中学校・高等学校の統計教育のポイント

青木 敏

大阪府教職員統計教育セミナー
令和元年 8 月 1 日

この講演の目的

中学数学の「資料の活用」，高校数学（数学 I）の「データの分析」で学ぶ内容のいくつかの項目について，考え方のポイントや，教える側として知っておくとよいこと，気に留めていただきたいことを解説する．

1. ヒストグラム（度数分布表）
2. 代表値（平均値，中央値）
3. データの散らばり（分散，標準偏差，四分位偏差）
4. データの相関（散布図，相関係数）

中学校学習指導要領

第 3 節 各学年の内容

[第 1 学年]

D 資料の活用

(1) 目的に応じて資料を収集し，コンピュータを用いたりするなどして表やグラフに整理し，代表値や資料の散らばりに着目してその資料の傾向を読み取ることができるようにする.

ア　ヒストグラムや代表値の必要性と意味を理解すること.

イ　ヒストグラムや代表値を用いて資料の傾向をとらえ説明すること.

[用語・記号]

平均値 中央値 最頻値 相対度数 範囲 階級

高等学校学習指導要領

第 1 部 数学
第 2 章 各科目
第 1 節 数学 I
3 内容と内容の取扱い
(4) データの分析

(4) データの分析

統計の基本的な考えを理解するとともに，それを用いてデータを整理・分析し傾向を把握できるようにする.

ア　データの散らばり

四分位偏差，分散及び標準偏差などの意味について理解し，それらを用いてデータの傾向を把握し，説明すること.

イ　データの相関

散布図や相関係数の意味を理解し，それらを用いて二つのデータの相関を把握し説明すること.

1. ヒストグラム（度数分布表）

- 1 変量の量的データに対する統計処理: まず**集計して眺める.**
- 目的: データの傾向を読み取る.

例: 60 人の身長 (cm)

149	153	153	155	153	165	155	160	165	155
157	161	161	155	155	155	155	151	159	157
157	154	165	157	154	157	151	164	155	155
158	163	160	154	160	163	156	163	155	147
162	152	158	156	166	158	160	155	161	155
159	158	159	158	148	153	167	172	168	153

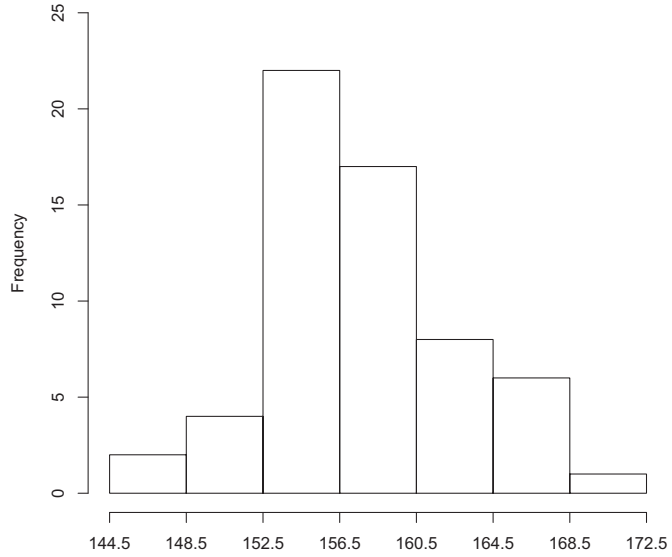
$\Longrightarrow$ 範囲: 147cm $\sim$ 172cm

4cm 刻みで集計する.

度数分布表

階級	度数	相対度数
145 ∼ 148	2	0.033
149 ∼ 152	4	0.067
153 ∼ 156	22	0.367
157 ∼ 160	17	0.283
161 ∼ 164	8	0.133
165 ∼ 168	6	0.100
169 ∼ 172	1	0.017
合計	60	1

ヒストグラム



- データの傾向: 単峰型．ほぼ対称（僅かに右の裾が重い）
- 注意: 階級「145 ∼ 148」は,「144.5 以上 148.5 未満」と書くべき（データの数値は四捨五入によって得られたもの）．階級に隙間があってはならない（棒グラフとの違い）．

- 度数分布表，ヒストグラムの作成手順

(1) 階級の数，および，階級幅を決める．

(2) 度数分布表における最小値を決める．

(3) データを数え上げて度数分布表を作成する．

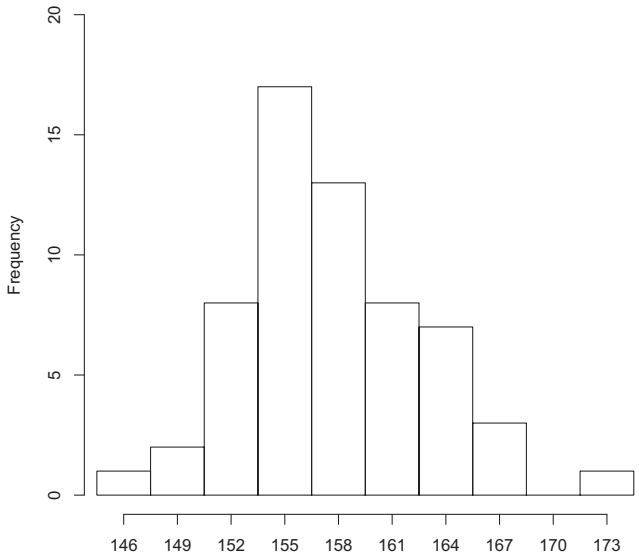
(4) ヒストグラムを作成する．

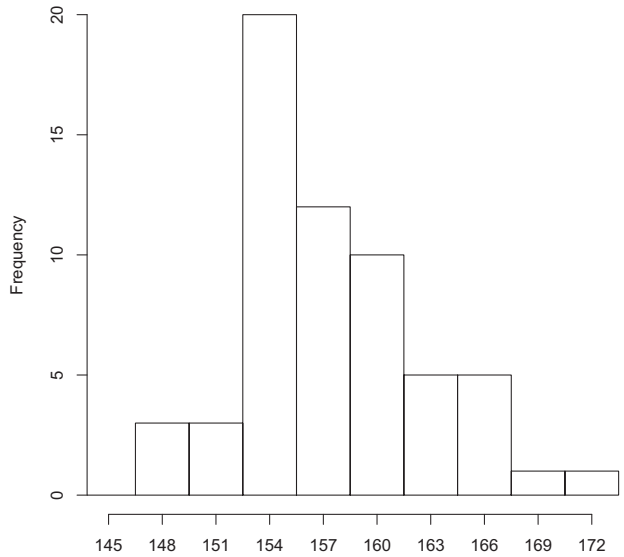
最も重要なのは，(1) の階級数・階級幅の決定．

出来上がるヒストグラムから，データの傾向が把握できるか，納得いくまで (1) と (2) を変化させて確認し，適切なものを採用する．

- 例: 60 人の身長のデータ
  - 3cm 刻み

    $145 \sim 147, 148 \sim 150, \cdots$　　$147 \sim 149, 150 \sim 152, \cdots$



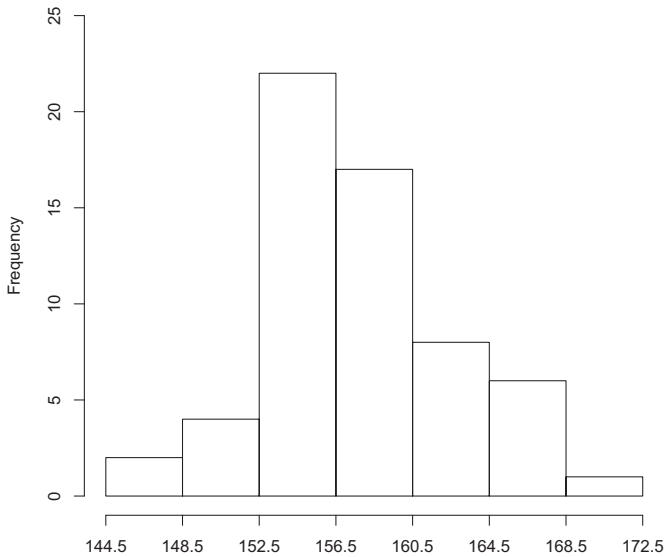


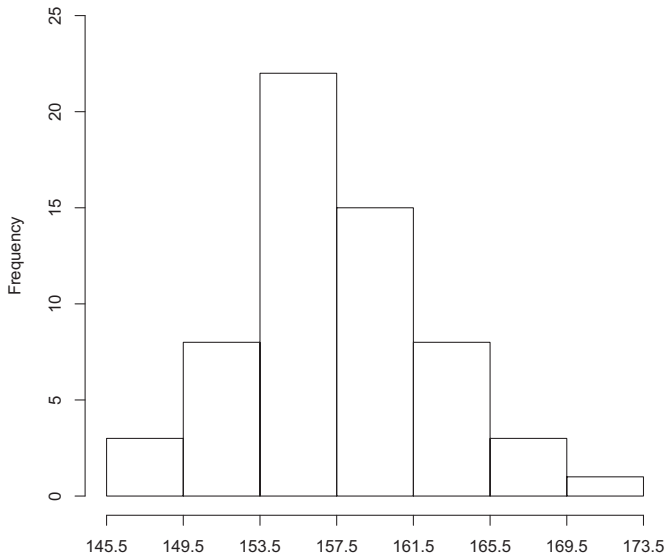
左図は，172cm の 1 名が外れ値に見える?

同じ階級幅で，階級をずらすことでデータの特徴の印象が変わるなら，その特徴には意味がない（むしろ，ミスリーディング）

○ 先ほどの 4cm 刻みで階級をずらす

$145 \sim 148, 149 \sim 152, \cdots$　　$146 \sim 149, 150 \sim 153, \cdots$

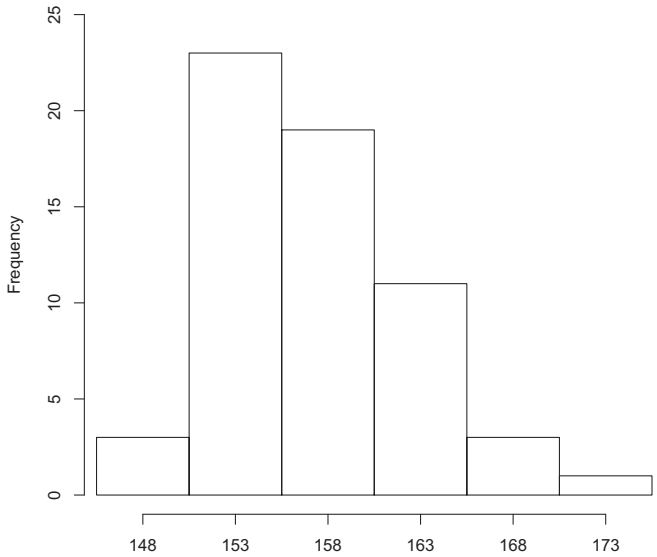


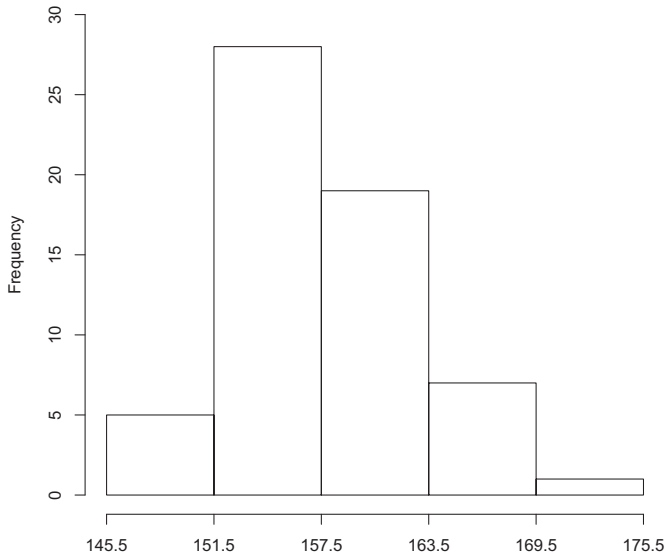


階級をずらした影響が少ない.

