2-1-2.データの分布と代表値

2. データリテラシー
2-1.データを読む

この項目で学ぶこと

- データの分布
  - ＝どのようなデータがどれぐらい出ているか
  - ＝データがどのように広がっているか
- ヒストグラム
  - データの頻度を可視化する
- 代表値
  - データ全体を分布中心のデータ１つで表す方法
  - 平均値，中央値，最頻値

この項目を学ぶ意義

- 大量のデータに潜む傾向をつかむ基本テクニックが身に付く

- テクニック1: 散布図
  - 1データを1点として数直線や平面上に可視化
  - データの広がりを「目」で見える

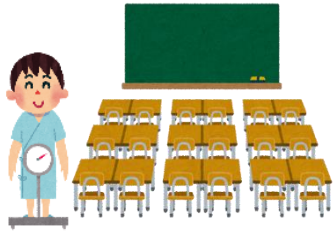
- テクニック2: ヒストグラム
  - 同じ(ような)値を持つデータをまとめて可視化→データ数が膨大でもOK

- テクニック3: 代表値
  - 膨大なデータを全部見なくても，一つの代表値でそこそこ傾向をつかめる
    - ex. クラスA,Bの「平均」が各々60点,75点なら，クラスBに点数が高い学生が多そう(断言は無理だが)

データと分布

たくさんのデータ(数値)の傾向はわかりにくい

- ex. 3年10組の30人の体重

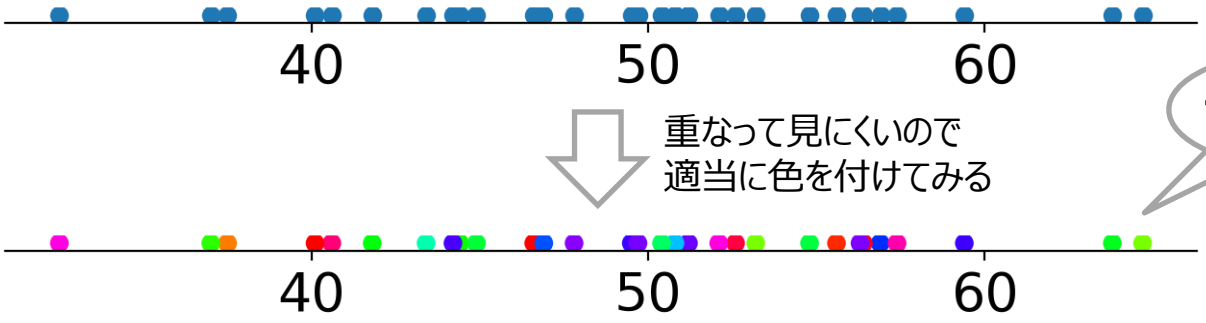


49.5	63.8	56.4	64.7	44.9	40.1
46.6	50.8	52.1	56.3	41.8	55.6
56.9	40.6	57.4	54.8	53.2	59.4
47.8	43.4	37.5	44.4	49.7	44.2
51.2	52.6	32.5	37.0	46.9	50.4

データを数直線上に並べてみる

49.5	63.8	56.4	64.7	44.9	40.1
46.6	50.8	52.1	56.3	41.8	55.6
56.9	40.6	57.4	54.8	53.2	59.4
47.8	43.4	37.5	44.4	49.7	44.2
51.2	52.6	32.5	37.0	46.9	50.4

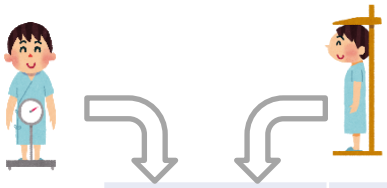
40～60kgが多い感じ?



