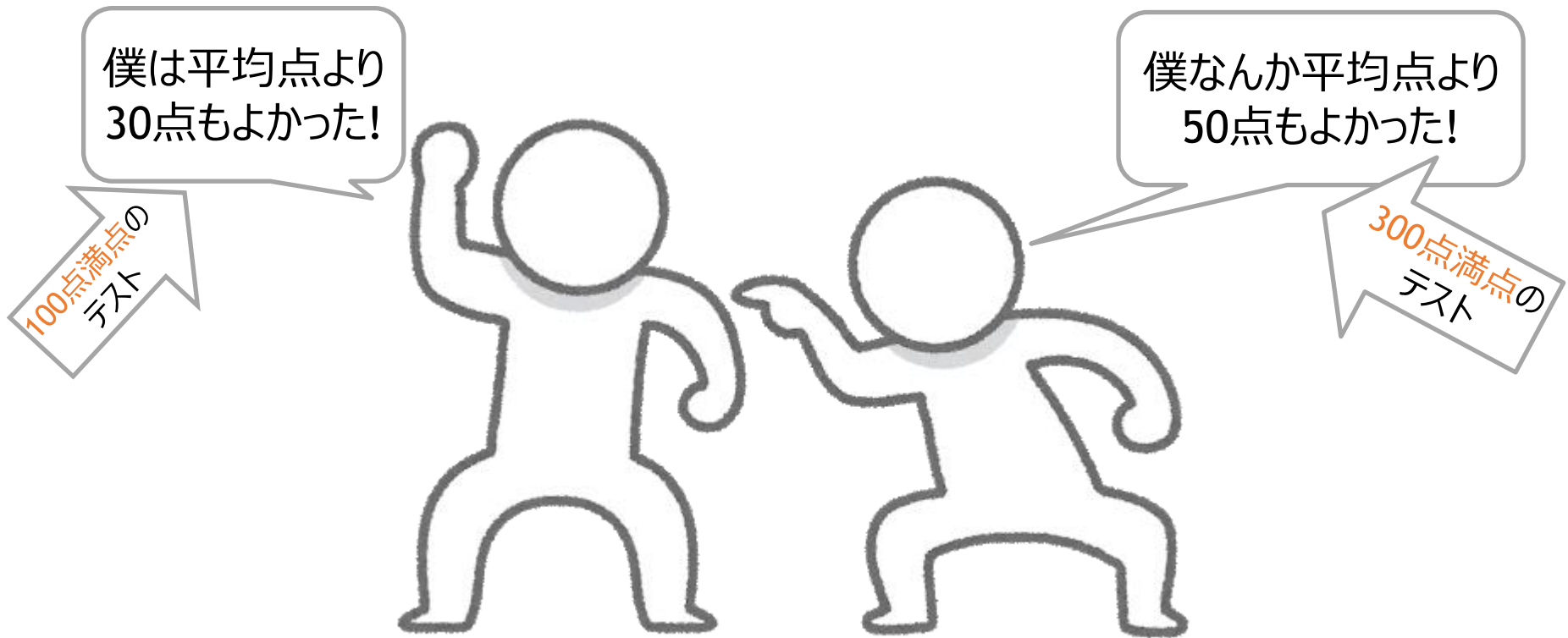
100m競争で1位になれるんだから、
数学のテストで1位になりなさい!



分散に限らず、「性質が違うものを 数字の上で比べる」トリックには要注意 (2/2)



- 全く同じ「テストの点数」ですら...