集団の特徴（単峰型，ほぼ対称でわずかに右の裾が重い）を捉えた，適切な階級設定だといえそうだ.

○ 5cm 刻み，6cm 刻み．だんだん大味になる．

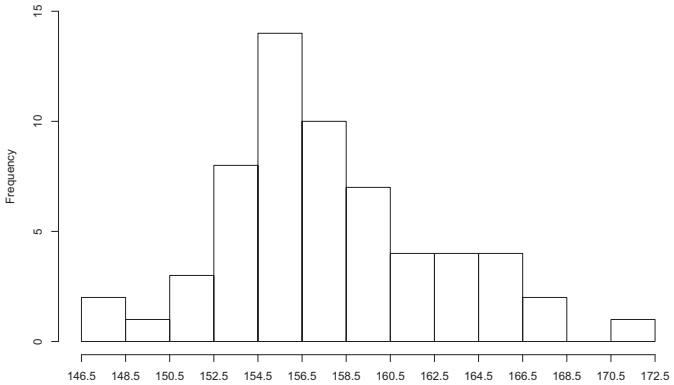


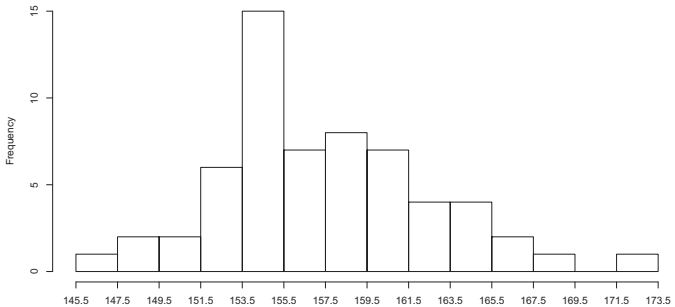


失われる情報が多すぎて，せっかく取ったデータがもったいない．

○ 2cm 刻み.

$147 \sim 148, 149 \sim 150, \cdots$　　$146 \sim 147, 148 \sim 149, \cdots$



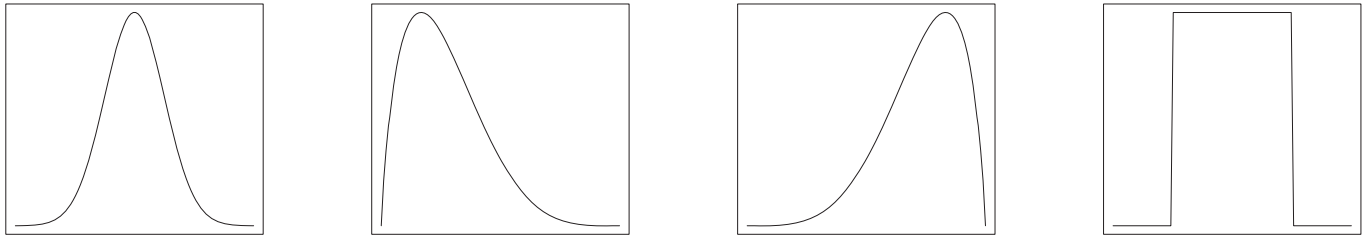


階級が多すぎて見にくい．階級をずらすことによる影響も大きい．
情報を適切に整理できていない．

この例では，階級は 4cm 刻みがよさそうだ．

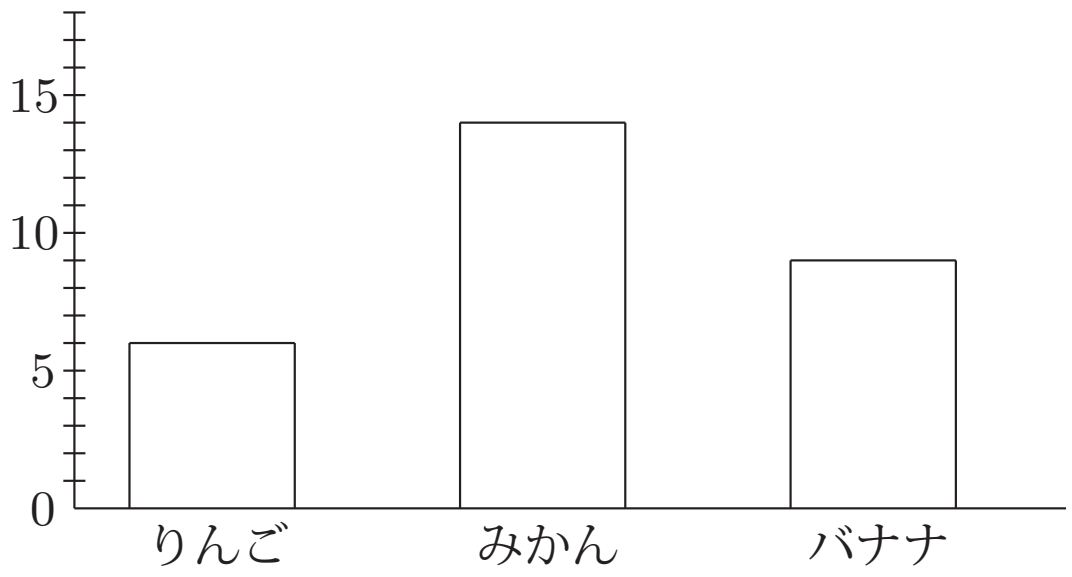
データの傾向は，ほぼ対称な単峰型（僅かに右の裾が重い）で，外れ値は無さそう．

- データの傾向: 代表的な 4 つのパターン
  - ベル型: 左右対称になっているベルのような形
  - 右に裾が重い: 全体に対して，値の大きい方にもデータがある
  - 左に裾が重い: 全体に対して，値の小さい方にもデータがある
  - 一様: ある範囲内でどの値も同程度に出現する



ヒストグラムに 2 個以上の峰がある場合，**「複数の異なる特徴を持つ集団が混ざっていないか」,「データの大きさが不十分ではないか」**などと疑う.

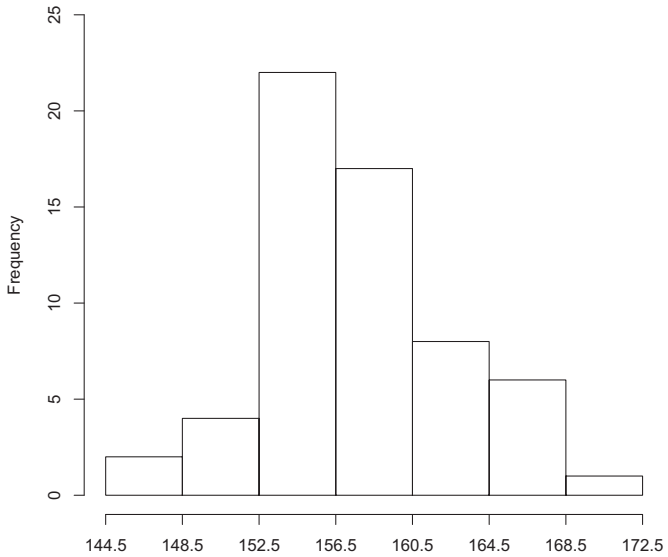
- ヒストグラムと棒グラフの違い
  - 棒グラフ: 質的データに利用．度数を棒の**高さで**表現する．



項目の順番は入れ替えてもよい．

見やすさのため，棒と棒の間には隙間を開けるのが普通

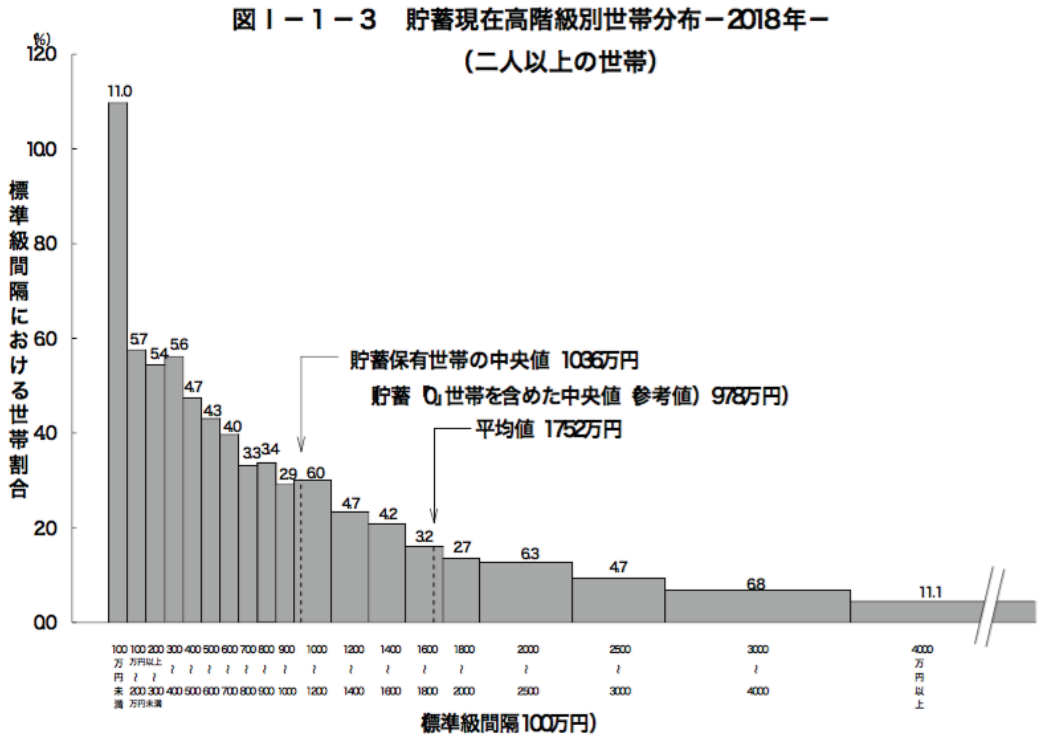
- ヒストグラム: 量的データに利用．度数を「階級幅 × 高さ」の柱状の**面積で**表現する．



横軸の順序は変更できない．階級に隙間があってはダメ．横軸は，**小さい値から大きな値まで連続しており**，その動きに応じて度数がどのように変化するかを見る．(変化の様子 = 「分布する (distribute)」)

- 階級幅が一定でない例:

