統計の分析と利用
(旧カリ:データ分布と予測)

堀田　敬介

- 1次元のデータ
  - 度数分布・ヒストグラム
  - 代表値と散らばり
- 2次元のデータ
  - 散布図，相関関係・共分散

x	11	9	-3	14	5	23

x	11	9	-3	14	5	23
y	3	0	5	-2	7	-4

2008/4/18, Fri.

一次元のデータ

$$x = (x_1,\ x_2,\ \cdots,\ x_n)$$ （n個）

- 度数分布
- ヒストグラム
- 幹葉プロット
- 箱ひげ図

	x_1	x_2	x_3	x_4	x_5	x_6
x	11	9	-3	14	5	23

$(n = 6)$

度数分布

- データ [土日の来店客数の1年間のデータ]

292	373	282	251	322	392	366	300	226	314
325	300	356	319	213	229	244	347	283	372
253	317	306	390	287	268	257	247	318	232
306	274	231	370	275	186	327	297	260	300
285	365	272	335	167	289	352	321	341	313
319	351	299	327	405	259	376	360	259	252
339	301	337	229	244	279	243	272	211	303
316	311	287	248	199	274	286	367	317	311
434	346	329	338	319	244	329	329	274	262
288	306	189	248	344	262	385	302	366	249
250	297	292	261						

$x = (x_1, x_2, \cdots x_{104})$ $(n = 104)$

データが多すぎて**全体の傾向**がよくわからない！

度数分布

- 度数分布表[土日の来店客数の1年間のデータ]

来店客数	日　数
150–179	1
180–209	3
210–239	7
240–269	20
270–299	20
300–329	28
330–359	11
360–389	10
390–419	3
420–449	1
	0
計	104

階級 (class)
階級数:10
階級幅:30

度数 (frequency)

階級値
各階級の上限・下限値の中間値
〔例〕344.5←330-359
〔例〕345 ←330-360

全体の傾向がよくわかる！

度数分布

度数分布にすると全体の傾向がわかりやすくなるが，生データと比べて情報量が少なくなるため，このようなことがおこる．

- 度数分布表[土日の来店客数の1年間のデータ]

来店客数	日　数
150–179	1
180–209	3
210–239	7
240–269	20
270–299	20
300–329	28
330–359	11
360–389	10
390–419	3
420–449	1
計	104

階級数:10
階級幅:30

来店客数	日　数
150–199	4
200–249	15
250–299	32
300–349	36
350–399	15
400–449	2
計	104

階級数:6
階級幅:50

来店客数	日　数
160–169	1
170–179	0
180–189	2
190–199	1
200–209	0
210–219	2
220–229	3
230–239	2
240–249	8
250–259	7
260–269	5
270–279	7
280–289	8
290–299	5

階級数:28
階級幅:10

来店客数	日　数
300–309	9
310–319	11
320–329	8
330–339	4
340–349	4
350–359	3
360–369	5
370–379	4
380–389	1
390–399	2
400–409	1
410–419	0
420–429	0
430–439	1
計	104

階級数(階級幅)をどうするかが問題

度数分布

- スタージェスの公式 [階級数の目安]

$$k \approx 1 + \log_2 n = 1 + \frac{\log_{10} n}{\log_{10} 2}$$

（k: 階級数，n: データ数）

例では

$$k \approx 1 + \frac{\log_{10} 104}{\log_{10} 2} \approx 1 + \frac{2.0170}{0.3010} \approx 7.7004$$

より，階級数は8ぐらいで十分

度数分布

- 階級数8(階級幅38)で書くと…

来店客数	日数	相対度数
150–187	2	1.9
188–225	4	3.8
226–263	24	23.1
264–301	25	24.0
302–339	28	26.9
340–377	16	15.4
378–415	4	3.8
416–453	1	1.0
計	104	100.0

なるほど，週末の来店客数の全体傾向はだいたいわかったぞ

でも，度数の多い階級は全体からみてどのぐらいの割合なの？

相対度数
(relative frequency)

度数分布

- 度数分布表[相対度数]

Bさんのお店と比べて，うちのお客さんの来店傾向はどうなのか比較したいな…

来店客数	日　数	相対度数
150–179	1	1.0
180–209	3	2.9
210–239	7	6.7
240–269	20	19.2
270–299	20	19.2
300–329	28	26.9
330–359	11	10.6
360–389	10	9.6
390–419	3	2.9
420–449	1	1.0
計	104	100

来店客数	日　数	相対度数
150–179	2	1.0
180–209	6	3.0
210–239	21	10.5
240–269	24	12.0
270–299	40	20.0
300–329	54	27.0
330–359	32	16.0
360–389	13	6.5
390–419	6	3.0
420–449	2	1.0
計	200	100.0

データ数が異なる2つのグループの比較ができる

度数分布

- 累積度数分布表[累積度数，累積相対度数]

来店客数	日　数	相対度数	累積度数	累積相対度数
150–179	1	1.0	1	1.0
180–209	3	2.9	4	3.8
210–239	7	6.7	11	10.6
240–269	20	19.2	31	29.8
270–299	20	19.2	51	49.0
300–329	28	26.9	79	76.0
330–359	11	10.6	90	86.5
360–389	10	9.6	100	96.2
390–419	3	2.9	103	99.0
420–449	1	1.0	104	100.0
計	104	100.0		

累積度数
(cumulative frequency)

累積相対度数
(cumulative relative frequency)