標準偏差

分散をわかりやすくしたもので、やっぱり「ばらつき」

標準偏差 = 分散の平方根

- 例
 - 分散が100なら, 標準偏差は10
 - 分散が2なら, 標準偏差は $\sqrt{2}$
 - 分散が0なら, 標準偏差も0
- なので, 標準偏差を二乗したら分散
 - 標準偏差が10なら, 分散は100
 - 標準偏差が $\sqrt{2}$ なら, 分散は2

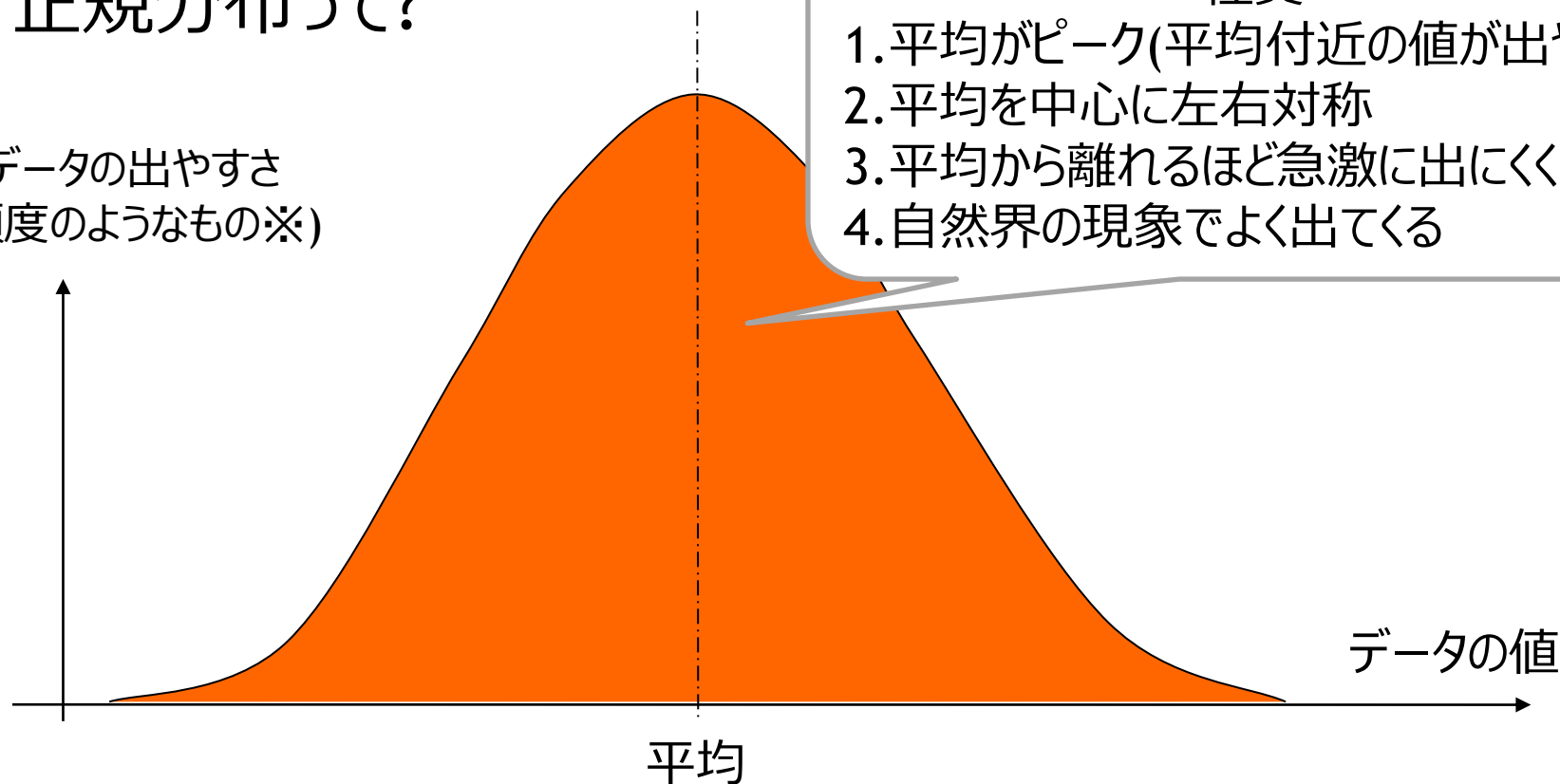
標準偏差は何のためにある？

- 分散が大きい（小さい）ければ標準偏差も大きい（小さい）
 - よって，分散と同じように，全データの「ばらつき」を表す値
- 標準偏差は分散よりもわかりやすい
 - 分散を求める際，二乗して「ずれ」を求めている
 - 体重のデータや平均には50kgのように単位kgがつく
 - しかし分散を求めるときには二乗したので， 62.0kg^2 のような妙な単位に
 - 標準偏差は，その平方根なので，より「ずれ」をシンプルに表す
 - 単位は再びkgに戻る （ $62.0\text{kg}^2 \rightarrow \sqrt{62.0}\text{kg}$ ）
 - なので， $50\text{kg} \pm \sqrt{62.0}\text{kg}$ のような感じで，ばらつきの範囲を使う際にも使える！