（図Ⅰ－１－３）

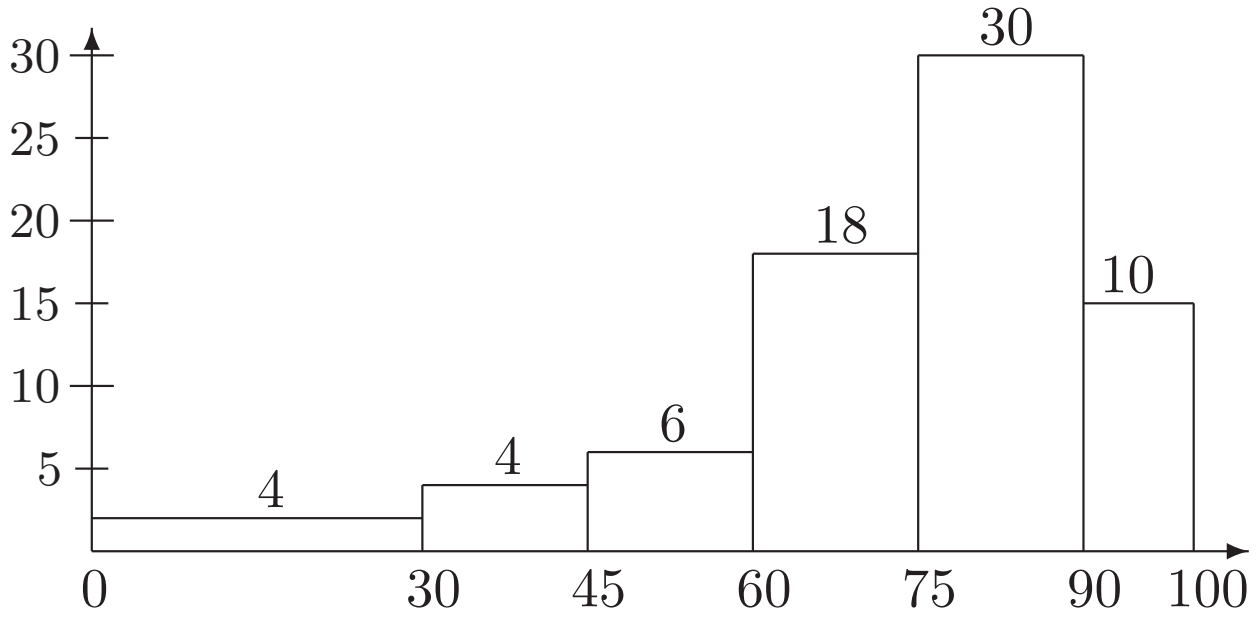


1000 万円未満の階級幅は 100 万円（標準階級幅）. 1000 万円以上は階級幅を大きくしている（見やすさのため）.

出典: 総務省統計局「貯蓄現在高階級別世帯分布（二人以上の世帯）2018 年」

- 練習問題
  以下の度数分布表からヒストグラムを作成せよ.

試験成績		人数
0 点以上	30 点未満	4
30 点以上	45 点未満	4
45 点以上	60 点未満	6
60 点以上	75 点未満	18
75 点以上	90 点未満	30
90 点以上	100 点以下	10
合計		72



15 点を標準の階級幅とすれば,

- 「0 点以上 30 点未満」の階級の度数 4 の高さは $4 \times \dfrac{15}{30} = 2$
- 「90 点以上 100 点以下」の階級の度数 10 の高さは $10 \times \dfrac{15}{10} = 15$

縦軸の目盛は「標準階級幅における点数」

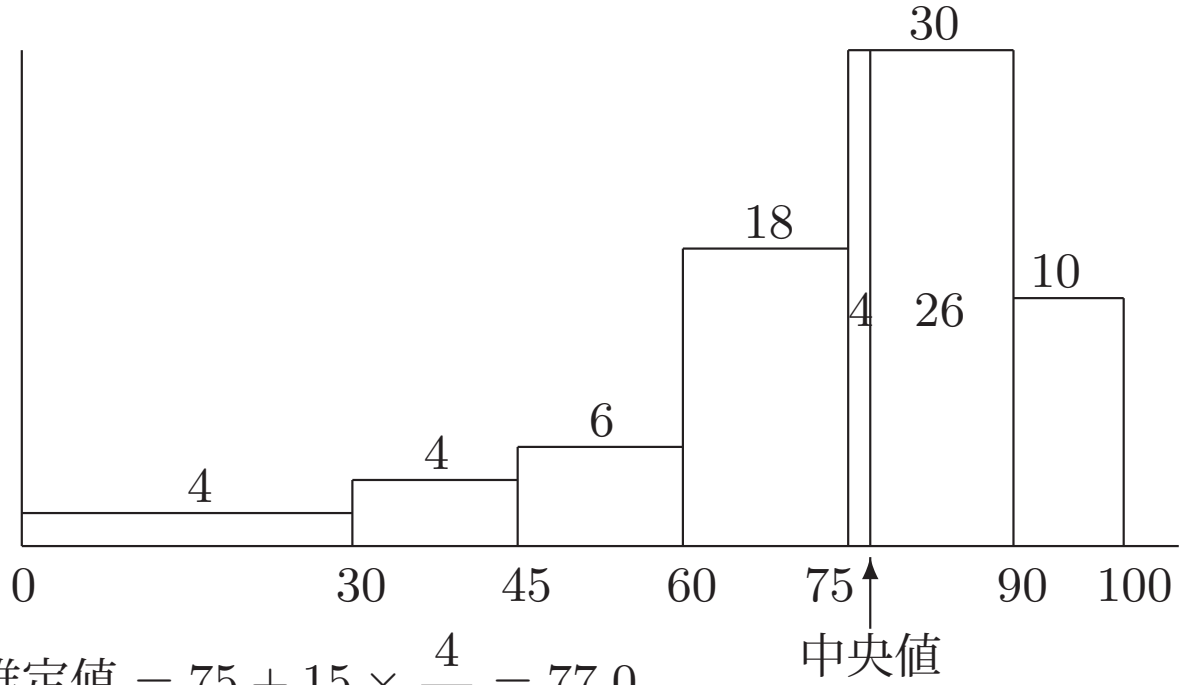
発展: 度数分布表（ヒストグラム）から中央値を推定する

- 高校数学では，中央値が含まれる階級（の階級値）で中央値を推定する．

試験成績		人数	
0 点以上	30 点未満	4	
30 点以上	45 点未満	4	
45 点以上	60 点未満	6	
60 点以上	75 点未満	18	
75 点以上	90 点未満	30	← 中央値を含む階級
90 点以上	100 点以下	10	
合計		72	

$$\text{中央値の推定値} = \frac{75+90}{2} = 82.5$$

- 「面積を 2 分する点」として，より精度のよい推定を行うことができる．



$$\text{中央値の推定値} = 75 + 15 \times \frac{4}{30} = 77.0$$

発展: スタージェスの公式

- データのサイズ n から適切な階級数を定めるための古典的な指針

$$\text{階級の数} \approx 1 + \log_2 n$$

- 身長のデータ ($n = 60$) であれば

$$6 = 1 + \log_2 32 < 1 + \log_2 60 < 1 + \log_2 64 = 7$$

  より，6 または 7 くらい.（4cm 刻みの階級数が 7 であった.）

- 多くの統計ソフトウェアのデフォルト設定などに使われている.
- **あくまでも目安．基本は試行錯誤.**

2. 代表値（平均値，中央値）

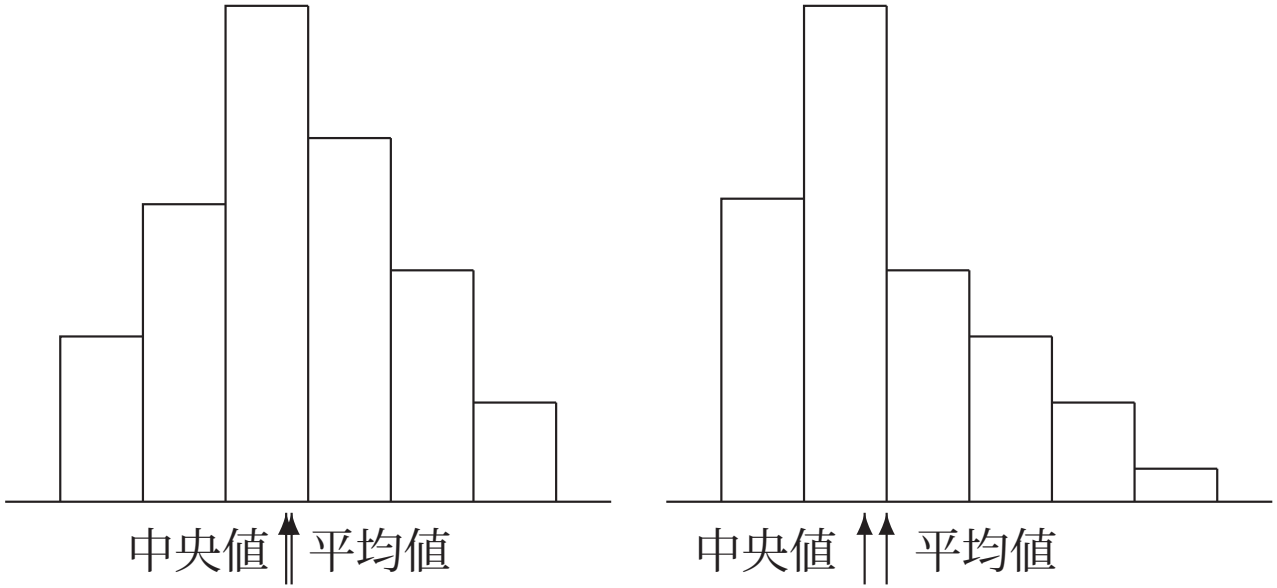
- データの代表値として用いられる，基本統計量.
- サイズ n のデータ $x_1, x_2, \ldots, x_n$ の平均値 $\bar{x}$ は

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 中央値の定義: 「データを小さい順に並べたときの中央の値」
  - $x_1 = 70,\ x_2 = 85,\ x_3 = 93 \qquad \Rightarrow$ 中央値 $= 85$
  - $x_1 = 70,\ x_2 = 85,\ x_3 = 93,\ x_4 = 99 \Rightarrow$ 中央値 $= \dfrac{85+93}{2} = 89.0$

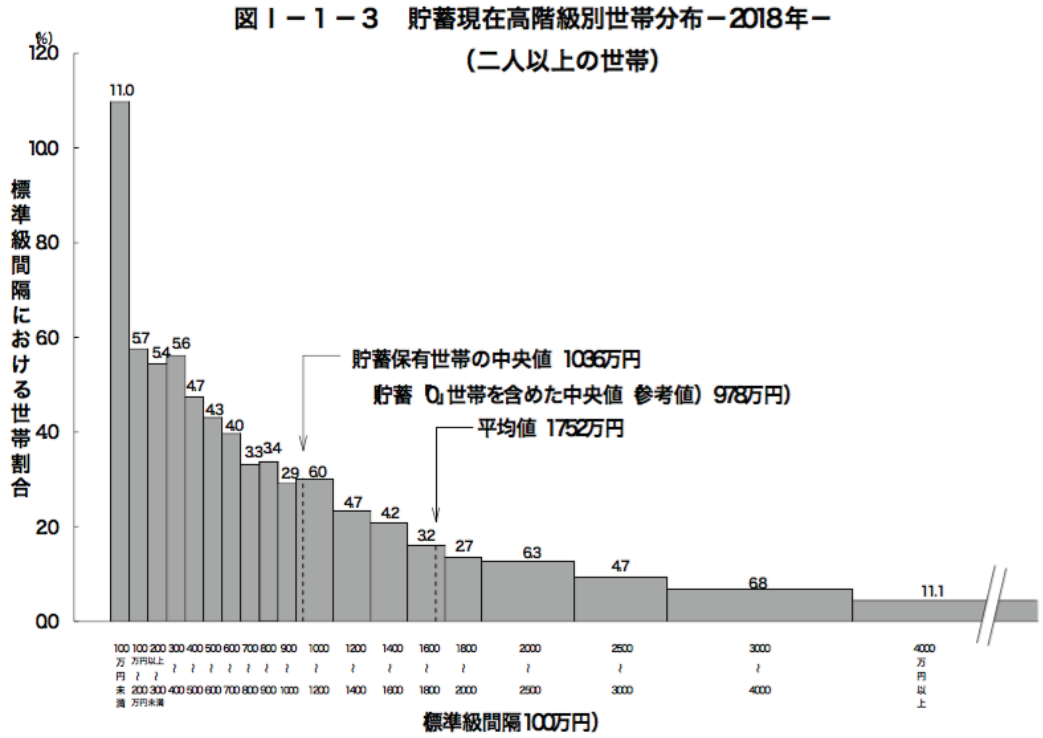
  データのサイズ n が奇数のときは中央の値.
  偶数のときは「中央を挟む 2 つの値の平均」.