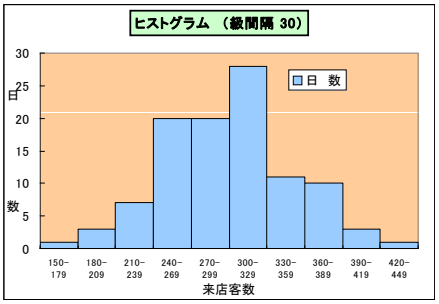
度数分布

- 問題：以下の度数分布が与えられているとき，平均来店客数を求めなさい．

来店客数	日数
150–187	2
188–225	4
226–263	24
264–301	25
302–339	28
340–377	16
378–415	4
416–453	1
計	104

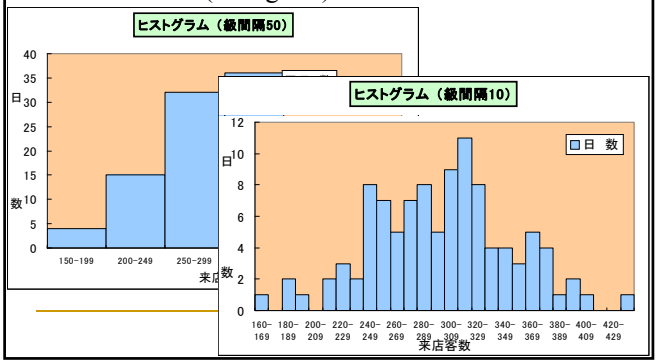
ヒストグラム

- ヒストグラム(histogram)・柱状グラフ



ヒストグラム

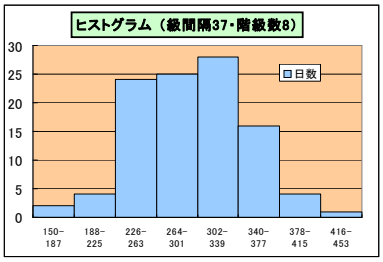
- ヒストグラム(histogram)・柱状グラフ



度数分布

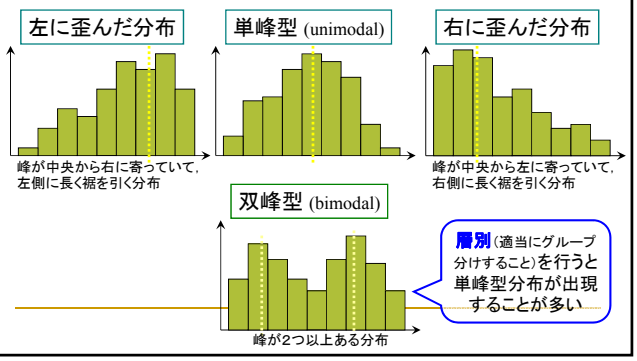
- 階級数8で書くと…

来店客数	日数
150-187	2
188-225	4
226-263	24
264-301	25
302-339	28
340-377	16
378-415	4
416-453	1
計	104



ヒストグラム

- ヒストグラムの形状



その他の手法1

幹葉プロットがヒストグラムより優れているのはどんなところ？

- 幹葉プロット，ステムプロット（stem-and-leaf diagram[plot]）
  - 野球選手の打率一覧
    - Aチーム

0.275	0.347	0.266	0.263
0.271	0.225	0.283	0.324
0.286	0.351	0.346	0.342
0.388	0.319	0.303	0.279
0.217	0.273	0.244	0.234
0.277	0.392	0.326	0.32
0.282	0.289	0.218	0.285
0.316	0.335	0.34	0.31
0.346	0.239	0.127	0.263
0.317	0.341	0.34	0.253

    - Bチーム

0.317	0.327	0.37	0.355
0.291	0.28	0.297	0.311
0.317	0.306	0.245	0.366
0.232	0.342	0.335	0.263
0.304	0.311	0.294	0.214
0.327	0.327	0.252	0.331
0.268	0.291	0.279	0.296
0.363	0.33	0.329	0.246
0.354	0.249	0.332	0.333
0.256	0.418	0.268	0.305

4	21	7 8
	22	5 7
2	23	4 9
9 6 5	24	4
6 2	25	3
8 8 3	26	3 6
9	27	1 3 5 7 9
0	28	2 3 5 6 9
7 6 4 1 1	29	
6 5 4	30	3
7 7 1 1	31	0 6 7 9
9 7 7 7 7	32	0 4 6
5 3 2 1 0	33	5
2	34	0 0 1 2 6 6 7
5 4	35	1
6 3	36	
0	37	
	38	8
	39	2
	40	
8	41	

幹 葉

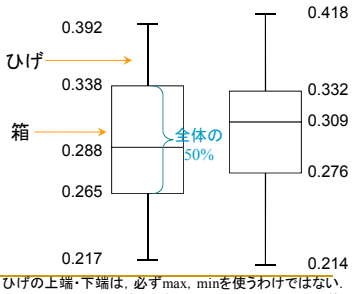
その他の手法2

- 箱ひげ図，箱型図（box plot）
  - 野球選手の打率一覧
    - Aチーム
    - Bチーム

（打率データは上と同じ）

〔Aチーム〕
max.0.392
Q_3 0.338
med.0.288
Q_1 0.265
min. 0.217

〔Bチーム〕
0.418 max.
0.332 Q_3
0.309 med.
0.276 Q_1
0.214 min.