参考：データの分布(ヒストグラム→2.1.2のようなもの)が正規分布の場合、標準偏差が大活躍(1/3)

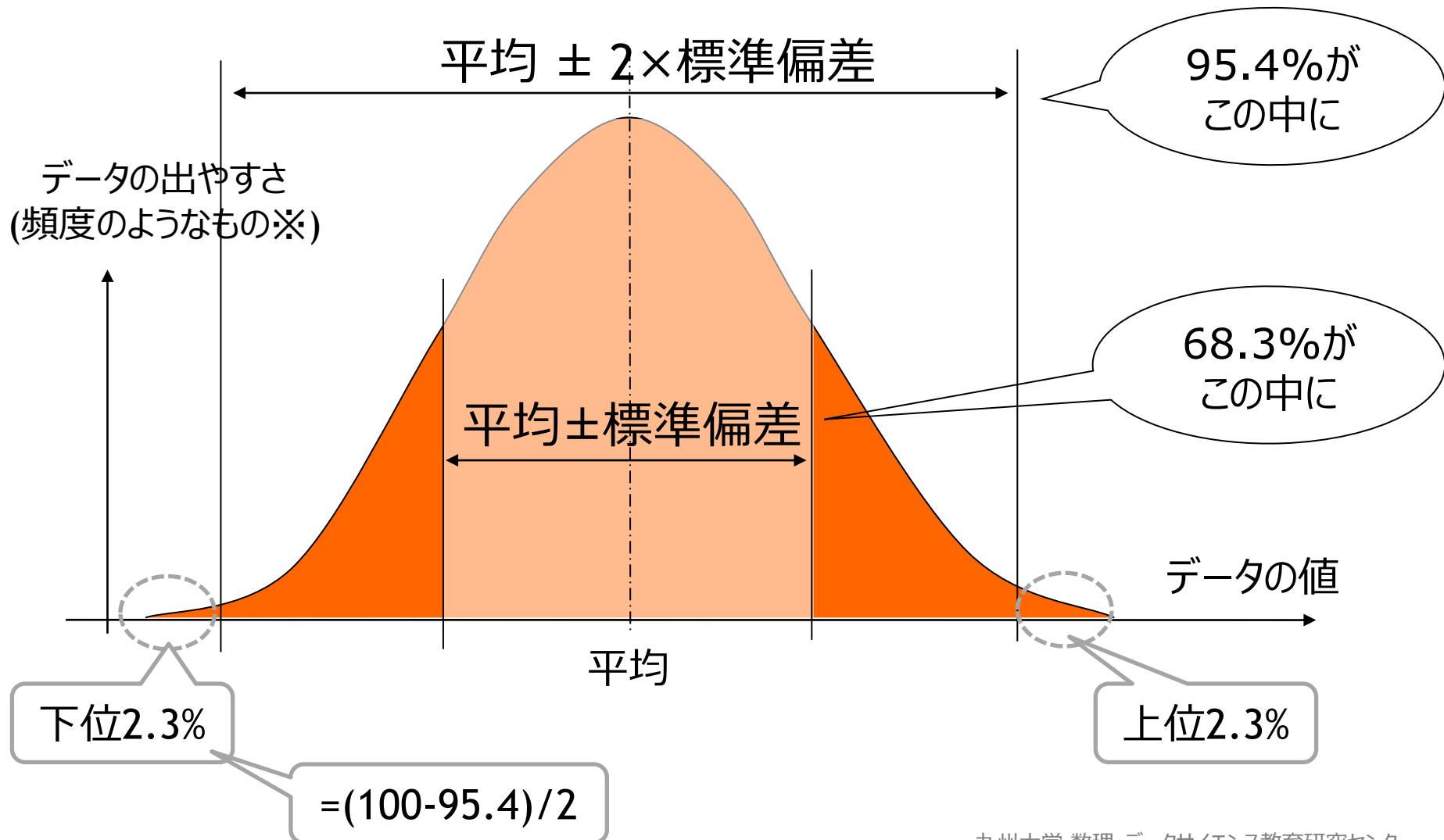
● 正規分布って？

データの出やすさ
(頻度のようなもの※)



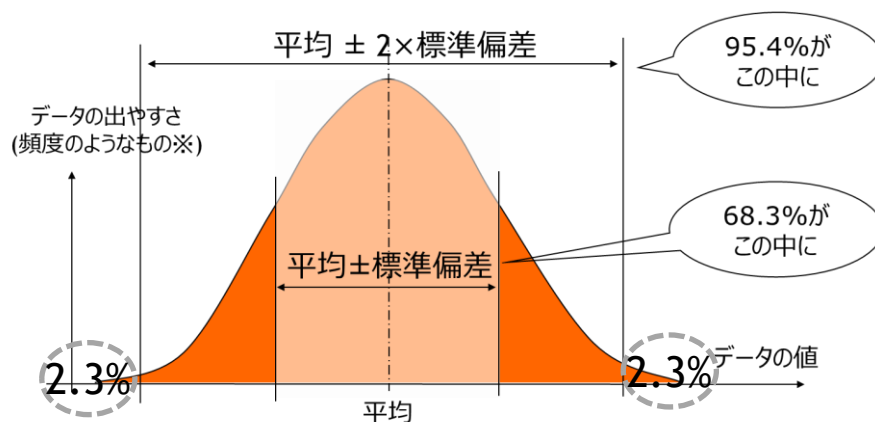
※正確には「確率密度」と呼ばれる

参考：データの分布(ヒストグラム→2.1.2のようなもの)が 正規分布の場合，標準偏差が大活躍(2/3)



参考：データの分布(ヒストグラム→2.1.2のようなもの)が正規分布の場合，標準偏差が大活躍(3/3)

- もしテストの点数が正規分布だったとして，
 - 平均点=50点
 - 標準偏差=10点 (すなわち分散が100)とすると，
 - 75点取っていたら，平均+2×標準偏差=70点より上なので，上位2.3%に入っている！
 - 85点だと，平均+3×標準偏差=80点より上なので，もっと上位！



この辺の話は，
データ群AとBに差があるか
どうかをチェックする時に使われる
「統計的検定」でも
大活躍

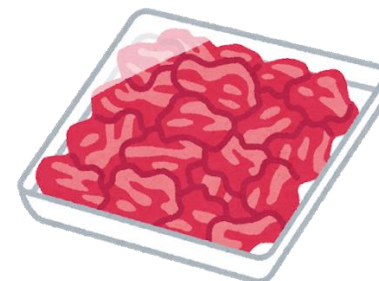
再掲：全国物価統計調査

平成9年全国物価統計調査 大規模店舗編

● 福岡県の牛肉価格だけ取り出してみた

ここ出てます

	地域	調査数	平均価格	標準偏差
牛肉, 国産品	福岡県 町村	21	517.2	122.9
	福岡県 福岡	211	530.3	129.2
	福岡県 北九州	139	507.7	148.8
	福岡県 筑豊	35	514.9	181
	福岡県 筑後	52	488.3	143.2
牛肉, 輸入品	福岡県 町村	19	268.1	96.9
	福岡県 福岡	145	304.1	106
	福岡県 北九州	123	266.8	112.5
	福岡県 筑豊	27	251.7	86.6
	福岡県 筑後	42	309.3	122.9