- 平均値は，データの代表値として最もよく使われる．
  - 「数学の学力試験で 80 点を取ったよ!」
    （平均点: 55 点）$\longrightarrow$ 「すごい! よく頑張ったね!」
    （平均点: 70 点）$\longrightarrow$ 「... 頑張ったね.　（塾に通わせないと!）」
- 平均値が「データの中心」の値になるのは，ヒストグラムが左右対称な場合のみ．



- 裾が重い分布では，平均値よりも中央値の方が「データの代表値」の意味に近くなる．平均値は，あまり意味をなさない．

（図I－1－3）



貯蓄の平均値: 1752 万円．中央値: 1036 万円．

「貯蓄現在高が平均値 (1752 万円) を下回る世帯が約 3 分の 2 を占める.」

- 平均値は「外れ値」の影響が大きい．中央値は「外れ値」に強い．
  - 例: 不動産物件

$$x_1 = 80000,\ x_2 = 75000,\ x_3 = 92000,\ x_4 = 78000,\ \underline{x_5 = 400000}$$

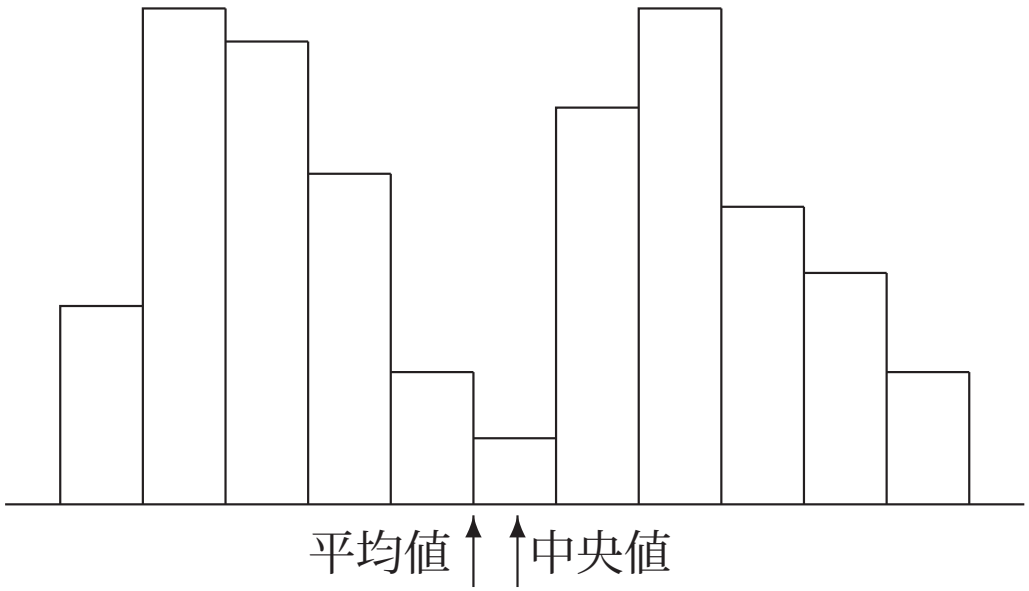
$$\Longrightarrow \text{平均値} = 145000, \quad \text{中央値} = 80000.$$

  - 例: 小麦粉における銅の含有量 ($\mu g\ g^{-1}$)(Abbey, 1998, Analytical Methods Committee より引用．)

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90	3.03	3.03
3.10	3.37	3.40	3.40	3.40	3.50	3.60	3.70	3.70	3.70
3.70	3.77	<u>5.28</u>	<u>28.95</u>						

$$\Longrightarrow \text{平均値} = 4.28, \quad \text{中央値} = 3.385.$$

- データに 2 個以上の峰があるときは，平均値，中央値，いずれも「分布の代表値」としては不適切.



データに「複数の異なる特徴を持つ集団が混ざっていないか」を検討し，混ざっていれば集団を分け（**層別**という），それぞれの均質な集団ごとに代表値を求める.

発展: 頑健性

- 「外れ値の影響を受けにくい」という性質を，頑健性 (robustness) といいます．中央値は，平均値に比べて，頑健な統計量です．

- 一方で，中央値は，データの中央の値しか見ていないので，平均値に比べて情報を損失しています.「頑健性」と「情報の量」にはトレードオフの関係があります.

  （次に扱う,「分散，標準偏差」と「四分位偏差」の関係も同様.）

	大 ⟵ 情報の量 ⟶ 小 弱 ⟵ 頑健性 ⟶ 強	
代表値	平均	中央値
散布度	標準偏差	四分位偏差

- 平均値と中央値の中間の性質をもつ代表値には，刈り込み平均があります．

> <u>刈り込み平均 (trimmed mean)</u>
>
> データを大きさの順に並べ，小さい方の $100\alpha\%$ と大きい方の $100\alpha\%$ を取り除き，残りについての平均値を計算する．これを $100\alpha\%$ 刈り込み平均という．

例： 2.20 2.20 2.40 2.40 2.50 2.70 2.80 2.90 3.03 3.03
3.10 3.37 3.40 3.40 3.40 3.50 3.60 3.70 3.70 3.70
3.70 3.77 <u>5.28</u> <u>28.95</u>

α	0(平均値)	0.1	0.2	0.3	0.4	0.5(中央値)
刈り込み平均	4.28	3.205	3.239	3.273	3.283	3.385

- 複数の審査員による採点競技などで使われます．
例: 飛び込みの採点．7 人の審査員の得点のうち，低い点 2 つと高い点 2 つを除き，残る 3 人の得点の平均点（に難易率を掛けたもの）で採点する（$\alpha = 2/7 = 0.286$ の刈り込み平均．参照：兵庫県水泳連盟 HP）．

3. データの散らばり（分散，標準偏差，四分位偏差）

- データの散らばりの指標として用いられる.
- サイズ n のデータ $x_1, \ldots, x_n$ の分散 s^2 は

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

分散の正の平方根

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

を，標準偏差という.

- 分散: 「個々の観測値と平均との差を 2 乗したものの平均」

- 分散の性質

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

  - 分散 = 「2 乗の平均」− 「平均の 2 乗」
  - 基本的な関係式ですので，きちんと証明できることが望ましいですが，$\displaystyle\sum_{i=1}^{n}$ の扱いに不慣れな場合は難しいかもしれません．

「データの分析」を扱う数学 I は，数学 B で数列を学ぶよりも早いのが普通だと思います．数学 B を学んで Σ の扱いに慣れた後，センター試験，大学入学共通テスト（および個別試験）対策で改めて学ぶのが現実的だと思います．

- 分散と標準偏差の例

A, B, C, D の 4 つの班の試験成績

	成績 (点)					平均点	分散	標準偏差 (点)
A 班	30	40	50	60	70	50	200	14.1
B 班	10	30	50	70	90	50	800	28.3
C 班	0	40	50	60	100	50	1040	32.2
D 班	0	10	50	90	100	50	1640	40.5

  - 標準偏差は，観測値と同じ単位をもつ.
  - 各グループ内の成績の偏差値は?

- データを，平均 50，標準偏差 10 に**標準化**した値を偏差値という．

  データ $x_1, \ldots, x_n$，平均値 $\bar{x}$，標準偏差 s

  $\Rightarrow$ 個体 i の偏差値 $= 10 \times \left(\dfrac{x_i - \bar{x}}{s} \right) + 50$

	成績（上）と偏差値（下）					平均点	分散	標準偏差
A	30	40	50	60	70	50	200	14.1
	35.9	42.9	50	57.1	64.1			
B	10	30	50	70	90	50	800	28.3
	35.9	42.9	50	57.1	64.1			
C	0	40	50	60	100	50	1040	32.2
	34.5	46.9	50	53.1	65.5			
D	0	10	50	90	100	50	1640	40.5
	37.7	40.1	50	59.9	62.3			

- 偏差値の考え方: **標準化**

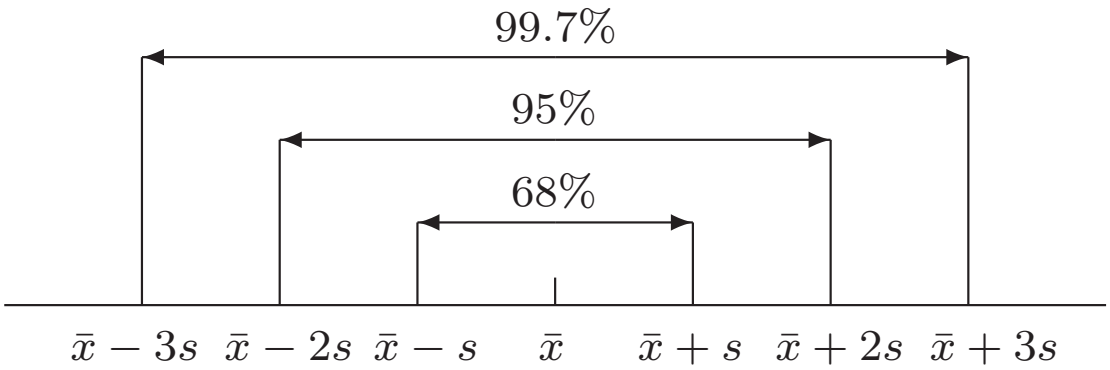
データ $x_1, \ldots, x_n$, 平均値 $\bar{x}$，標準偏差 s
$\Rightarrow$ 個体 i の標準化得点: $z_i = \dfrac{x_i - \bar{x}}{s}$

  $z_1, \ldots, z_n$ は，平均値が 0，分散（標準偏差）が 1 となる．

- 標準化により，平均と分散が異なる集団の中の相対的な位置を比較することができる．
  - 例: 同じ 60 点でも，A 班の太郎くん（偏差値 57.1）の方が C 班の次郎くん（偏差値 53.1）よりも集団内の相対的な出来が良い (??)