注：ひげの上端・下端は，必ずmax，minを使うわけではない．
r:=q3-q1 としたとき，上端は区間(q3, q3+1.5r]内の最大値，
下端は区間[q1-1.5r, q1)内の最小値を用いる，など．

演習

- 男女20人の身長のデータがある．
  - 男女それぞれのデータについて，10の位までを幹，1の位を葉として幹葉プロットを描け．
  - 男女それぞれのデータについて，箱ひげ図を描け．

男	
167	176
157	155
172	178
183	178
182	181
187	188
148	159
168	173
181	177
159	169

女	
165	145
155	162
159	162
155	159
167	159
160	162
175	162
157	177
150	166
149	168

一次元のデータ

$$\boldsymbol{x} = (\underbrace{x_1,\ x_2,\ \cdots,\ x_n}_{n\text{個}})$$

- データ測定の尺度

$x_1, x_2, x_3, x_4, x_5, x_6$

x	11	9	-3	14	5	23

$(n = 6)$

データの測定尺度による分類

- 測定(measurement)と尺度(scale)
  - 名義(名目)尺度 nominal scale　質的(カテゴリ)データ
    - 属性を表す基準(対象に区別がつけられる)
      - 例:性別(男, 女, それ以外), パソコン保有(保有, 非保有)
  - 順序尺度 ordinal scale　質的(カテゴリ)データ
    - 対象間に順序がつけられる基準
      - 例:成績(A>B>C>D), 居住性(住みやすい>まあまあ>すみにくい)
  - 間隔尺度 interval scale　量的(数値)データ
    - 間隔のみが意味を持つ基準
      - 例:温度(摂氏℃, 華氏°F), 時刻(午後3時から1時間後)
  - 比率尺度 ratio scale　量的(数値)データ
    - 比が意味を持つ基準
      - 例:身長(父は子の1.5倍の背), 体重(5kg重い), 絶対温度(°K, 絶対零度)

測定が厳密 ↓

データの測定尺度による集計例

- 質的データと量的データの集計例

質的データ　量的データ

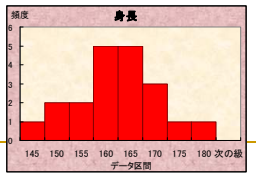
データ例

性別	成績
(男, 女)	(A, B, C, D)

女性身長				
165	155	159	155	167
160	175	157	150	149
145	162	162	159	159
162	162	177	166	168

集計例

	A	B	C	D	計
男	3	2	1	0	6
女	1	0	2	2	5
計	4	2	3	2	11



一次元のデータ

$$\boldsymbol{x} = (\underbrace{x_1,\ x_2,\ \cdots,\ x_n}_{n\text{個}})$$

- **データの代表値**
  - 算術平均
  - 幾何平均, 調和平均
  - 中央値, 最頻値
  - 四分位点
  - ミッド・レンジ

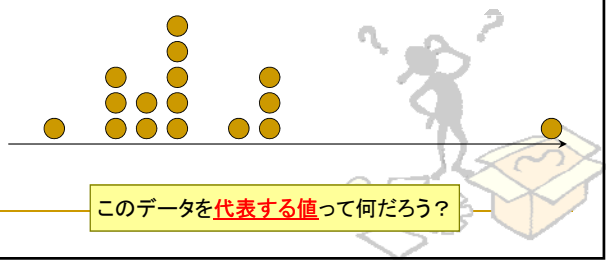
	x_1	x_2	x_3	x_4	x_5	x_6
$\boldsymbol{x}$	11	9	-3	14	5	23

$(n = 6)$

データの代表値を考える

- 例:16個のデータ

$\boldsymbol{x}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10



このデータを代表する値って何だろう？

代表値 averages

- **平均(算術平均, 相加平均) arithmetic mean**

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

$\boldsymbol{x}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10

$$\rightarrow \bar{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = \frac{1}{16}(10 + 7 + \cdots + 10) = 9.625$$

Coffee Break

- 記号の定義
  - 和を表す記号:Σ(しぐま)

$$\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$$

x_i を i を 1から n まで動かして足す

  - 積を表す記号:Π(ぱい)

$$\prod_{i=1}^{n} x_i = x_1 \times \cdots \times x_n$$

x_i を i を 1から n まで動かして掛ける

使用例)

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

$$\sum_{k=1}^{5} k = 1 + 2 + 3 + 4 + 5$$

$$\sum_{j=2}^{4} 5j = 5\cdot 2 + 5\cdot 3 + 5\cdot 4$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$$

$$\prod_{t=1}^{6} t = 1\times 2\times 3\times 4\times 5\times 6$$

代表値 averages

- **幾何平均 geometric mean**

幾何平均 = n個の積のn乗根

$$x_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \times \cdots \times x_n}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10

$$\rightarrow x_G = \sqrt[16]{\prod_{i=1}^{16} x_i}$$

$$= \sqrt[16]{10 \times 7 \times 3 \times 5 \times \cdots \times 10} \approx 7.51$$

補足：対数を利用すると計算が楽になる

$$\log x_G = \log \sqrt[n]{x_1 \times \cdots \times x_n} = \frac{\log x_1 + \cdots + \log x_n}{n}$$

☆どんなときに幾何平均が役に立つ？

例題：次の表から平均地価上昇率を求めよ

年度	2002	2003	2004	2005	2006
地価上昇率	1%	2%	3%	4%	5%

$\bar{x} = \frac{1+2+3+4+5}{5} = 3 \rightarrow 3\%$ ✗

$x_G = \sqrt[5]{1.01 \times 1.02 \times 1.03 \times 1.04 \times 1.05}$

$\approx 1.029 \rightarrow 2.9\%$ ○

代表値 averages

- **調和平均 harmonic mean**

調和平均 = 逆数の算術平均の逆数

$$x_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10

$$\rightarrow x_H = \frac{1}{\frac{1}{16}\sum_{i=1}^{16}\frac{1}{x_i}} = \frac{1}{\frac{1}{16}\sum_{i=1}^{16}\left(\frac{1}{10} + \frac{1}{7} + \cdots \frac{1}{10}\right)} \approx 6.63$$