- たくさんのデータの広がり方（＝どの辺の値がどれぐらい出やすいか）を「分布」と言います

データと分布：
データの「ペア」の場合

- ex. 3年10組の30人の（体重，身長）のペア

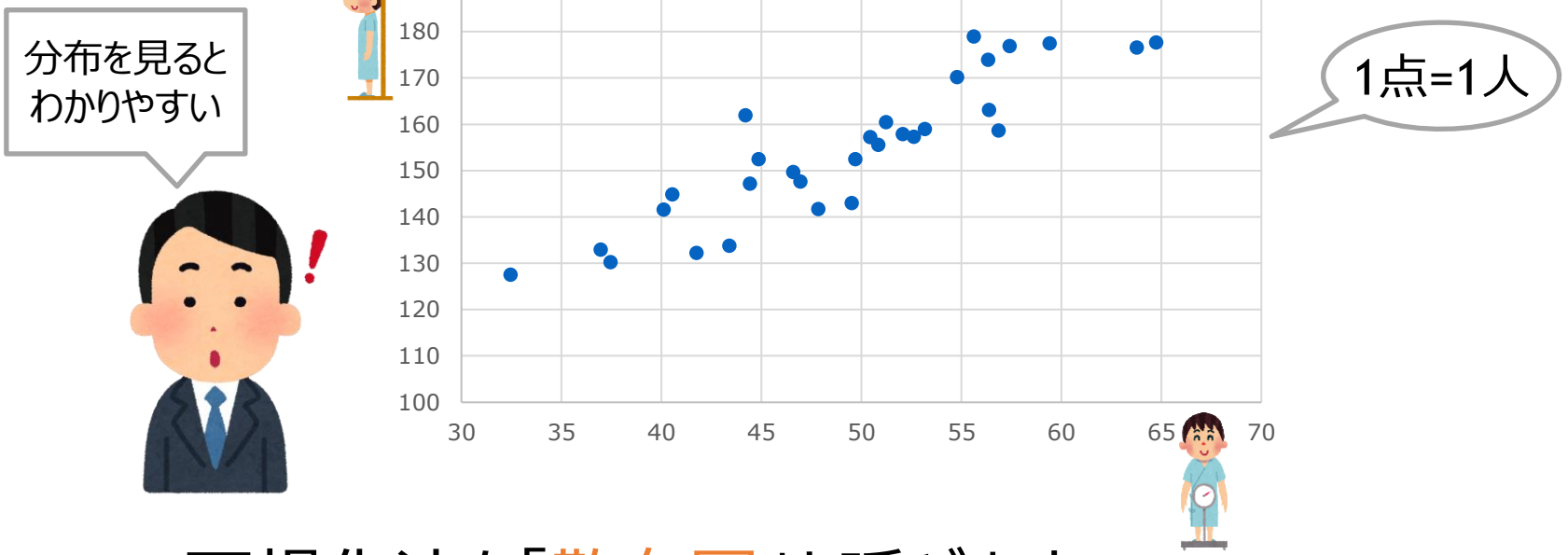


(49.5, 143.0)	(63.8, 176.6)	(56.4, 163.1)	(64.7, 177.7)	(44.9, 152.5)	(40.1, 141.6)
(46.6, 149.7)	(50.8, 155.6)	(52.1, 157.9)	(56.3, 173.9)	(41.8, 132.3)	(55.6, 179.0)
(56.9, 158.7)	(40.6, 144.9)	(57.4, 176.9)	(54.8, 170.2)	(53.2, 159.0)	(59.4, 177.5)
(47.8, 141.7)	(43.4, 133.8)	(37.5, 130.3)	(44.4, 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)

データと分布：
データの「ペア」の場合

- ex. 3年10組の30人の（体重，身長）のペア

(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)