もし筑豊地方のスーパーの牛肉価格が正規分布しているなら、68.3%のスーパーが 514.9 ± 181 円で国産牛を売ってます

↓ 全データはこちらから入手可能

<https://www.e-stat.go.jp/dbview?sid=0000100087>

↓ 最新の物価統計調査も手に入りますよ

<http://www.stat.go.jp/data/kouri/doukou/index.html>

偏差値

状況が違ってても、どれぐらいずれているかを比べたい

A君は〇〇模試で80点, B君が△△模試で80点,
どちらも自分がどれぐらいスゴイのか知りたい

- A君の受けた〇〇模試の平均点=60点



- B君の受けた△△模試の平均点=70点



- そうすると平均より20点上だった, A君のほうがスゴそう

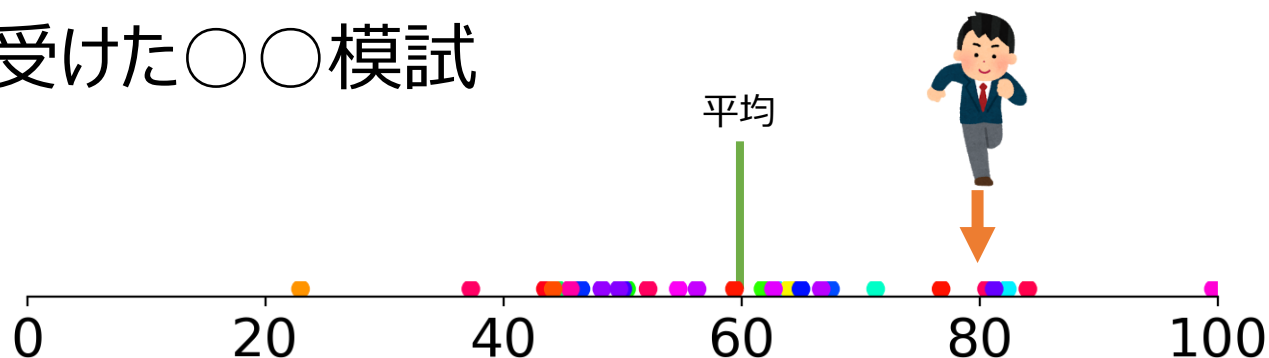


- でも, もし, △△模試のほぼ全員が70点ぐらいだったら, その状況で80点とったB君のほうがすごくない?

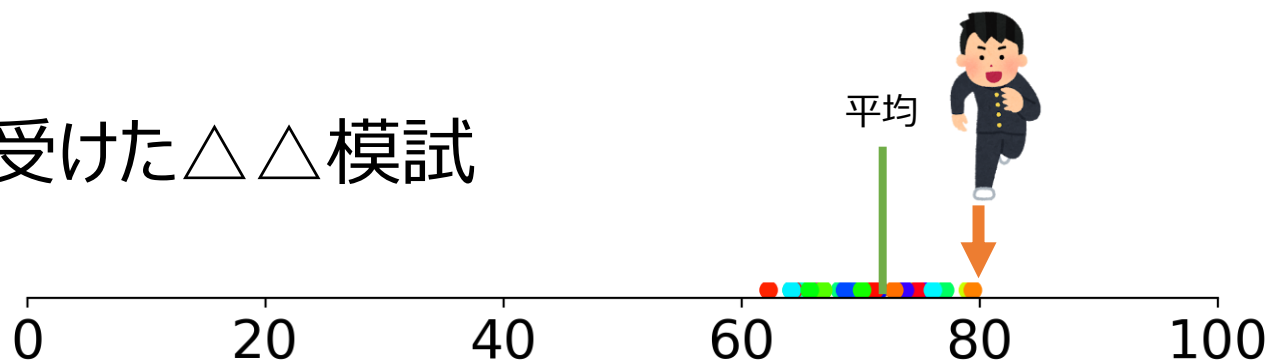


点数の分布を見てみると...

- A君の受けた〇〇模試



- B君の受けた△△模試



- 確かに, B君のほうがスゴイかも...