☆どんなときに調和平均が役に立つ？

例題：行きが時速25km，帰りが時速15kmで走ったバスの平均時速を求めよ

$\bar{x} = \frac{25+15}{2} = 20 \rightarrow 20\text{km/h}$ ✗

$x_H = \frac{1}{\frac{1}{2}\left(\frac{1}{15} + \frac{1}{25}\right)} = 18.75 \rightarrow 18.75\text{km/h}$ ○

代表値 averages

- **中央値 median**
  - データをソートして，ちょうど真ん中にある値
- **最頻値 mode**
  - データの中で最も頻繁に出てくる値

補足：ソート sort とは？ データを値の小さい（大きい）順に並べること

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10
ソート後	3	5	5	5	6	6	7	7	7	7	7	9	10	10	10	50

$\rightarrow x_{med} = \frac{7+7}{2} = 7$

$\rightarrow x_{mode} = 7$

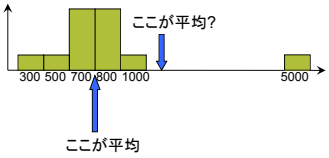
補足：データ数が偶数の場合は，中央値は真ん中2つの算術平均

補足：最も頻繁に出てくる値がない場合は最頻値はなし

代表値 averages

- 中央値や最頻値は何故必要なのか？
  - 例：年収（単位：万円）の代表値は？

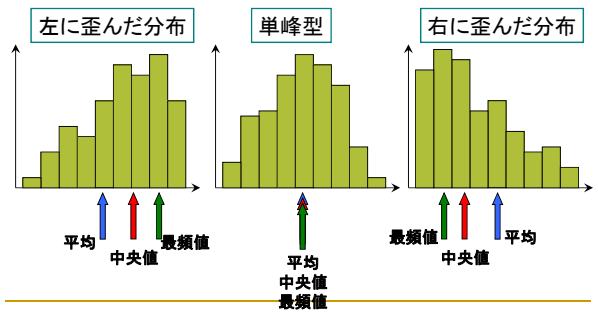
700 500 1000 800 5000 700 300 800 700 800



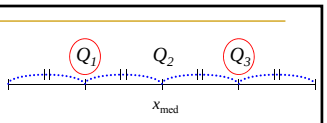
  - 算術平均
    - 1130万円
  - 中央値
    - (700+800) / 2 = 750万円
  - 最頻値
    - 700万円，800万円

代表値 averages

- 算術平均，中央値，最頻値の関係



代表値 averages



- **四分位点 quartile**
  - データをソートし，4等分したときの3つの分割点の値
    - Q_1：第1四分位点，Q_3：第3四分位点

補足：Q_2：第2四分位点は中央値 x_{med} である

  - 注意：四分位数の定義は複数ある
    - $k_1 := 0.25 \times (n\text{-}1)$, $k_3 := 0.75 \times (n\text{-}1)$ とし，

$$\begin{cases} Q_1 = x_{\lfloor k_1 \rfloor + 1} + (k_1 - \lfloor k_1 \rfloor) \times (x_{\lfloor k_1 \rfloor + 2} - x_{\lfloor k_1 \rfloor + 1}) \\ Q_3 = x_{\lfloor k_3 \rfloor + 1} + (k_3 - \lfloor k_3 \rfloor) \times (x_{\lfloor k_3 \rfloor + 2} - x_{\lfloor k_3 \rfloor + 1}) \end{cases}$$

    - $Q_1 = x_{\lfloor 0.25 \times n \rfloor}$, $Q_3 = x_{n+1-\lfloor 0.25 \times n \rfloor}$ など

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10
ソート後	3	5	5	5	6	6	7	7	7	7	7	9	10	10	10	50

※quartile：四分位数 quantile：分位数

MS Excel の 関数QUARTILE() では，Q_1=5.75, Q_3=9.25
Mathematica の関数quantile[]では，Q_1=5, Q_3=9
Rの関数quantile() では，Q_1=5.75, Q_3=9.25

代表値 averages

- **ミッド・レンジ mid-range**
  - データの最大値と最小値の算術平均

$$x_{MR} = \frac{\max\{x_1, \cdots, x_n\} + \min\{x_1, \cdots, x_n\}}{2}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10
ソート後	3	5	5	5	6	6	7	7	7	7	7	9	10	10	10	50

$$\rightarrow x_{MR} = \frac{\max(10,7,\cdots,10) + \min(10,7,\cdots,10)}{2} = \frac{50+3}{2} = 26.5$$

演習2

- 統計データを使って代表値を計算する
  - 総務省統計局（http://www.stat.go.jp）から世帯収入，世帯貯蓄などのデータを取得し，グラフ化せよ．グラフの形状はどのようになるか？
  - このデータの「算術平均」「中央値」「最頻値」を計算し，分布の代表値として最も適切だと思われるのはどれか考察せよ．
  - 「第1四分位数」「第3四分位数」「ミッドレンジ」を求めよ．
- 簡単なデータを使って代表値を計算する
  - 以下の10個のデータがある

1	20	20	22	23	24	25	26	26	50

  - 「算術平均」「中央値」「最頻値」を求めよ．
  - 「第1四分位数」「第3四分位数」「ミッドレンジ」を求めよ．

一次元のデータ

$$\boldsymbol{x} = (\underbrace{x_1, \quad x_2, \quad \cdots, \quad x_n}_{n\text{個}})$$

- **データの散らばり**
  - 範囲
  - 四分位偏差
  - 平均偏差
  - 分散，標準偏差

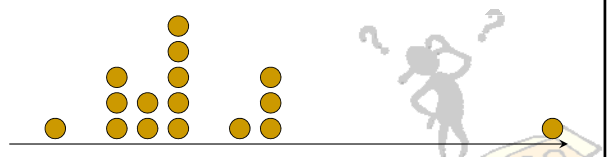
	x_1	x_2	x_3	x_4	x_5	x_6
x	11	9	-3	14	5	23