- この可視化法を「散布図」と呼びます

ヒストグラム

頻度を目に見えるようにする

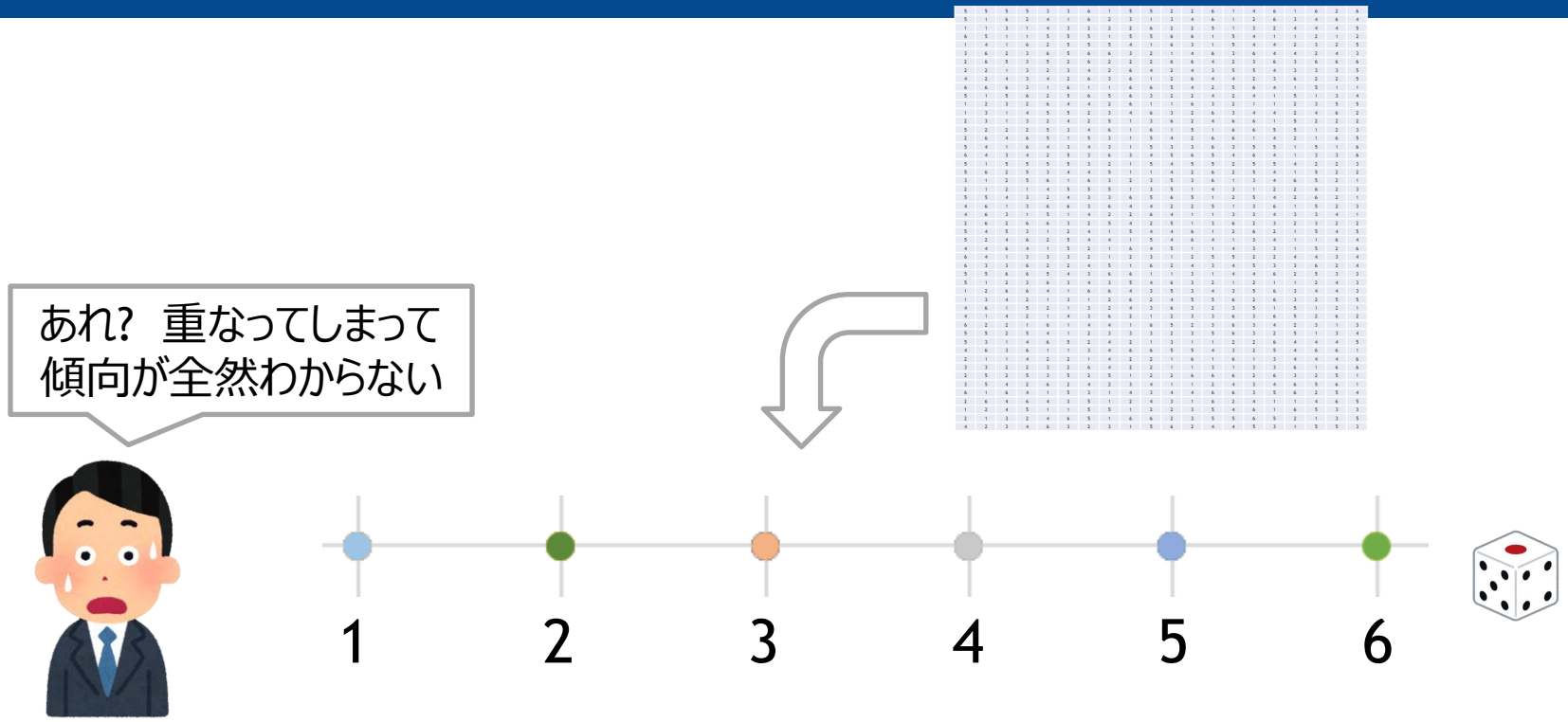
たくさんのデータ(数値)の傾向はわかりにくい

- さて，今度はサイコロを1000回振ってみた

5	5	5	5	3	3	6	1	5	5	2	2	6	1	4	6	1	6	2	6
5	1	6	2	4	1	6	2	3	1	3	4	6	1	2	6	3	4	6	4
1	1	3	1	4	3	2	2	2	6	2	2	5	1	3	2	4	4	4	5
6	5	1	1	5	5	5	1	5	5	6	6	1	5	4	1	1	2	1	2
1	4	1	6	2	5	5	5	4	1	6	3	1	5	4	4	2	3	2	5
3	6	2	3	6	5	6	6	3	2	1	4	6	3	6	4	4	2	4	3
2	6	5	3	5	2	6	2	2	2	6	6	4	2	3	6	3	6	6	6
2	2	1	3	2	3	4	2	6	4	2	4	3	5	5	4	3	3	3	5
4	2	4	3	4	2	6	3	6	1	2	6	4	4	2	3	6	2	2	5
6	6	6	3	1	6	1	1	6	6	5	4	2	5	6	4	1	5	1	1
5	1	5	6	2	5	6	5	6	3	2	2	4	2	4	1	5	1	3	4
1	2	3	2	6	4	4	2	6	1	1	6	3	2	1	1	2	3	5	5
1	3	1	4	5	5	2	3	4	6	3	2	6	3	4	4	2	4	6	2
2	3	1	3	2	4	2	5	1	3	6	2	4	6	6	1	5	2	2	2
5	2	2	2	5	3	4	6	1	6	1	5	1	6	6	5	5	1	2	3
2	6	4	6	5	1	5	3	1	5	4	2	6	6	1	4	2	1	6	5
5	4	1	6	4	3	4	3	1	5	3	3	6	3	5	5	1	5	1	6
6	4	3	4	2	5	3	6	3	4	5	6	5	4	6	4	1	3	3	6
5	1	5	5	5	5	3	2	1	5	4	5	5	2	5	5	4	2	2	3
5	6	2	5	3	4	4	5	1	1	4	2	6	2	5	4	1	5	2	2
3	1	2	5	6	1	6	3	2	3	5	3	6	1	3	4	6	5	2	1
2	1	2	1	4	5	5	5	1	3	5	1	4	3	1	2	2	6	2	3
5	5	4	3	2	4	3	3	6	5	6	5	1	2	5	4	2	6	2	1
4	6	1	3	6	6	3	6	4	4	2	2	5	1	3	6	1	5	2	3
4	6	3	1	5	1	4	2	2	6	4	1	1	3	3	4	3	3	4	1
2	6	2	6	6	3	2	5	4	2	5	1	3	6	2	3	2	3	2	2
5	4	5	3	1	2	4	1	5	4	4	6	1	2	6	2	1	5	4	5
5	2	4	6	2	5	4	4	1	5	4	6	4	1	3	4	1	1	6	4
4	4	6	4	1	5	2	1	6	4	5	1	1	4	3	3	1	5	2	6
6	4	1	3	3	3	2	1	2	3	1	2	5	5	2	2	4	4	3	4
6	3	3	6	2	2	4	5	1	6	2	4	3	4	5	3	3	6	2	4
5	5	6	6	5	4	3	6	6	1	1	3	1	4	4	6	2	5	3	3
5	1	2	3	6	3	4	3	5	4	6	3	2	1	2	1	1	2	4	3
1	2	6	6	4	1	6	6	4	3	5	3	4	3	5	6	3	4	4	3
1	3	4	2	1	3	1	2	6	2	4	5	5	6	2	6	3	2	5	5
4	6	1	5	2	1	3	2	4	3	6	3	2	3	5	1	5	1	2	1
4	1	4	2	1	4	3	6	2	1	2	3	3	6	3	6	5	2	6	2
6	2	2	1	6	1	4	4	1	6	5	2	3	6	3	4	2	3	1	3
5	5	2	5	4	1	2	3	3	3	2	3	5	6	3	2	5	1	3	4
5	3	1	4	6	5	2	4	2	1	3	1	1	2	2	6	4	4	4	5
4	6	3	6	1	1	3	4	6	6	5	5	4	3	2	5	4	6	6	1
2	1	1	4	2	2	1	4	2	2	1	6	1	6	1	3	4	4	4	6
3	3	2	2	3	2	6	4	2	2	1	1	3	1	3	3	6	1	6	6
2	5	2	5	3	5	2	5	1	2	2	6	6	6	2	6	3	2	5	1
3	5	4	2	6	2	4	2	3	4	1	1	2	4	3	4	6	5	6	1
6	1	6	4	1	5	3	1	4	3	4	4	6	6	3	5	6	2	5	4
2	6	4	6	4	3	5	1	2	4	3	1	6	2	4	1	1	4	6	5
1	2	4	5	1	1	5	5	1	2	2	3	5	4	6	1	6	5	3	3
2	1	3	2	4	6	5	1	6	6	2	3	5	5	6	5	2	1	3	5
4	2	3	4	6	3	2	3	1	5	6	2	4	4	5	3	1	5	5	3