- サイズの小さい集団では，分散や標準偏差，偏差値の値は，しばしば直感に反することもある．**サイズがある程度大きく，分布の形がベル型の場合には**，平均と分散・標準偏差から，ある範囲に含まれる観測値のおおよその割合が分かる．$\Rightarrow$ <u>68 - 95 - 99.7 **ルール**</u>

発展: 68 - 95 - 99.7 ルール



- 正規分布の分布型にもとづく目安.
- 68 - 95 - 99.7 ルールが正確に成り立つのは，集団の分布が正規分布にしたがうときであるが，標本サイズがある程度大きい均質な集団では，近似的に正規分布と考えてよいことが多い.
- このルールによれば,「偏差値 40 ～ 60 の生徒は全体の 68%」「偏差値 30 ～ 70 の生徒は全体の 95%」「偏差値 70 を超えるのは全体の 2.5%（受験者 20 万人なら上位 5000 人）」「偏差値 80 を超えるのは全体の 0.15%（受験者 20 万人なら上位 300 人）」などとなる.

- 四分位偏差の定義:
  データを小さい順に並べ，小さい方から 1/4 づつの場所にある値を第 1 四分位数（Q_1），第 2 四分位数（Q_2．これは中央値に等しい），第 3 四分位数（Q_3）としたとき，

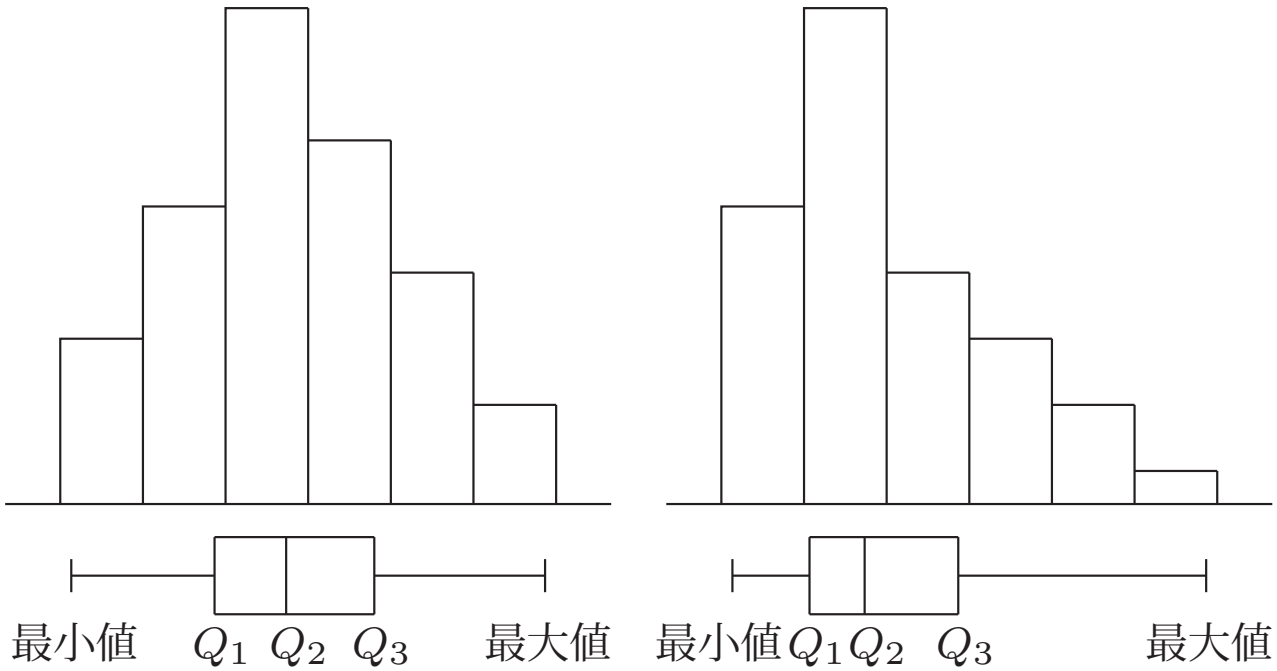
$$\text{四分位偏差} = \frac{Q_3 - Q_1}{2}.$$

  $Q_3 - Q_1$ を四分位範囲という．

- 発展: Q_1, Q_2, Q_3 の厳密な定義
  標本サイズ n を 4 で割った余り (4 通り) で定義が変わる．

データ	Q_1	Q_2	Q_3
$1, 2, 3, 4, 5, 6, 7, 8$	2.5	4.5	6.5
$1, 2, 3, 4, 5, 6, 7, 8, 9$	2.5	5	7.5
$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$	3	5.5	8
$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$	3	6	9

- 四分位偏差の性質: 分散，標準偏差よりも頑健な散らばりの指標
- 四分位偏差，四分位範囲を視覚的に表したもの: **箱ヒゲ図**



箱ヒゲ図から，データの範囲，分布の概形（対称か，右または左の裾が重いか）が分かる．

- 分散・標準偏差は，外れ値に弱い．四分位偏差は外れ値に強い（頑健）．

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90	3.03	3.03
3.10	3.37	3.40	3.40	3.40	3.50	3.60	3.70	3.70	3.70
3.70	3.77	<u>5.28</u>	<u>28.95</u>						

$\Rightarrow$ 平均値 $= 4.28$，中央値 $= 3.385$，分散 $= 26.89$，四分位偏差 $= 0.475$

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90	3.03	3.03
3.10	3.37	3.40	3.40	3.40	3.50	3.60	3.70	3.70	3.70
3.70	3.77								

$\Rightarrow$ 平均値 $= 3.11$，中央値 $= 3.235$，分散 $= 0.268$，四分位偏差 $= 0.45$

	大 $\longleftarrow$ 情報の量 $\longrightarrow$ 小 弱 $\longleftarrow$ 頑健性 $\longrightarrow$ 強	
代表値	平均	中央値
散布度	標準偏差	四分位偏差

発展: 分散は，なぜ，平均を引くのか

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

- より一般に,「a の周りの分散」を定義して,

$$g(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$$

を a の関数とみたとき，$g(a)$ が最小となるときの a の値が $\bar{x}$ です.

$$g'(a) = -\frac{2}{n}\sum_{i=1}^{n}(x_i - a) = -2(\bar{x} - a) = 0 \ \Rightarrow a = \bar{x}$$

発展: 頑健な散らばりの指標

- 外れ値に強い散らばりの指標には，四分位偏差以外にもあります.
- MAD (Median Absolute Deviation) :
  「中央値からの差の絶対値」の中央値

  例: 2.20, 2.40, 2.70, 3.03, 3.60, 5.28, 28.95
  $\Rightarrow$ 中央値 $= 3.03$
  $\Rightarrow$ 中央値からの差: -0.83, -0.63, -0.33, 0, 0.57, 2.25, 25.92
  $\Rightarrow$ 中央値からの差の絶対値: 0, 0.33, 0.57, 0.63, 0.83, 2.25, 25.92
  $\Rightarrow$ 中央値からの差の絶対値の中央値 (MAD): 0.63
- 頑健統計学の分野では，最もよく用いられる散らばりの指標です.

4. データの相関（散布図，相関係数）

- 2 次元のデータ:

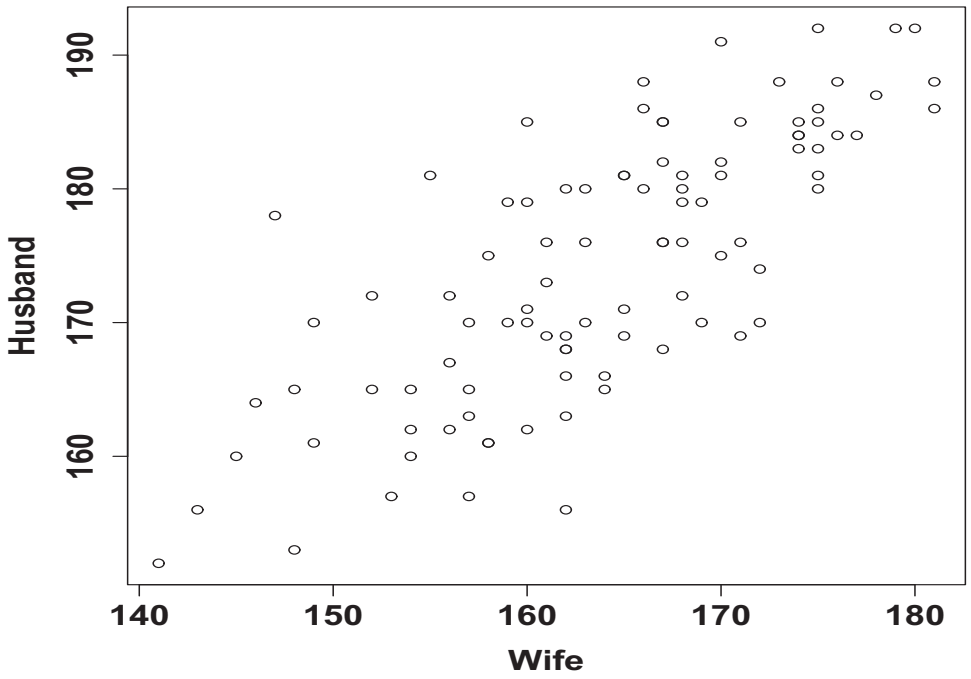
$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

例: 新婚のカップル 96 組の，夫と妻の身長のデータ[a]

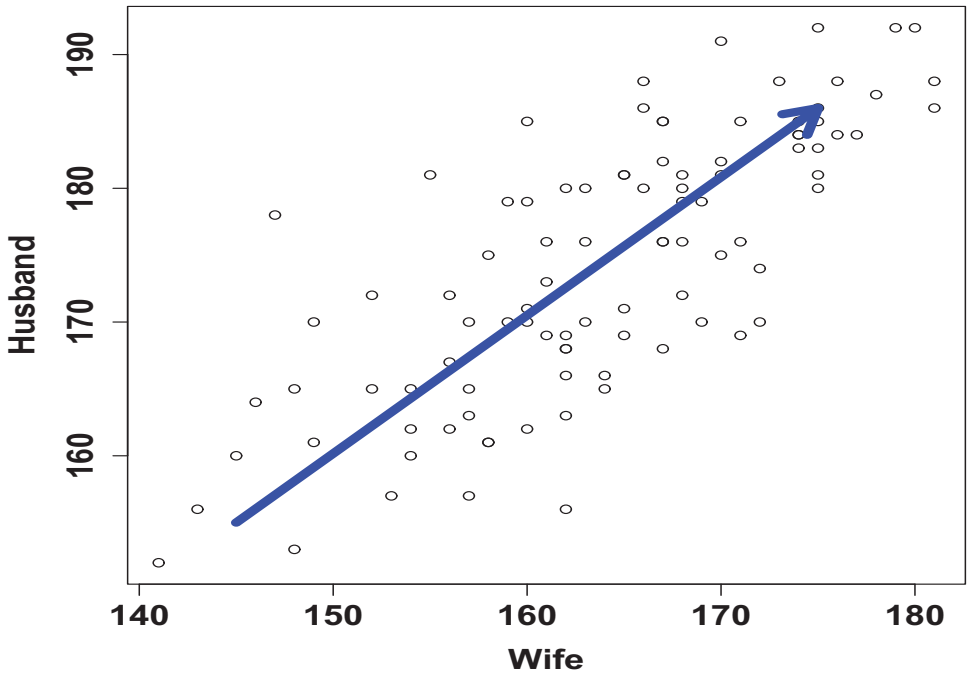
No.	Husband	Wife
1	186	175
2	180	168
3	160	154
⋮	⋮	⋮
96	181	170

[a] S. Chatterjee and A. S. Hadi. (2012). *Regression Analysis by Example*, 5th ed. Wiley の Table 2.11 から転載．この本は Web 上で pdf が公開されています．

- まず，**散布図**を描く．縦軸: 妻 (Wife)，横軸: 夫 (Husband).