$(n = 6)$

データの値らばりを考える

- 例：16個のデータ

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10



このデータの散らばり具合はどのように測るの？

散らばりの度合いを一つの数値で示し，利用したい

散らばり dispersion

- **範囲 range**
  - 最大値と最小値の差

$$R = \max\{x_1, \cdots, x_n\} - \min\{x_1, \cdots, x_n\}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10
ソート後	3	5	5	5	6	6	7	7	7	7	7	9	10	10	10	50

$$\rightarrow R = \max(x_1, \cdots, x_{16}) - \min(x_1, \cdots, x_{16}) = 50 - 3 = 47$$

散らばり dispersion

- **四分位偏差 quartile deviation**
  - 第3四分位点 Q_3 と第1四分位点 Q_1 の差の半分

$$Q = \frac{Q_3 - Q_1}{2}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10
ソート後	3	5	5	5	6	6	7	7	7	7	7	9	10	10	10	50

$$\rightarrow Q = \frac{Q_3 - Q_1}{2} = \frac{9.75 - 5.25}{2} = 2.25$$

散らばり dispersion

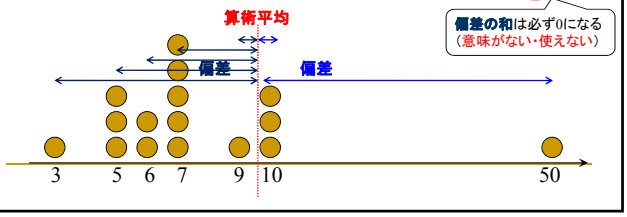
散らばり具合の度合い
＝
平均値からの平均的な差

- **偏差 deviation**
  - 各データと平均との差

$$x_i - \bar{x} \quad (i = 1, \ldots, n)$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10	9.63 平均
偏差	0.38	-2.63	-6.63	-4.63	-2.63	-4.63	0.38	-0.63	-3.63	-2.63	40.38	-2.63	-4.63	-2.63	-3.63	0.38	0.0 偏差の和

偏差の和は必ず0になる
（意味がない・使えない）



散らばり dispersion

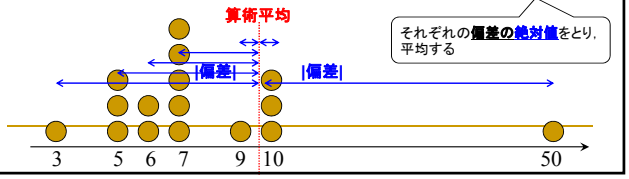
散らばり具合の度合い
＝
平均値からの平均的な差

- **平均偏差 mean deviation**
  - 偏差の絶対値の合計を平均化した値

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10	9.63 平均
偏差	0.38	-2.63	-6.63	-4.63	-2.63	-4.63	0.38	-0.63	-3.63	-2.63	40.38	-2.63	-4.63	-2.63	-3.63	0.38	0.0 偏差の和
|偏差|	0.38	2.63	6.63	4.63	2.63	4.63	0.38	0.63	3.63	2.63	40.38	2.63	4.63	2.63	3.63	0.38	5.19 平均偏差

それぞれの偏差の絶対値をとり，平均する



散らばり dispersion

補足：分散は，データの2乗の平均から平均の2乗を引いても計算できる

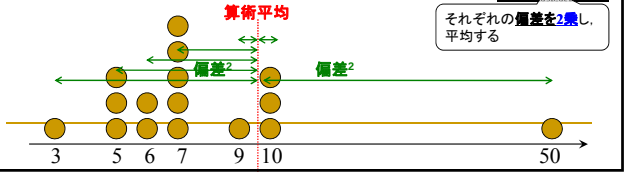
$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

- **分散 variance**
  - 偏差の2乗の合計を平均化した値

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

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10	9.63 平均
偏差	0.38	-2.63	-6.63	-4.63	-2.63	-4.63	0.38	-0.63	-3.63	-2.63	40.38	-2.63	-4.63	-2.63	-3.63	0.38	0.0 偏差の和
|偏差|	0.38	2.63	6.63	4.63	2.63	4.63	0.38	0.63	3.63	2.63	40.38	2.63	4.63	2.63	3.63	0.38	5.19 平均偏差
(偏差)^2	0.14	6.89	43.89	21.39	6.89	21.39	0.14	0.39	13.14	6.89	1630.14	6.89	21.39	6.89	13.14	0.14	112.48 分散
																	10.61 標準偏差

それぞれの偏差を2乗し，平均する



散らばり dispersion

- **標準偏差 standard deviation**
  - 分散の平方根

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

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	
データ	10	7	3	5	7	5	10	9	6	7	50	7	5	7	6	10	9.63 平均
偏差	0.38	-2.63	-6.63	-4.63	-2.63	-4.63	0.38	-0.63	-3.63	-2.63	40.38	-2.63	-4.63	-2.63	-3.63	0.38	0.0 偏差の和
|偏差|	0.38	2.63	6.63	4.63	2.63	4.63	0.38	0.63	3.63	2.63	40.38	2.63	4.63	2.63	3.63	0.38	5.19 平均偏差
(偏差)^2	0.14	6.89	43.89	21.39	6.89	21.39	0.14	0.39	13.14	6.89	1630.14	6.89	21.39	6.89	13.14	0.14	112.48 分散
																	10.61 標準偏差