特定の目が出やすい
「ズルいサイコロ」
じゃないかチェックしよう

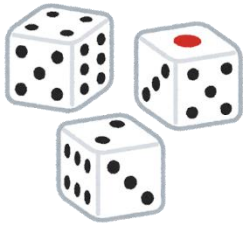
データを数直線上に並べてみる



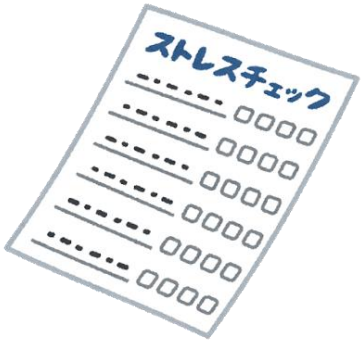
- 全く同じデータが何度も出る場合，この方法で分布はわからない

そこで「頻度」を考える

- さいころを1000回振って出た目の回数
  - 「1」が168回，「2」が164回，…，「6」が164回

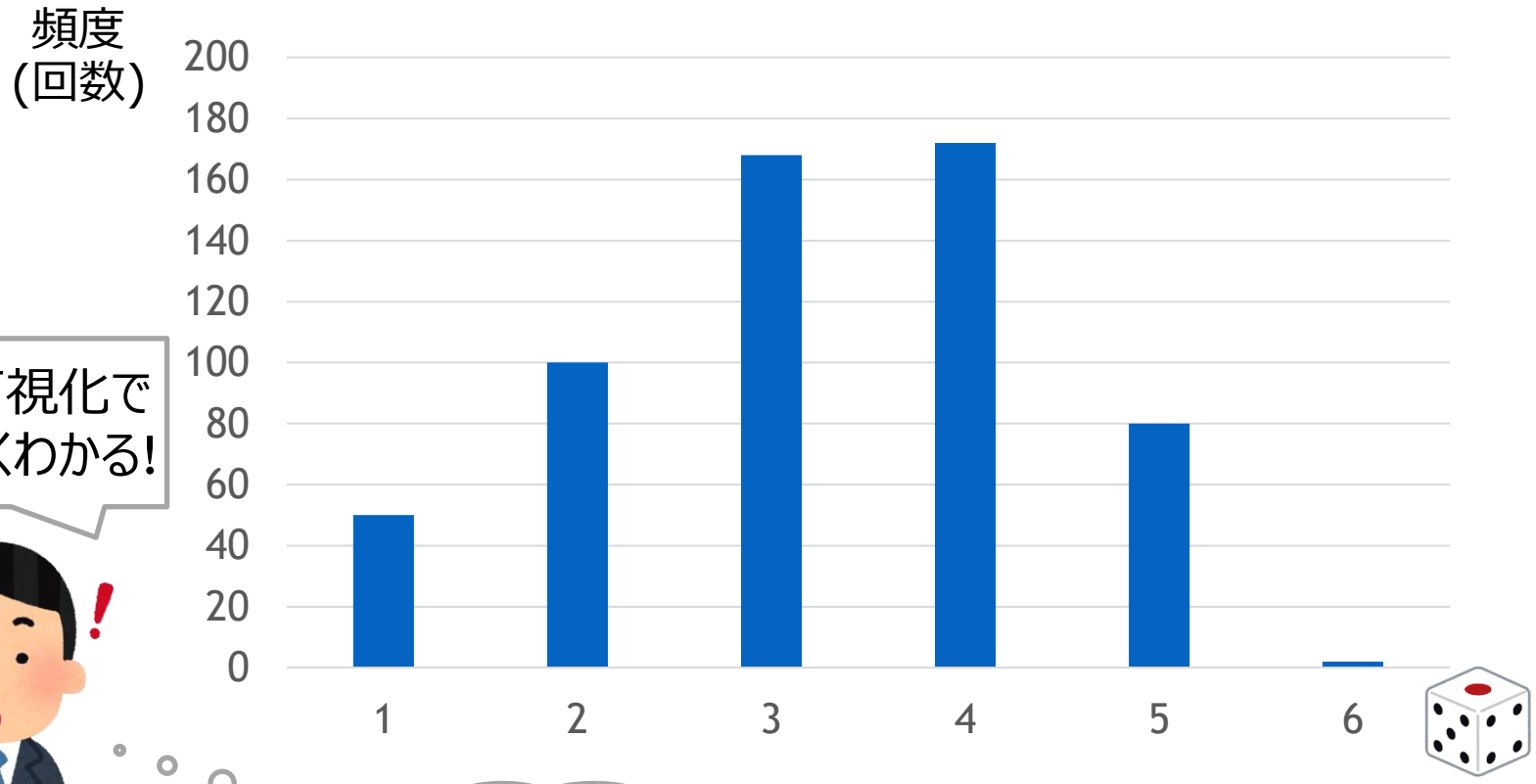


- 5段階アンケートの回答結果の集計
  - 「非常によい」が103名，「よい」が30名，…，「非常に悪い」が0名



- 今日のメニュー注文者数
  - 「かつ丼」が58食，「ラーメン」が102食，…，「高菜めし」が21食

ヒストグラムによる頻度の可視化:
さいころを1000回振って出た目のヒストグラム



頻度→ヒストグラムはよくわかる！ ところが…

- 3年10組の30人の体重の「頻度」を見ようとすると…