- 右上がりの傾向が見られる.



身長が高い者は高い相手と，低い者は低い相手と結婚したい?

この「右上がりの傾向」を定量的に表す量が**相関係数**.

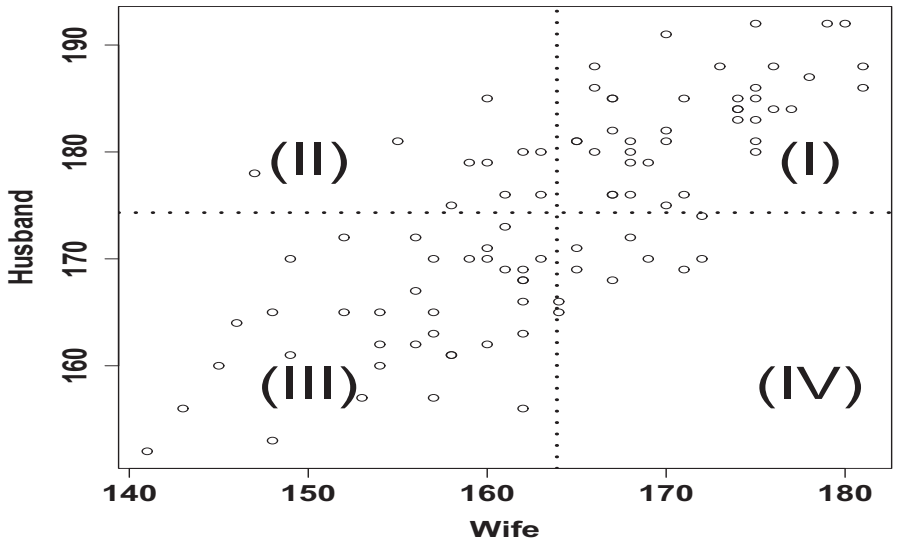
- 相関係数の定義

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad : \quad x, y \text{ の}\textbf{共分散}$$

相関係数 r_{xy} は，共分散 s_{xy} を，各変数の標準偏差 s_x, s_y で割って，$-1 \leq r_{xy} \leq 1$ となるように標準化したもの.

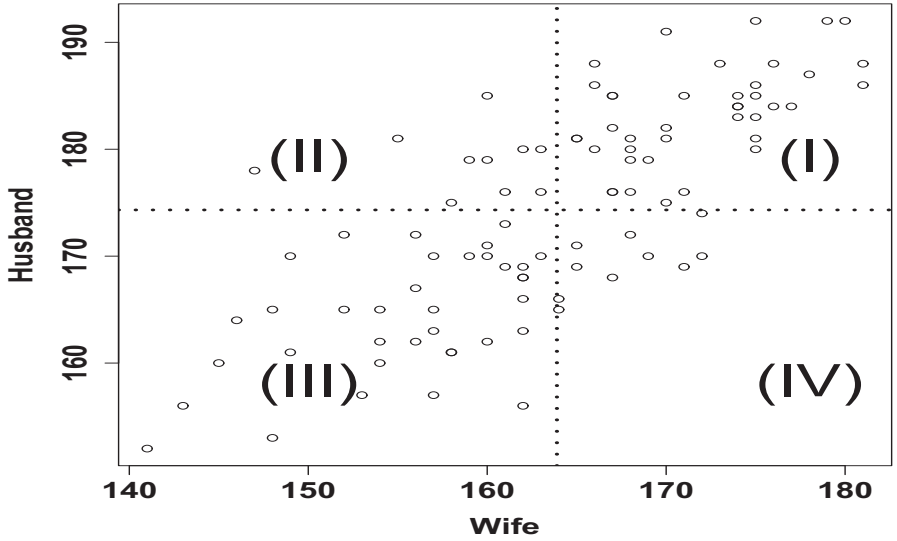
- 共分散の考え方: $x = \bar{x}, y = \bar{y}$ で <u>散布図を 4 分割</u> する.



$$\begin{cases} (x_i - \bar{x})(y_i - \bar{y}) > 0, & \text{領域 (I),(III)} \\ (x_i - \bar{x})(y_i - \bar{y}) < 0, & \text{領域 (II),(IV)} \end{cases}$$

$s_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ は，領域 (I)(III), (II)(IV) のどちらの観測値の寄与が大きいかをはかる量（符号付き平均）.

- 新婚カップルデータ



$$\bar{x} = 163.90, \quad \bar{y} = 174.32$$

$$s_{xy} = \frac{1}{96}\sum_{i=1}^{96}(x_i - 163.90)(y_i - 174.32) = 68.69$$

つまり，正の相関．標準偏差 $s_x = 9.08$，$s_y = 9.91$ より，相関係数は

$$r_{xy} = \frac{68.69}{9.08 \times 9.91} = 0.76$$

発展: $|r_{xy}| \leq 1$ の証明

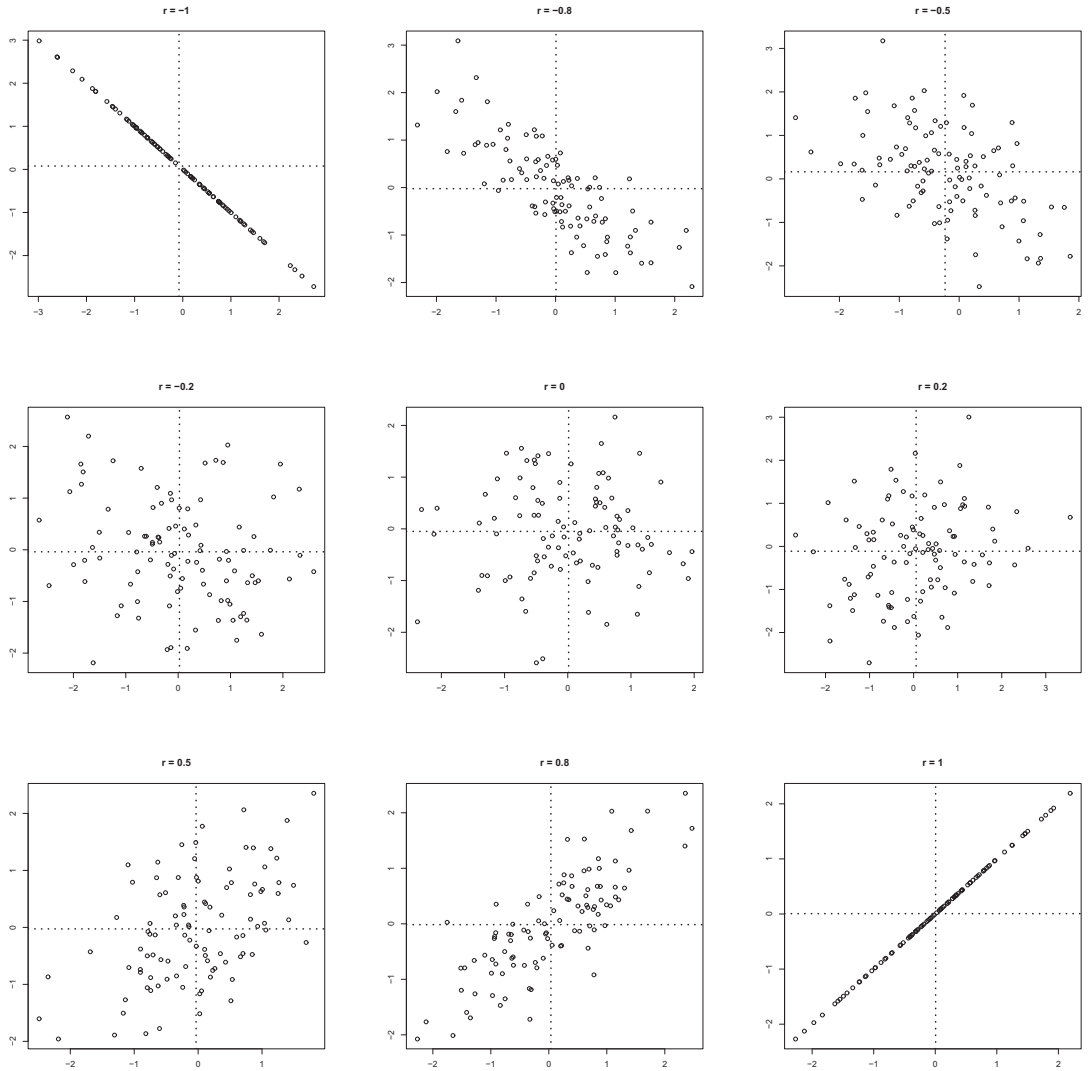
- 高校数学の教科書には載っていませんが，以下の方法であれば，**式変形して定義に当てはめるだけで**示すことができます.
- 証明の方針:

$$f_i = \left(\frac{x_i - \bar{x}}{s_x}\right) \pm \left(\frac{y_i - \bar{y}}{s_y}\right)$$

とおき，$\displaystyle\sum_{i=1}^{n} f_i^2 \geq 0$ を式変形する.