分散の平方根

演習3

- 以下の10個のデータについて散らばりを計算せよ．（式だけでもよい）

1	20	20	22	23	24	25	26	26	50

  - このデータの「範囲」を計算せよ．
    - 例）data[1, 5, 7, 9, 3] → 範囲：9 – 1= 8
  - このデータの「四分位偏差」を計算せよ．
  - このデータの「偏差」をだし，合計が0になることを確かめよ．
  - このデータの「平均偏差」を計算せよ．
  - このデータの「分散」を計算せよ．
  - このデータの「標準偏差」を計算せよ．

一次元のデータ

$$\boldsymbol{x} = (\underbrace{x_1, \quad x_2, \quad \cdots, \quad x_n}_{n\text{個}})$$

- データの変換
  - 標準化（正規化）
    Cf. 偏差値

	x_1	x_2	x_3	x_4	x_5	x_6
x	11	9	-3	14	5	23

$(n = 6)$

データの一次変換

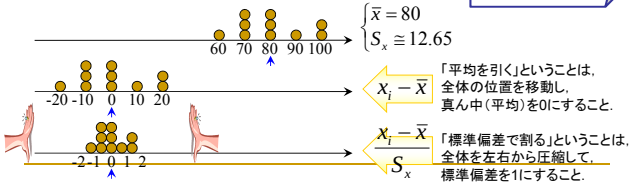
どんな1次元データも
標準化しちゃえば
同じ土俵で比較できるね！

- 標準化 standardization
  - 各データについて，平均を引き標準偏差で割る

$$z_i = \frac{x_i - \bar{x}}{S_x} \quad (i = 1, \cdots, n)$$

標準得点 standard score，Z得点

変換後のデータは
平均0，
標準偏差1
となる．



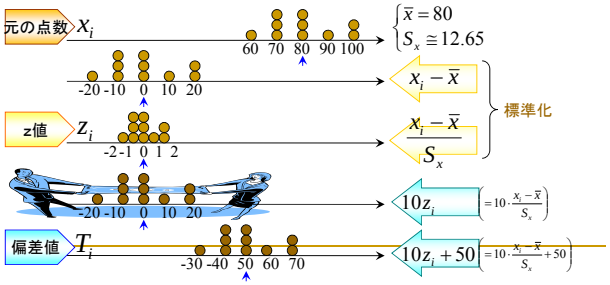
データの一次変換

変換後のデータは
平均50，
標準偏差10
となる．

- 偏差値
  - 標準得点に以下の一次変換を施す

$$T_i = 10z_i + 50 \quad (i = 1, \cdots, n)$$

偏差値得点，T得点



データの一次変換

- 例：10人の中間・期末試験の得点，z得点と偏差値

平均88，標準偏差9.8

中間試験											
	得点	100	90	80	80	90	100	80	90	100	70
	z得点	1.2	0.2	−1	−1	0.2	1.2	−1	0.2	1.2	−2
	偏差値	62	52	42	42	52	62	42	52	62	32

$1.2 = \frac{100 - 88}{9.8}$，
$62 = 1.2 \times 10 + 50$

平均33，標準偏差16

期末試験											
	得点	40	20	60	20	40	10	50	45	25	15
	z得点	0.5	−1	1.7	−1	0.5	−1	1.1	0.8	−0	−1
	偏差値	55	42	67	42	55	36	61	58	45	39

演習4

- 偏差値を計算しよう
  - 以下のデータはある試験の16人の学生の結果である．
  - 英語の結果について，各学生の得点を標準化し，z得点を出せ．
  - 国語の結果について，各学生の偏差値を計算せよ．
  - 3教科合計点について，各学生の偏差値を計算せよ．

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
英語	22	28	36	74	49	88	65	29	50	57	56	85	92	42	85	67
国語	78	50	51	33	28	23	80	97	88	66	25	72	79	44	81	29
数学	26	74	38	26	95	61	80	84	48	63	68	24	70	54	62	63

統計の分析と利用
（旧カリ：データ分布と予測）

堀田　敬介

- 一次元のデータ
  - 度数分布・ヒストグラム
  - 代表値と散らばり
- 二次元のデータ
  - 散布図，相関関係・共分散

x	11	9	-3	14	5	23

x	11	9	-3	14	5	23
y	3	0	5	-2	7	-4

二次元のデータ

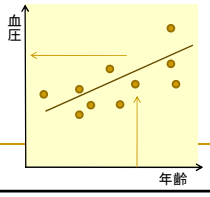
- 相関と回帰
- 共分散
- 相関係数

二次元のデータ

- 2次元データ x , y の比較
  - 相関 correlation
    - x と y との間に区別をつけず対等に見る見方・方法
      - 例：身長と体重，数学の成績と英語の成績

身長	165	175	184	172	169	170	172	168	178
体重	59	68	75	72	69	65	60	68	74

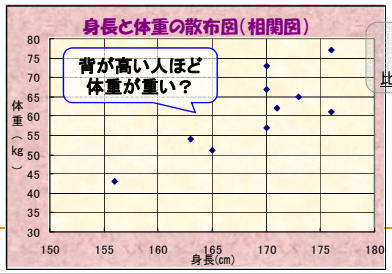
  - 回帰 regression
    - x から y を見る見方・方法
    - ある一方が他方を左右する場合
      - 例：年齢と血圧，所得と消費，人口と商業，気候と住環境



散布図 scattergram

- 2つを同時に見る
  - 例：身長と体重

身長	176	170	163	173	170	171	165	170	176	156
体重	61	73	54	65	67	62	51	57	77	43



相関関係

- 共分散 covariance

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

（2次元データ $\{x_1,\ldots,x_n\},\{y_1,\ldots,y_n\}$ について）

ある i 番目のデータについて，x_i と平均 $\bar{x}$ との差と，y_i と平均 $\bar{y}$ との差が**共に大きい**とき，共分散の値は**大きく**なり，**そうではない**とき共分散の値は**小さく**なる．すなわち，2種類のデータの**関係の強さ**を表している．