49.5	63.8	56.4	64.7	44.9	40.1
46.6	50.8	52.1	56.3	41.8	55.6
56.9	40.6	57.4	54.8	53.2	59.4
47.8	43.4	37.5	44.4	49.7	44.2
51.2	52.6	32.5	37.0	46.9	50.4

- ピッタリ同じ体重の人はほぼいないので，頻度は高々1

「値が連続的に変化しうるデータ」の場合，頻度は計りにくい

頻度をそのままでは計りにくい場合は，区間を考えればOK

- 3年10組の30人の体重で，5kg幅の区間を考えてみると

49.5	63.8	56.4	64.7	44.9	40.1
46.6	50.8	52.1	56.3	41.8	55.6
56.9	40.6	57.4	54.8	53.2	59.4
47.8	43.4	37.5	44.4	49.7	44.2
51.2	52.6	32.5	37.0	46.9	50.4

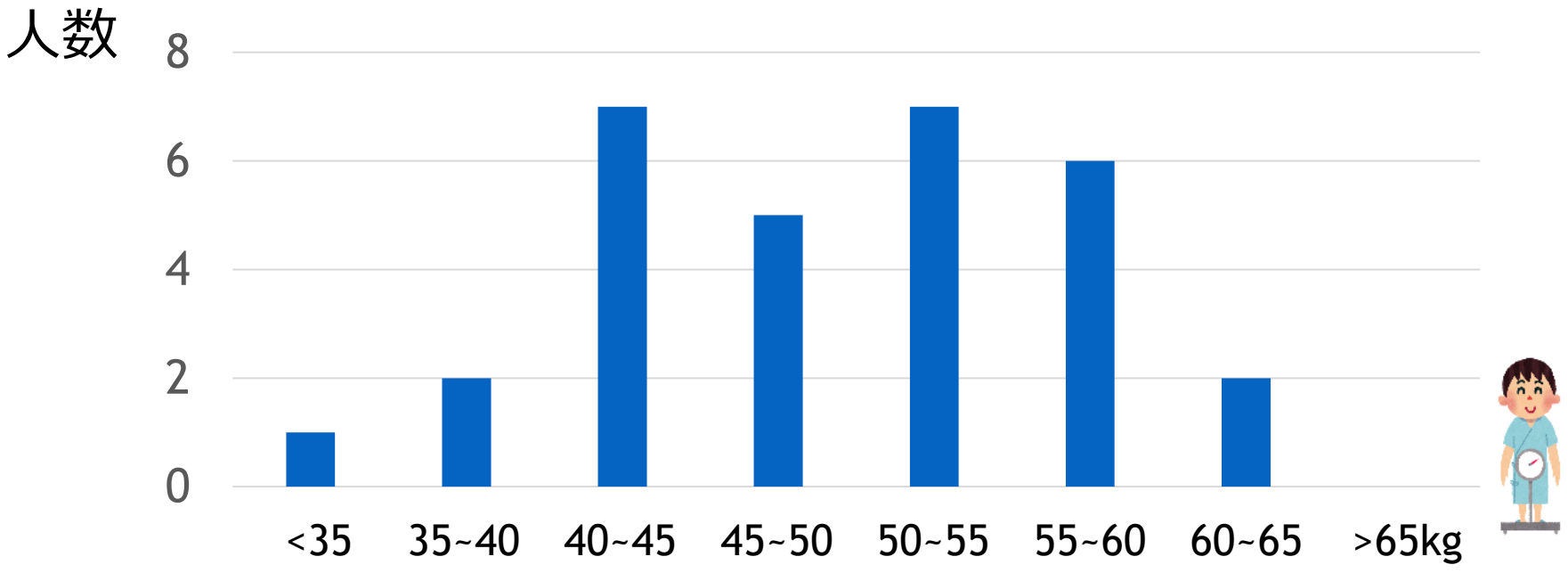
どれぐらいの体重の人が
何人いるのか
よくわかる!