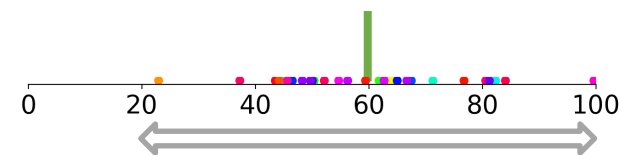
ではどうやって二人のすごさを比較するか？

- 平均との差だけではダメっぽい

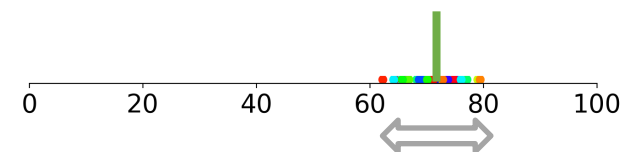
- そこで、「ばらつき」具合を使う！



- ○○模試のように、「ばらつき」が大きい
→ 平均との差を小さめに評価
= 平均と結構違っても「あまり変わらない」



- △△模試のように、「ばらつき」が小さい
→ 平均との差を大きめに評価
= 平均と少しでも違えば「すごく違う」



実際にはどうする？

「平均からのずれ」を標準偏差で割ればOK

$$\frac{\text{データの値} - \text{平均}}{\text{標準偏差}}$$

平均からのずれ

ばらつき具合

- ○○模試の標準偏差=20 (ばらつき大)
- △△模試の標準偏差=5 (ばらつき小)
- A君の値 = $(80-60)/20 = 1$
- B君の値 = $(80-70)/5 = 2 \rightarrow$ B君のほうが, より平均から離れている!

偏差値



$$\text{偏差値} = 10 \times \frac{\text{データの値} - \text{平均}}{\text{標準偏差}} + 50$$

- ○○模試の標準偏差=20 (ばらつき大)
- △△模試の標準偏差=5 (ばらつき小)

- A君の値 = $10 \times (80 - 60) / 20 + 50 = 60$
- B君の値 = $10 \times (80 - 70) / 5 + 50 = 70$

B君のほうが
偏差値高い



偏差値の性質



$$\text{偏差値} = 10 \times \frac{\text{データの値} - \text{平均}}{\text{標準偏差}} + 50$$

- 平均点と同じなら偏差値50
- 標準偏差が小さい(ばらつきが小さい)ほど、データの少しの変化が大きく影響する

まとめ

- 代表値だけでなく、「ばらつき」は重要
- 分散や標準偏差の計算法
 - 分散 = 「平均からのずれ」の平均
 - 標準偏差 = $\sqrt{\text{分散}}$
- 「平均からのずれ」を「標準偏差」で割ることで、ばらつきの異なるデータ間でも、平均からのずれ具合を適切に評価できる
 - これが偏差値

付録：本資料で用いた体重-身長データ

(※乱数で生成したものであり，実データではありません)

3年5組

49.28898	151.4737
52.33736	158.1474
50.9156	151.2327
52.12094	153.9946
44.55929	138.3027
45.32656	144.7388
54.36184	165.2948
52.35728	149.9001
47.41078	156.3905
50.7175	154.937
53.49103	163.8364
47.38458	143.5537
55.41912	168.531
50.41957	153.0828
52.01225	166.9758
51.7823	147.8139
47.55464	155.2434
52.33595	161.5521
48.24293	147.7574
60.58072	177.0258
50.10005	165.2525
45.08798	141.6259
47.67298	149.5468
49.7364	163.772
59.14007	175.9572
57.26009	172.9919
49.98545	158.8914
53.87983	159.9219
52.00903	161.7652
46.52984	145.6229

3年10組

49.51165	143.0219
46.58579	149.7477
56.86254	158.6564
47.8431	141.728
51.22793	160.462
63.77215	176.5847
50.84549	155.5656
40.55279	144.9083
43.40138	133.7944
52.62262	157.3297
56.38814	163.1046
52.07485	157.88
57.41918	176.9334
37.4549	130.2867
32.45199	127.535
64.73897	177.6759
56.33934	173.9304
54.79376	170.1935
44.42531	147.227
36.97103	132.9765
44.85306	152.4797
41.75252	132.259
53.18034	159.0043
49.7005	152.4852
46.94904	147.6628
40.10991	141.6258
55.61957	178.993
59.41709	177.4966
44.20566	162.0014
50.44417	157.259