- 例：文教太郎君と湘南花子さんの昼食に掛けた費用

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき，花子さんは貧乏な食事で我慢してるの？

相関関係

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- 共分散 covariance
  - 例：文教太郎君と湘南花子さんの昼食に掛けた費用

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき，花子さんは貧乏な食事で我慢してるの？

	月	火	水	木	金	
太郎	¥400	¥300	¥100	¥200	¥200	¥240
偏差	160	60	−140	−40	−40	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	−140	−40	60	160	−40	
積	−22,400	−2,400	−8,400	−6,400	1,600	−7,600

平均

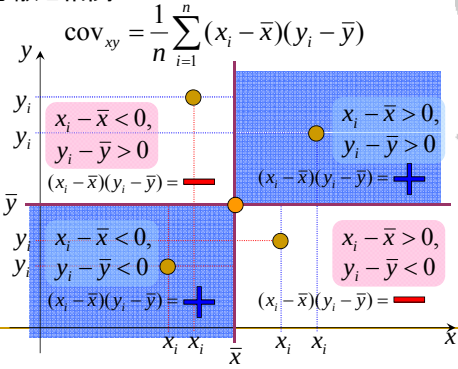
共分散

相関関係

- 共分散と相関

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

共分散って，一体何を測ってるの？



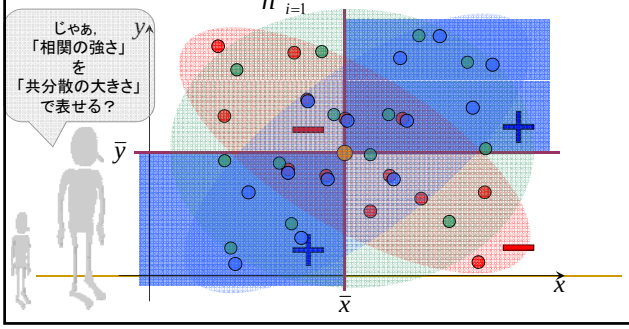
相関関係

$$\mathrm{cov}_{xy} = \begin{cases} + \longrightarrow \text{正の相関} \\ 0 \longrightarrow \text{無相関} \\ - \longrightarrow \text{負の相関} \end{cases}$$

- 共分散と相関

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

じゃあ，「相関の強さ」を「共分散の大きさ」で表せる？



相関関係

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- 共分散と関係の強さ
  - 例：文教太郎君と湘南花子さんの昼食費

	月	火	水	木	金
太郎	¥400	¥300	¥100	¥200	¥200
花子	¥100	¥200	¥300	¥400	¥200

太郎君がリッチな食事をとるとき，花子さんは貧乏な食事で我慢してるの？

  - 例：文教次郎君と湘南花子さんの昼食費

	月	火	水	木	金
次郎	¥40万	¥30万	¥10万	¥20万	¥20万
花子	¥100	¥200	¥300	¥400	¥200

超リッチな食事をとる次郎君と比べたら，花子さんの食事ってどうなの？

相関関係

測定単位が変わると，相関の度合い（強さ）が変わってしまう！

- 共分散と関係の強さ
  - 例：文教太郎君と湘南花子さんの昼食費 08dist1_資料.xls

	月	火	水	木	金	
太郎	¥400	¥300	¥100	¥200	¥200	¥240
偏差	160	60	−140	−40	−40	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	−140	−40	60	160	−40	
積	−22,400	−2,400	−8,400	−6,400	1,600	−7,600

平均　平均　共分散

	月	火	水	木	金	
次郎	¥40万	¥30万	¥10万	¥20万	¥20万	¥24万
偏差	16万	6万	−14万	−4万	−4万	
花子	¥100	¥200	¥300	¥400	¥200	¥240
偏差	−140	−40	60	160	−40	
積	−2,240万	−240万	−840万	−640万	160万	−760万

平均　平均　共分散

相関関係

（ピアソンの）積率相関係数
(Peason's) productmoment correlation coefficient

- **相関係数 correlation coefficient**

$$r_{xy} = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y} \qquad (-1 \le r_{xy} \le 1)$$

（2次元データ$\{x_1,\ldots,x_n\},\{y_1,\ldots,y_n\}$について）

$$r_{xy} = \begin{cases} 1 & \text{正の相関} \\ 0 & \text{無相関} \\ -1 & \text{負の相関} \end{cases}$$

共分散をそれぞれのデータ x_i, y_i の**標準偏差で割る**ことにより，測定単位を気にせずに，2種類のデータの**関係の強さ**を表せる．

- 注意
  - 相関係数は，2つの変数の直線的関係を見るためのもの．曲線関係が認められる場合等には向かない
  - 相関係数は，因果関係を保証するものではない．

相関関係

測定単位が変わっても，相関の度合い（強さ）は変わらない

- 共分散と関係の強さ
  - 例：文教太郎君と湘南花子さんの昼食費 08dist1_資料.xls

	月	火	水	木	金	Ave.	St.Dev.
太郎	¥400	¥300	¥100	¥200	¥200	¥240	101.98
偏差	160	60	−140	−40	−40		
花子	¥100	¥200	¥300	¥400	¥200	¥240	101.98
偏差	−140	−40	60	160	−40		
積	−22,400	−2,400	−8,400	−6,400	1,600	−7,600	−0.731
						Cov.	Corr.