35kg未満…1人，35～40kg…2人，40～45kg…7人，
45～50kg…5人，50～55kg…7人，55～60kg…6人，
60～65kg…2人，65kg以上…0人

頻度がよくわかるように！

区間を考えれば，連続的なデータでも ヒストグラムを作ることができる!

35kg未満…1人，35～40kg…2人，40～45kg…7人，
45～50kg…5人，50～55kg…7人，55～60kg…6人，
60～65kg…2人，65kg以上…0人



それぞれの区間のことを「ビン」と呼ぶ

ヒストグラムの作り方をまとめると

値は
6通りだけ

幾つかの特定の値
だけを持つデータ

各値の頻度をカウント

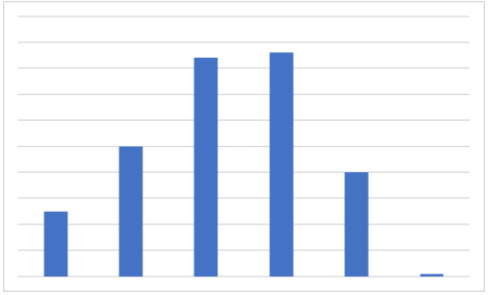
ぴったり同じ体重は
まずありえない

連続的な値を
持つデータ

区間にわける

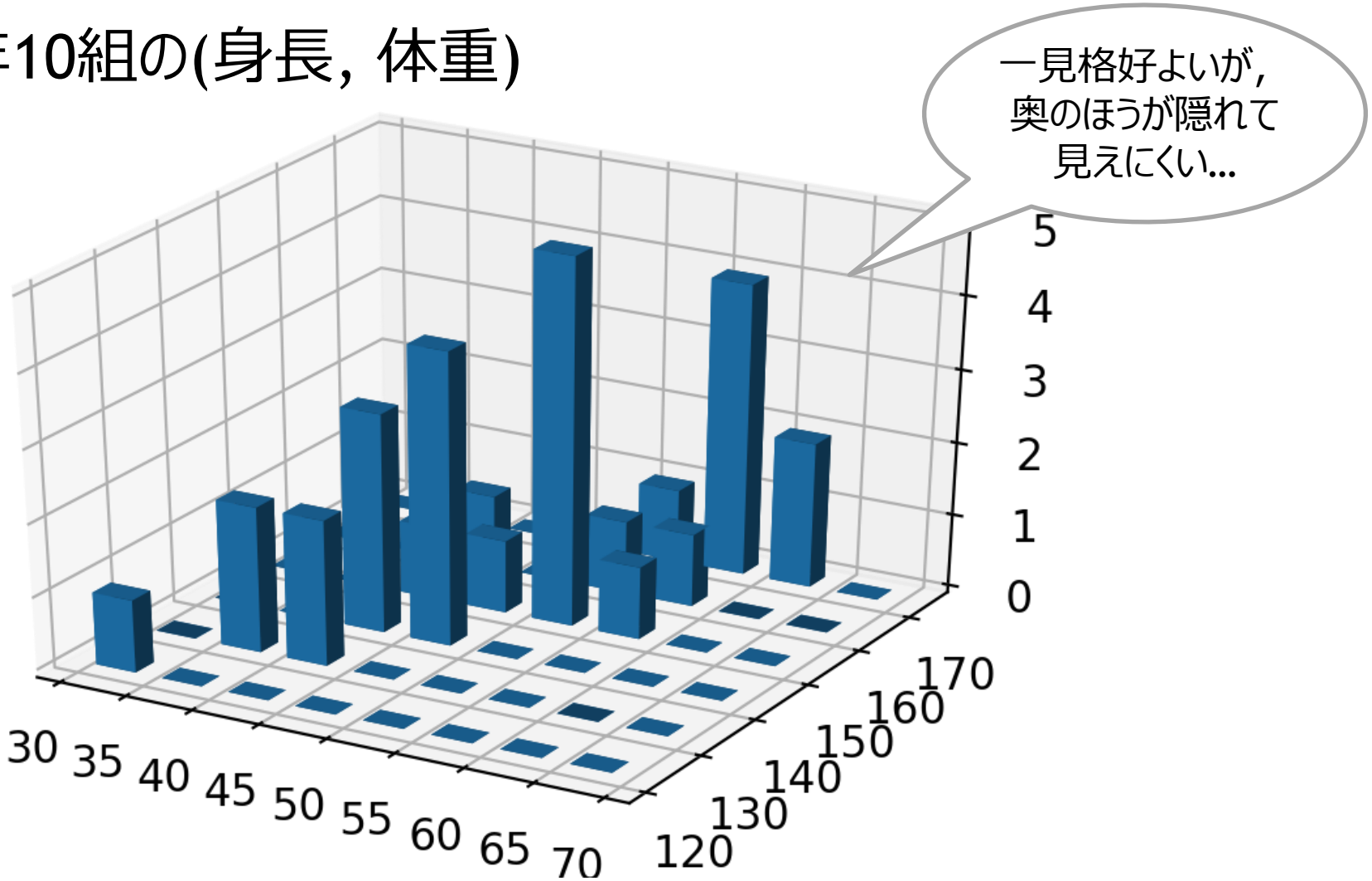
各区間に入るデータの
カウント

ヒストグラム



「ペア」のデータも，ヒストグラムで表現可能: 3D棒グラフによる表現

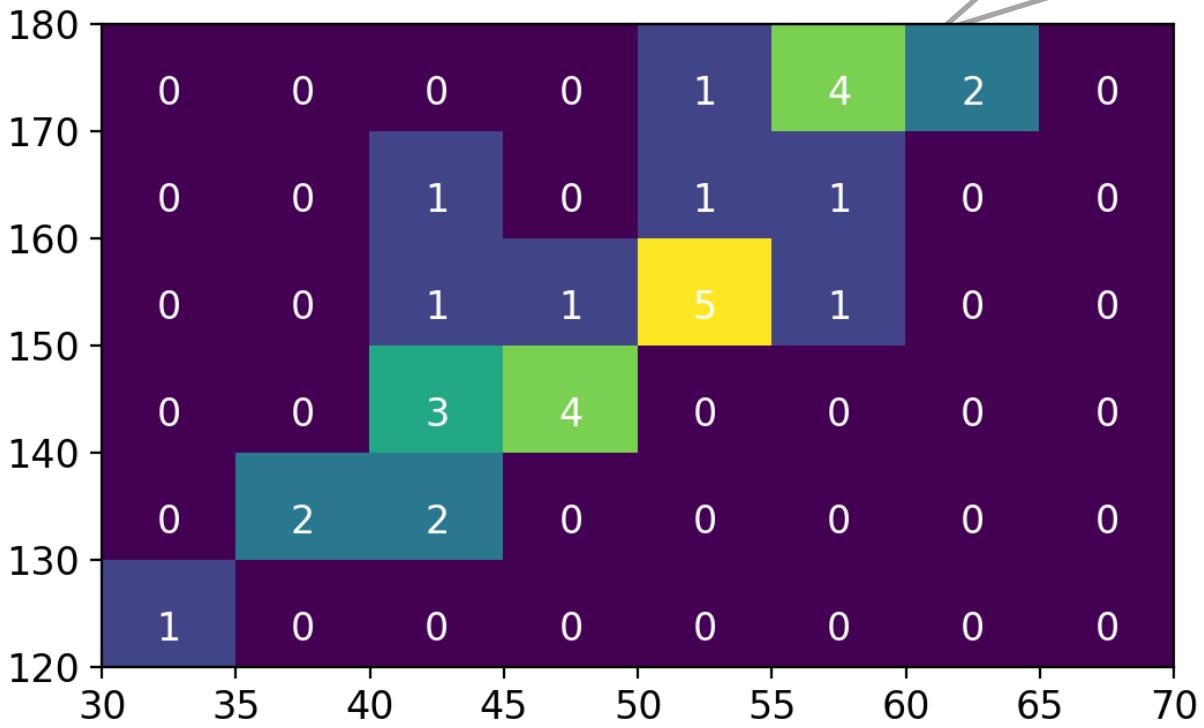
- 3年10組の(身長, 体重)



「ペア」のデータも，ヒストグラムで表現可能: ヒートマップ表現

- 3年10組の(身長, 体重)