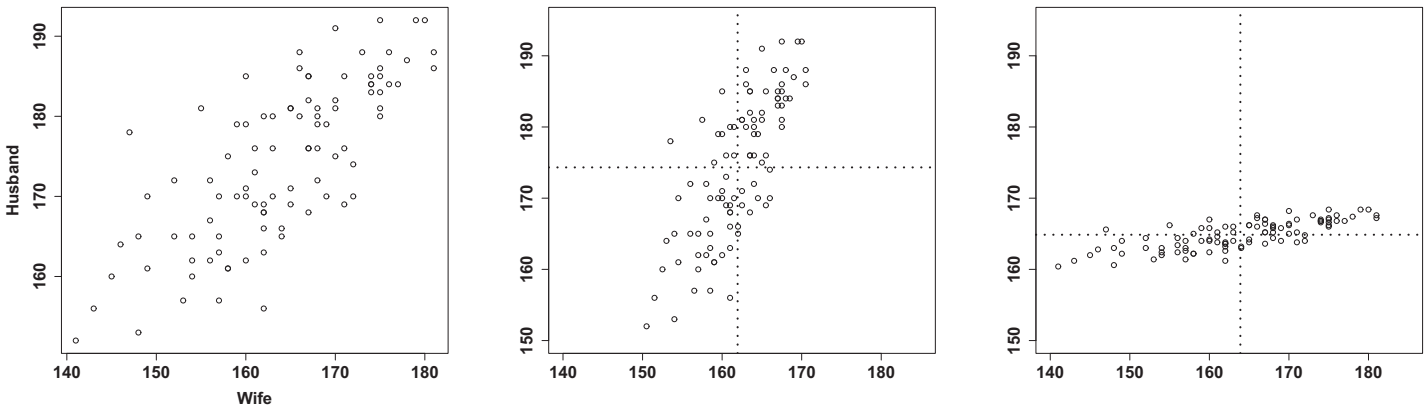
$$\left(\begin{aligned}
\sum_{i=1}^{n} f_i^2 &= \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 \pm 2\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) + \sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)^2 \\
&= \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \pm \frac{2}{s_x s_y}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{s_y^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= n \pm \frac{2}{s_x s_y} n s_{xy} + n \\
&= 2n(1 \pm r_{xy})
\end{aligned}\right)$$

- 相関の強さと散布図（順に $r_{xy} = -1,\ -0.8,\ -0.5,\ -0.2,\ 0,\ 0.2,\ 0.5,\ 0.8,\ 1$）



- 相関の強さ ⇔ 直線的な関連（**相関関係**）の強さ

  傾きの大きさには無関係



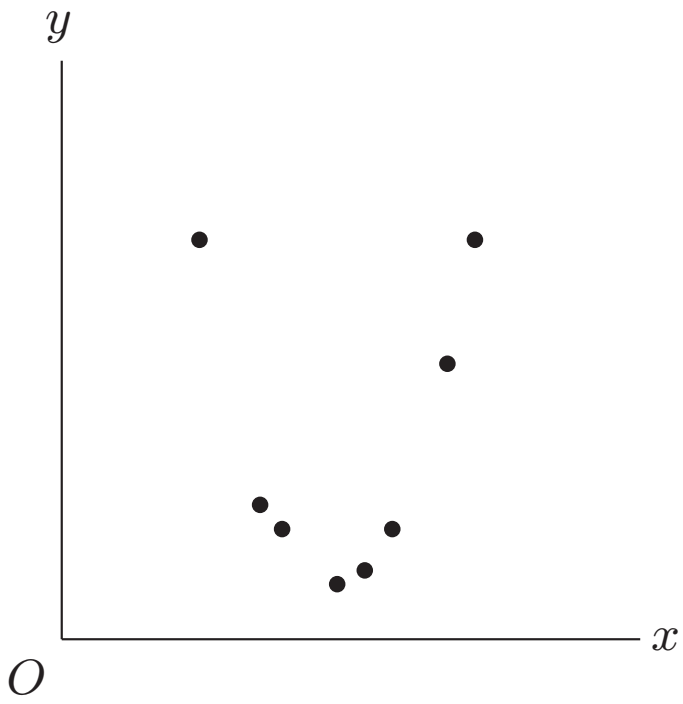
（オリジナル）　　$x_i \leftarrow \dfrac{1}{2}x_i + 80$　　$y_i \leftarrow \dfrac{1}{5}y_i + 130$

$\boxed{\text{相関係数はすべて } r_{xy} = 0.76}$

- 相関係数は，変数の 1 次変換に関して不変．つまり

$$u_i = ax_i + b, \quad v_i = cy_i + d \quad (a, c > 0) \;\Rightarrow\; r_{xy} = r_{uv}$$

- 相関がない（無相関である）とは？
  （誤）x, y に関連がない.　（正）x, y に <u>直線的な</u> 関連がない.



$r_{xy} = 0.11$
弱い正の相関（無相関に近い）だが，x, y には関係 ($y = ax^2 + bx + c$ の二次関数）がある.

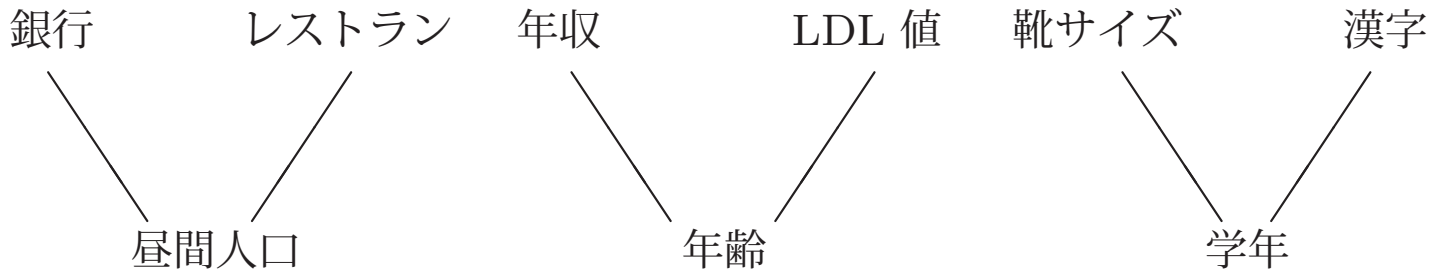
相関係数から判断できるのは，直線的な関係の大きさのみ.

- ふたつの変量の間に相関関係があっても，**必ずしも因果関係 (原因と結果の関係) があるとは限らない．** 相関関係があって因果関係がないときは，**みかけ上の相関**とよばれる．

「銀行が多い地域にはレストランが多い.」

「年収とコレステロール値は高い正の相関がある.」

「靴のサイズと，記憶している漢字の数には高い正の相関がある.」



背後にある影響力のある変数により，本来関係がない変数間に**見かけ上の相関**が現れる．見かけ上の相関は，問題の本質をとらえていないため，誤解や誤った結論を導く危険がある．

- 2 つ以上のグループの混在（層別の失敗）

  例:　（神戸大学理学部，計算数学 2 のレポートから[a]）

  マリオカート DS のカート性能の解析

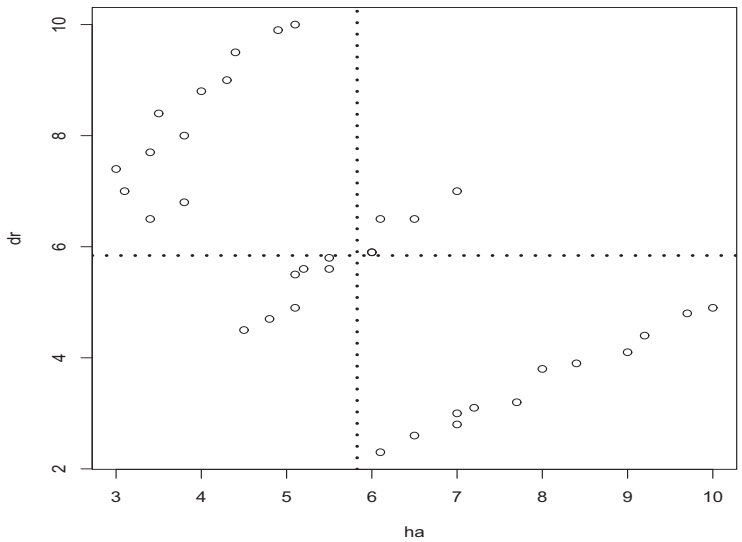
カート	最高速度	加速	重量	ハンドリング	ドリフト	アイテム
MR1	5.8	7.0	7.2	5.1	5.5	10.0
MR2	6.5	7.1	6.4	5.5	5.8	6.5
			⋮			
RB3	9.0	4.8	8.9	6.1	2.3	10.0

（36 種類のカートに関する 6 変数データ．出典は攻略 Wiki ページ.）

  - 統計手法を使って分析したところ，ハンドリング (ha)，ドリフト (dr) の二つの変数の主成分への寄与が，直感に反する結果となった.
  - この 2 変数の相関係数は -0.61（そこそこ強い負の相関）.

[a] 主成分分析の講義における，データ解析の自由課題．以前はプロ野球などのスポーツデータの解析が定番だったが，最近は題材にゲームを取り上げる者が多い.

○ 散布図:

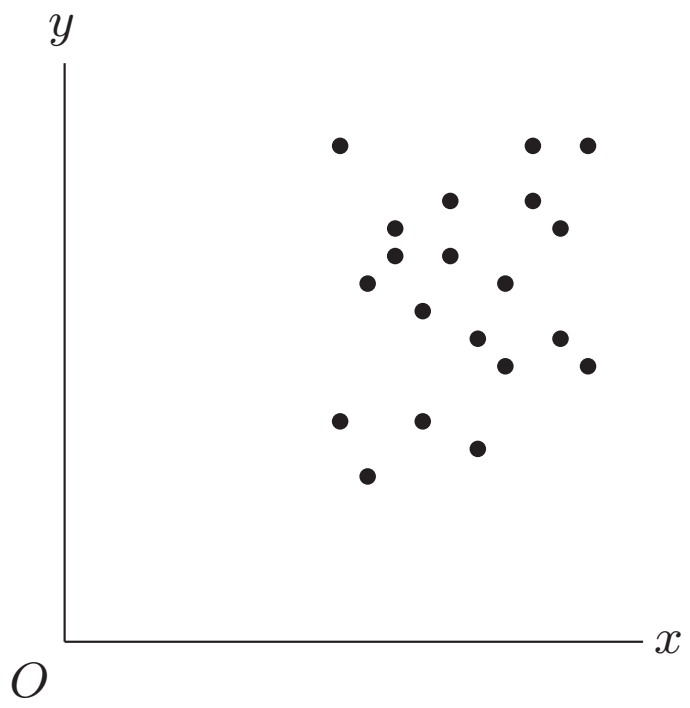


カートは 3 つの群に分けることができ，それぞれの群では，ハンドリングとドリフトに強い正の相関（直感に合う結果）がある.