	月	火	水	木	金	Ave.	St.Dev.
次郎	¥40万	¥30万	¥10万	¥20万	¥20万	¥24万	101,980
偏差	16万	6万	−14万	−4万	−4万		
花子	¥100	¥200	¥300	¥400	¥200	¥240	101.98
偏差	−140	−40	60	160	−40		
積	−2,240万	−240万	−840万	−640万	160万	−760万	−0.731
						Cov.	Corr.

相関関係

★順位相関係数を使うときとは？
データが選好順位（順序尺度）で与えられている場合

A	R_1,	R_2,	…,	R_n
B	Q_1,	Q_2,	…,	Q_n

（R_i：A が i を好きな順番）

- 参考：その他の相関係数
  - **（スピアマンの）順位相関係数 rank correlation coefficient**

$$r_S = 1 - \frac{6}{n^3-n}\sum_{i=1}^{n}(R_i-Q_i)^2 \qquad (-1 \le r_S \le 1)$$

順位が完全に一致しているとき $r_S=+1$
順位が完全に逆のとき $r_S=-1$

  - **（ケンドールの）順位相関係数 rank correlation coefficient**

$$r_K = \frac{G-H}{n(n-1)/2} \qquad (-1 \le r_K \le 1)$$

順位が完全に一致しているとき $r_K=+1$
順位が完全に逆のとき $r_K=-1$

  - 偏相関係数 partial correlation coefficient
  - （時系列データに対する）自己相関係数
    auto-correlation coefficient

相関関係

★順位相関係数を使うときとは？
データが選好順位（順序尺度）で与えられている場合

A	R_1,	R_2,	…,	R_n
B	Q_1,	Q_2,	…,	Q_n

（R_i：A が i を好きな順番）

- 参考：その他の相関係数
  - 例題：男女それぞれが好きな花の順番

	桜	菊	薔薇	梅	百合	鬱金香	カーネーション	椿
男	1	2	3	4	5	6	7	8
女	3	1	2	5	4	7	6	8

出展：（『統計学入門』p.55）

☆（スピアマンの）順位相関係数

$$r_s = 1-\frac{6}{n^3-n}\sum_{i=1}^{n}(R_i-Q_i)^2$$
$$= 1-\frac{6}{8^3-8}\left\{(1-3)^2+(2-1)^2+\cdots+(8-8)^2\right\}$$
$$= 1-\frac{1}{84}\cdot 10 = \frac{37}{42} \approx 0.881$$

ピアソンの積率相関係数を順序尺度に素直にあてはめたもの

☆（ケンドールの）順位相関係数

	菊	薔	梅	百	鬱	カ	椿
桜	×	×	○	○	○	○	○
菊		○	○	○	○	○	○
薔			○	○	○	○	○
梅				×	○	○	○
百					○	○	○
鬱						×	○
カ							○

桜 v.s. 椿
★男：1<8
★女：3<8
正順

鬱 v.s. カ
★男：6<7
★女：7>6
逆順

G：正順の数=24
H：逆順の数=4

全対（$n(n-1)/2$個）について，正順と逆順の個数の差を比較したもの

$$r_K = \frac{G-H}{n(n-1)/2} = \frac{24-4}{8(8-1)/2} = \frac{5}{7} \approx 0.714$$

演習5

- 相関係数を計算しよう
  - 右のデータ x, y について，
    - それぞれの分散 S_x^2, S_y^2 を計算せよ．
    - 共分散 cov_{xy} を計算せよ．
    - （ピアソンの積率）相関係数 r_{xy} を計算せよ．

x	1	3	5	7	9
y	4	6	2	0	3

  - 右のA君，Bさんの色の好みに関する選好順位データについて，
    - （スピアマンの）順位相関係数 r_S を計算せよ．
    - （ケンドールの）順位相関係数 r_K を計算せよ．

	赤	青	橙	緑	紫
A	1	2	3	4	5
B	4	5	2	1	3

参考：散らばりの比較

- 変動係数 coefficient of variation
  - 分布の中心が著しく異なる場合，分散で単純に散らばりを比較できない ⇒ 相対比を指標として用いる

$$C.V. = \frac{S_x}{\bar{x}}$$

（n個の観測値 $x_1, \cdots, x_n$ に対して）

  - 例：県民所得（単位：万円）の比較

	県民所得	
	平均	標準偏差
1965年	26.6	7.5
1975年	117.5	23.8

単純には所得格差は3倍に広がっているように見える

1965年：7.5/26.6 = 0.28 (28%)
1975年：23.8/117.5 = 0.20 (20%)

最後に…

- 統計解析・予測手法

記述統計学 descriptive statistics ← 度数分布，代表値，散らばり，相関関係，etc.

推測統計学 inferential statistics ← 確率分布，母集団・標本，推定，検定，etc.

多変量解析 multivariate analysis ← 重回帰分析，主成分分析，判別分析，数量化理論，etc.

参考文献

- 東大教養統計教室編「統計学入門」東大出版会(1991)
- 東大教養統計教室編「自然科学の統計学」東大出版会(1992)
- 村上雅人「なるほど統計学」海鳴社(2002)
- 大村平「改訂版 統計解析のはなし」日科技連(2006,1980)
- 大村平「QC数学のはなし」日科技連(2003)
- 丹慶勝市「図解雑学 統計解析」ナツメ社(2003)
- 高橋信「マンガでわかる統計学」オーム社(2004)
- 田栗正章ほか「やさしい統計入門」 講談社(2007)
- 桑田秀夫「経営・経済系のための統計学」日科技連(1992)
- J.アルバート&J.ベネット「メジャーリーグの数理科学」シュプリンガー(2004)
- 間瀬茂他「工学のためのデータサイエンス入門」数理工学社(2004)
- 荒木勉他「Excelで学ぶ統計解析」実教出版(2000)