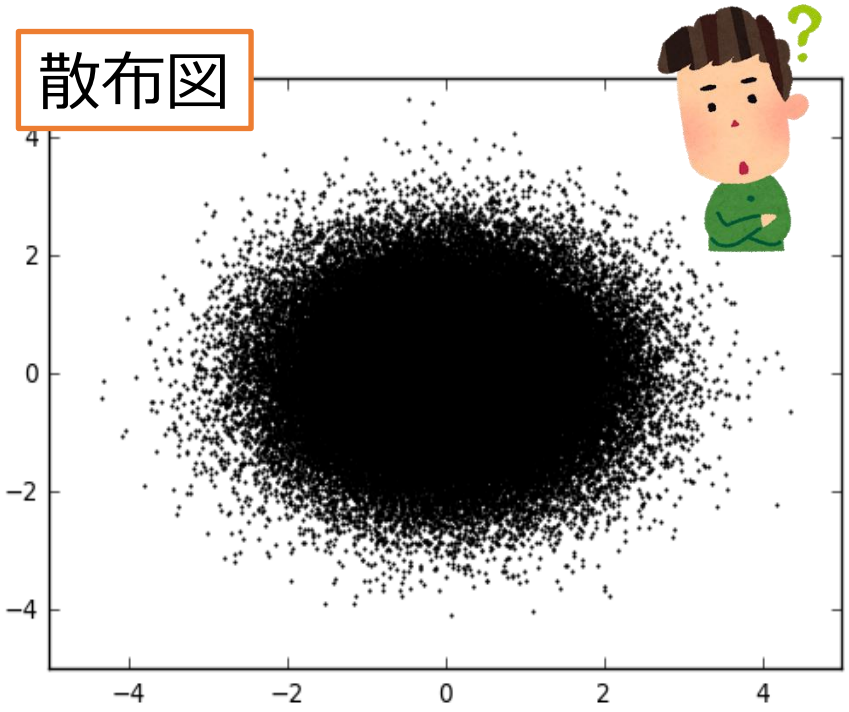
色で頻度を表す



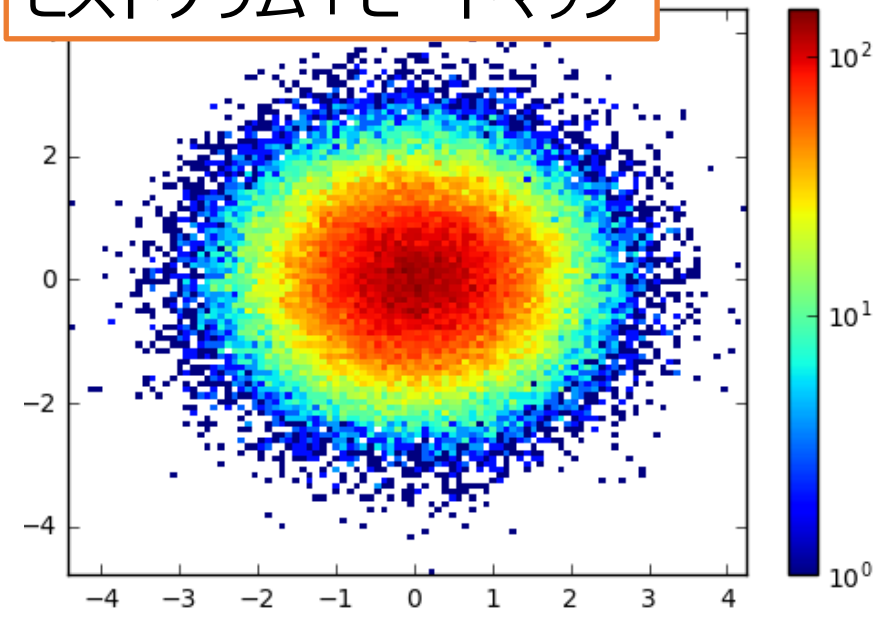
散布図のヒートマップ表現

- 散布図も，データが多すぎると，点が重なって見えなくなる
- ヒストグラムにして「ヒートマップ」にするとわかりやすい

散布図



ヒストグラム+ヒートマップ



代表値

平均値，中央値，最頻値

数多くのデータを「たった一つの数値」で代表させる

- Q: 3年10組の体重データを一言でいうとどんな感じ？

49.5	63.8	56.4	64.7	44.9	40.1
46.6	50.8	52.1	56.3	41.8	55.6
56.9	40.6	57.4	54.8	53.2	59.4
47.8	43.4	37.5	44.4	49.7	44.2
51.2	52.6	32.5	37.0	46.9	50.4

- A1: 最も重たいのは64.7kg
- A2: 最も軽いのは 32.5kg

そんな極端なケースを
言われてもなぁ..

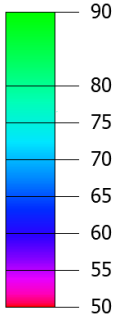
- A3: 平均すると約49.4kg

ああ，それぐらいの
体重の人が多いのね

要注意：
2-1-3も見ましょう

代表値その１：平均
色々なところで使われる！

- 平均寿命
  - もちろん各国には
    もっと短命・長命の
    人もいるが，傾向は
    わかる



https://ja.wikipedia.org/wiki/平均寿命#/media/ファイル:2018年の国・地域別平均寿命（CIAファクトブックより）.png

- 他にも色々
  - 平均年収
  - 英語テストの平均点
  - 平均身長

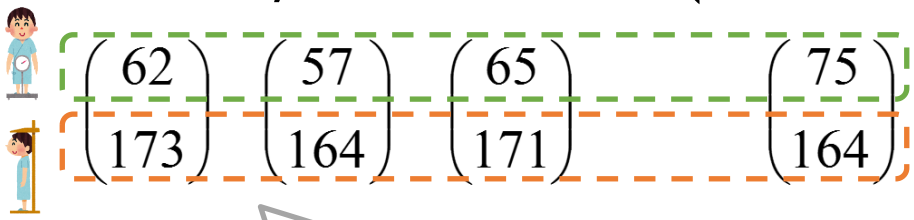
平均の求め方

- N 個のデータがあれば，基本は「全データを合計して」「N で割る」
  - 正式には「算術平均」とか「相加平均」という名前がついている

- 例：N = 5人の体重{62, 50, 49, 53, 73}の場合
  - 平均＝(62+50+49+53+73)/5 = 57.4kg

ペアになったデータの平均

- 要素ごとに足して，足した個数(=データ数N)で割るだけ



- 例： N = 5人の「(体重，身長)のペア」データの場合
  - 平均 $= \left[\begin{pmatrix}62\\173\end{pmatrix}+\begin{pmatrix}50\\162\end{pmatrix}+\begin{pmatrix}49\\158\end{pmatrix}+\begin{pmatrix}53\\156\end{pmatrix}+\begin{pmatrix}73\\176\end{pmatrix}\right]/5$

$$= \begin{pmatrix}(62+50+49+53+73)/5\\(173+162+158+156+176)/5\end{pmatrix} = \begin{pmatrix}57.4\\165\end{pmatrix}$$