- 打ち切りの効果

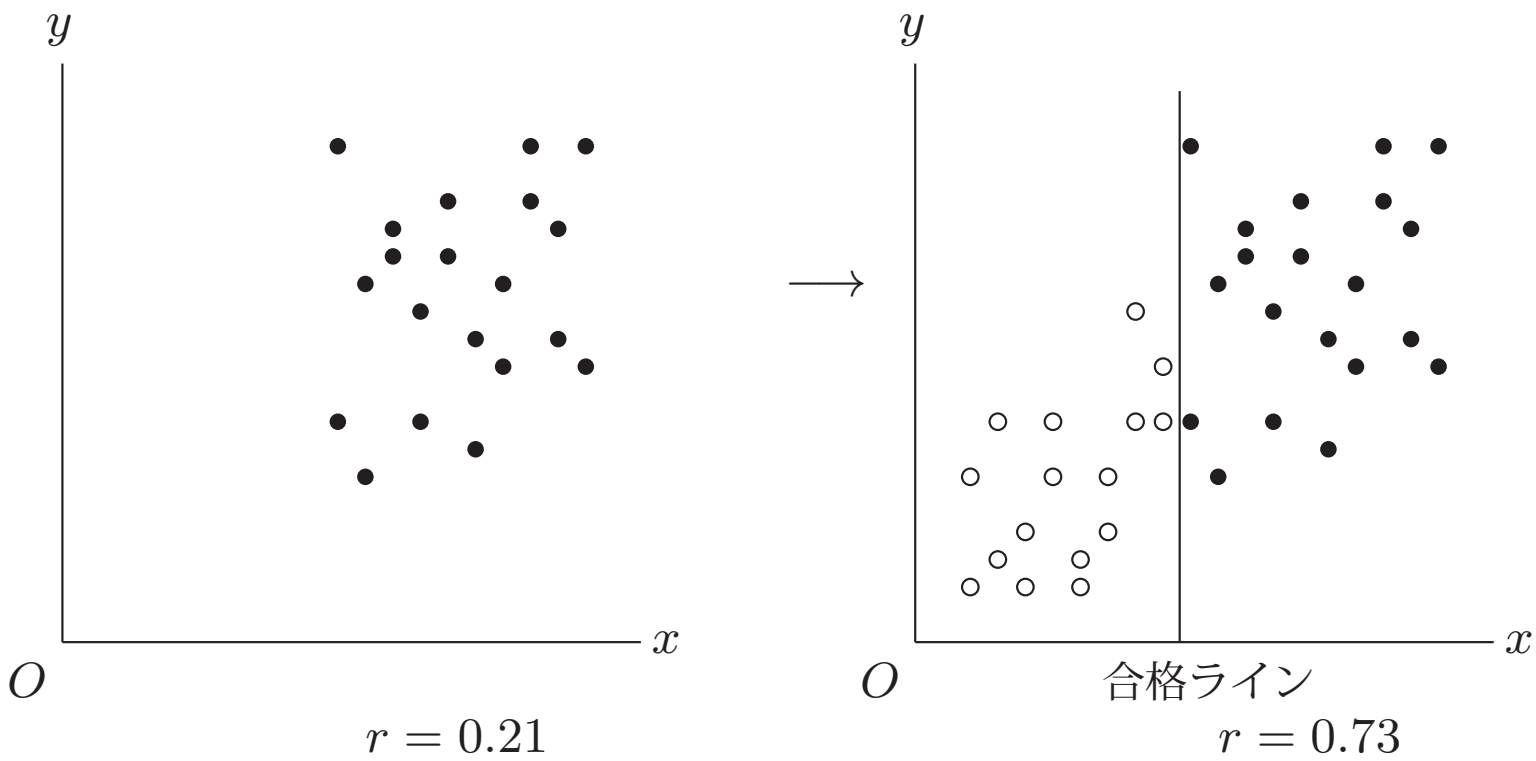
  例: 入試成績 (x) と入学後成績 (y) の相関を調べたところ，弱い正の相関（$r_{xy} = 0.21$）しか見られなかった.

  入学後の学力と入学時の学力には強い相関がない?

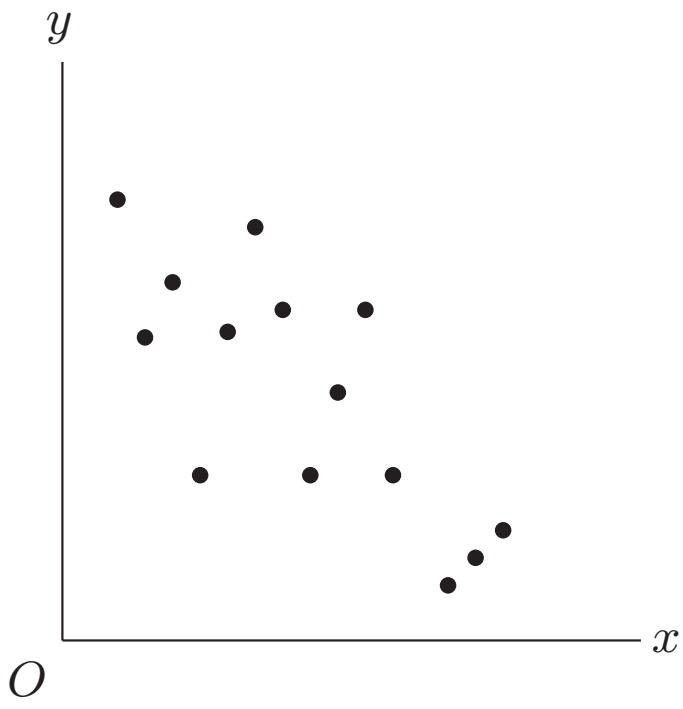


$r = 0.21$

入学後成績が測れるのは，入試成績が高得点の者だけ（**打ち切り**）．
不合格者の入学後成績が測れていたら，強い正の相関が観測できると予想できる．

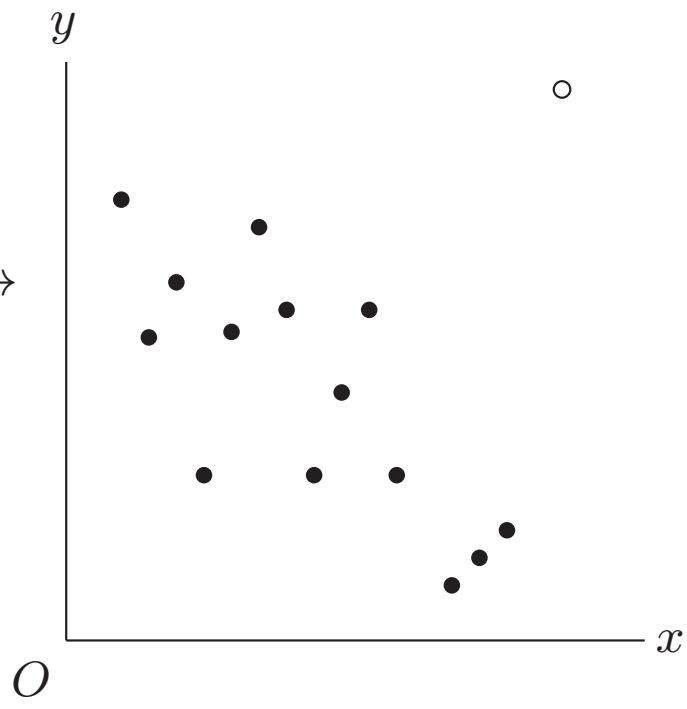


- 外れ値の影響



$r = -0.76$

$\longrightarrow$



$r = -0.29$

外れ値が 1 つあるだけで，相関係数の値は大きく変わる．

- 以上，相関係数の解釈の注意点を紹介した．

  共通して言えるのは，**相関係数の値のみを見て解釈を与えるのは危険**ということ．**必ず**散布図を描いて，打ち切りの有無，外れ値の有無，層別の必要性，等をチェックするべき．

  それでもなお，**見かけ上の相関**などの難しい問題がある．

まとめ

本講演で取り上げた項目のポイント

1. ヒストグラム，度数分布表
   - 階級幅，階級数をいくつにするかが重要．試行錯誤により，分布の特徴が読み取れるものを採用する．
   - 度数は柱の面積に比例する．
2. 平均値，中央値
   - 通常使われる代表値は平均値であるが，外れ値があるときや，分布の裾が重いときには，中央値の方が代表値として望ましい．
   - 2 つ以上の峰があるデータでは，平均値，中央値，いずれも代表値としてふさわしくない．

3. 分散，標準偏差，四分位偏差

- 通常使われる散らばりの尺度は，分散，標準偏差．特にデータの分布がベル型に近ければ，ある範囲に含まれる観測値のおおよその割合が分かる．
- 分散，標準偏差は外れ値に弱く，四分位偏差は外れ値に強い．四分位偏差を視覚的に表したものが箱ヒゲ図．

4. 相関係数

- 共分散の考え方は散布図の 4 分割．共分散を標準化したものが相関係数．
- 相関係数の解釈には注意すべき点が多い．
  - 相関関係で測れるのは「直線的な」関係のみ．
  - 見かけ上の相関．
  - 層別の失敗，打ち切りの影響，外れ値の影響．

発展: 今年の神戸大学の入試問題（後期日程，数学問 5，　抜粋）

2 つの科目 X と Y の試験を受けた 3 人の生徒の得点と，それぞれの科目の得点の平均値と標準偏差を以下のとおりとする．

	生徒 1	生徒 2	生徒 3	平均値	標準偏差
科目 X	x_1	x_2	x_3	$\bar{x}$	s_x
科目 Y	y_1	y_2	y_3	$\bar{y}$	s_y

ただし，$s_x \neq 0$ かつ $s_y \neq 0$ とする．科目 X と科目 Y の得点の相関係数は

$$r = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{3}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{3}(y_i - \bar{y})^2}}$$

で与えられる．

座標空間内に 3 点 $\mathrm{O}(0,0,0)$，$\mathrm{A}(x_1, x_2, x_3)$，$\mathrm{B}(y_1, y_2, y_3)$ を取る．O を通り，方向ベクトルが $(1,1,1)$ の直線を ℓ とする．ℓ 上の点 P を $\overrightarrow{\mathrm{PA}}$ と ℓ が垂直になるようにとり，ℓ 上の点 Q を $\overrightarrow{\mathrm{QB}}$ と ℓ が垂直になるようにとる．以下の問に答えよ．

(1) 点 P，点 Q の座標と内積 $\overrightarrow{\mathrm{PA}} \cdot \overrightarrow{\mathrm{QB}}$ を $\bar{x}, \bar{y}, s_x, s_y, r$ を用いて表せ．

(1) 略解

ℓ 上の点 P を $\mathrm{P}(t,t,t)$ とおけば，$\overrightarrow{\mathrm{PA}} \perp \ell$ より

$$(x_1-t, x_2-t, x_3-t)\cdot(1,1,1)=0 \quad \Rightarrow \quad t=\bar{x}$$

従って $\mathrm{P}(\bar{x},\bar{x},\bar{x})$．同様に $\mathrm{Q}(\bar{y},\bar{y},\bar{y})$．

$\overrightarrow{\mathrm{PA}}=(x_1-\bar{x}, x_2-\bar{x}, x_3-\bar{x})$, $\overrightarrow{\mathrm{QB}}=(y_1-\bar{y}, y_2-\bar{y}, y_3-\bar{y})$ より，

$$\|\overrightarrow{\mathrm{PA}}\|^2=\sum_{i=1}^{3}(x_i-\bar{x})^2=3s_x^2,\ \|\overrightarrow{\mathrm{QB}}\|^2=\sum_{i=1}^{3}(y_i-\bar{y})^2=3s_y^2$$

であるので，

$$\overrightarrow{\mathrm{PA}}\cdot\overrightarrow{\mathrm{QB}}=\sum_{i=1}^{3}(x_i-\bar{x})(y_i-\bar{y})=r\sqrt{3s_x^2}\sqrt{3s_y^2}=3rs_xs_y.$$

- 高校数学では $n=3$ までしか扱えませんが，大学の数学で n 次元空間を学べば，分散や相関係数を n 次元空間のベクトルのノルムや内積で特徴付けることができます．

  例えば，$\mathrm{O}(0,0,\ldots,0)$，$\mathrm{A}(x_1,x_2,\ldots,x_n)$，$\mathrm{P}(\bar{x},\bar{x},\ldots,\bar{x})$ とすれば，$\overrightarrow{\mathrm{PA}} \perp \overrightarrow{\mathrm{OP}}$ であり，分散の性質

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

  は，直角三角形 OAP に対する三平方の定理

$$\|\overrightarrow{\mathrm{PA}}\|^2 = \|\overrightarrow{\mathrm{OA}}\|^2 - \|\overrightarrow{\mathrm{OP}}\|^2$$

  （を $1/n$ 倍したもの）に他なりません．

- いま，データサイエンスが注目されており，早くから統計学に興味を持つ学生が増えています．高校数学での「データの分析」は数学 I に含まれますので，初習時は難しいですが，空間ベクトルを学んだときに，統計学とのつながりについても触れていただければ幸いに存じます．