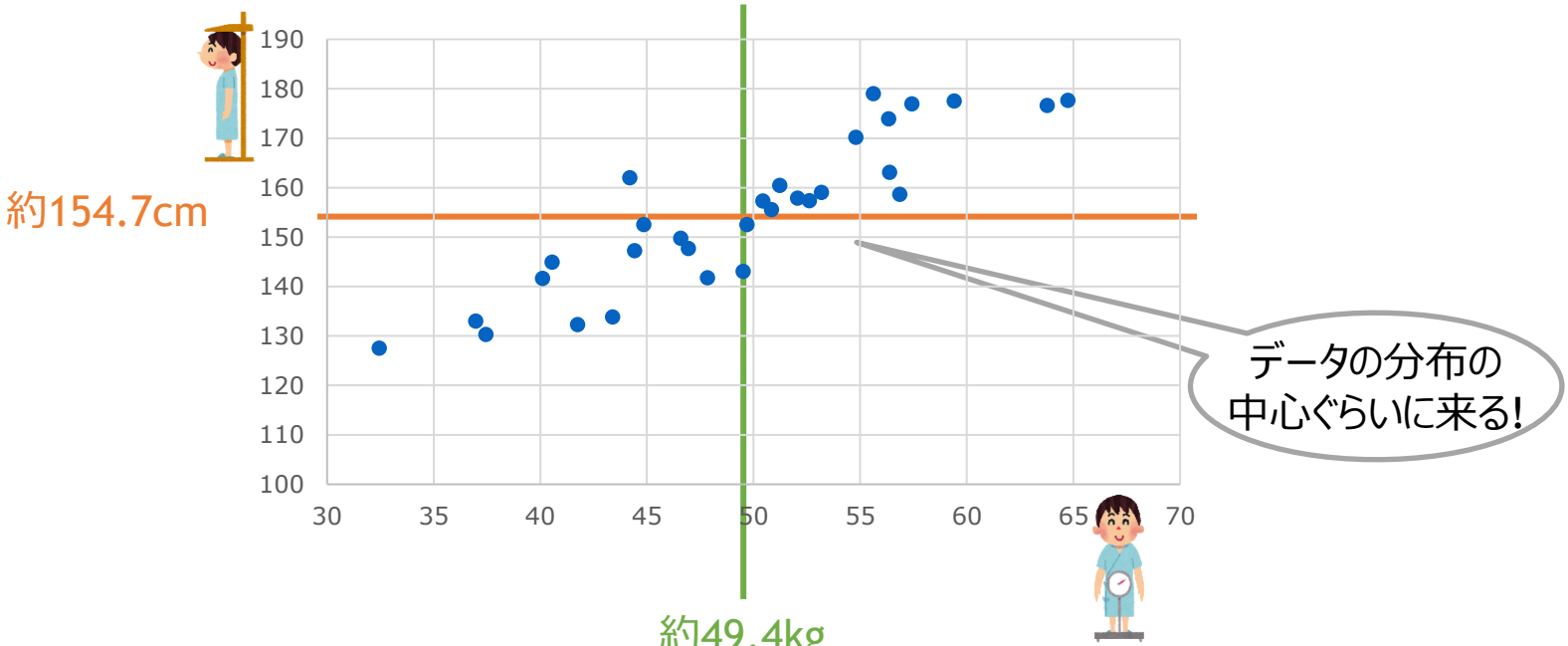
平均はどこに？

- 3年10組30人の(体重，身長)データの平均

(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 → 約49.4 kg

身長 → 約154.7cm



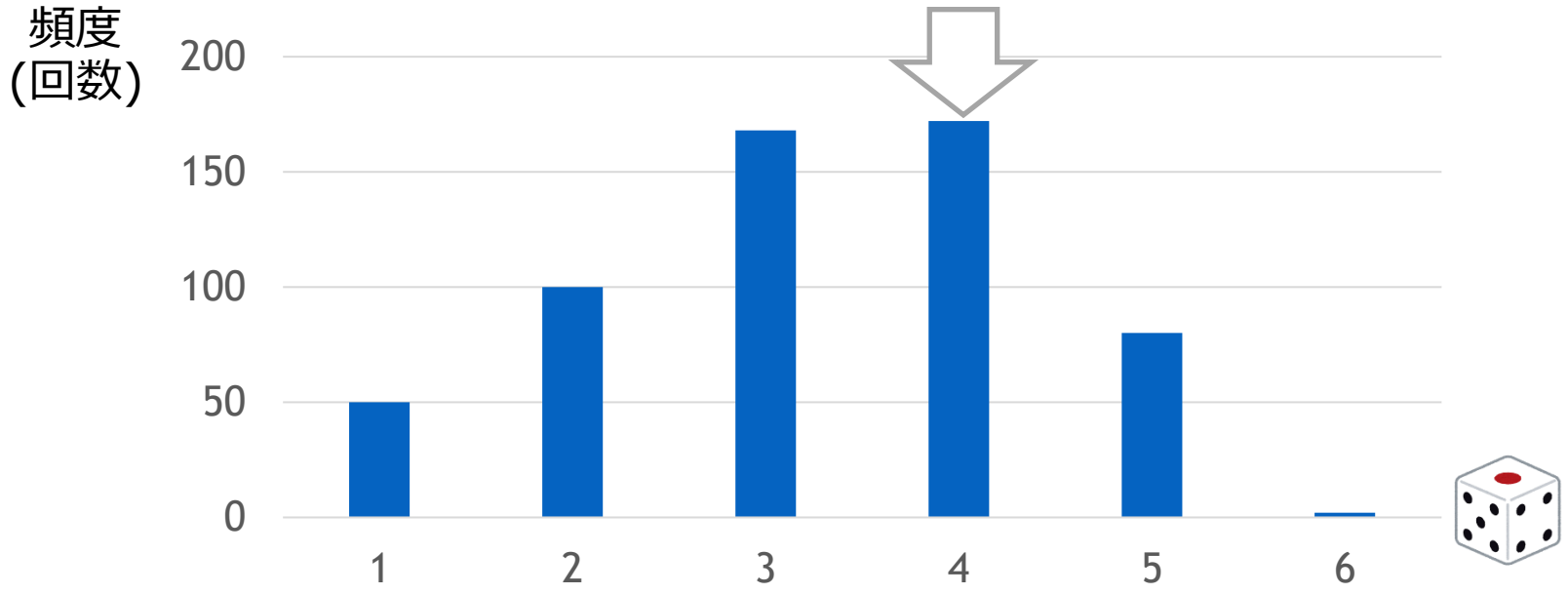
代表値その２：中央値（メディアン）

- 数値の大きさの順に並べた時に，真ん中に来る値
  - 例：N = 5人の体重{62, 50, 49, 53, 73}の場合
    - 並べ替えると，49, 50, 53, 62, 73
    - なので，中央値は53

- 利点
  - はずれ値に影響されにくい(→2-1-3)
  - （中央値はデータから選ばれるので）中央値と同じ値を持つデータが必ず存在

代表値その３：最頻値（モード）

- 最も頻度が高い=最も出やすいデータ=ヒストグラムのピーク
  - ズルいサイコロの最頻値は“4”



- どんなデータ（カテゴリデータ）にも使える！

まとめ

- データの分布
  - データを分析する前に，その傾向をまず把握したい
- 散布図やヒストグラム
  - （代表値ではなく）データ全体の広がり方をつかむ！

逆の性質!

- 代表値
  - （全体の広がり方ではなく）広がりの中心を数値で表す
    - 平均，中央値，最頻値
  - 扱いには注意が必要 → 2-1-3

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