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

2006/9/29,Fri.

1次元のデータ

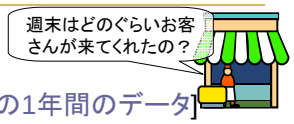
$$\boldsymbol{x} = (\underbrace{x_1,\ x_2,\ \cdots,\ x_n}_{n\text{個}})$$

- 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						

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

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

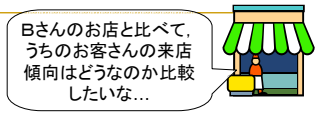
- 階級数8(階級幅38)で書くと…

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

相対度数
(relative frequency)

度数分布

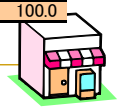
- 度数分布表[相対度数]



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)

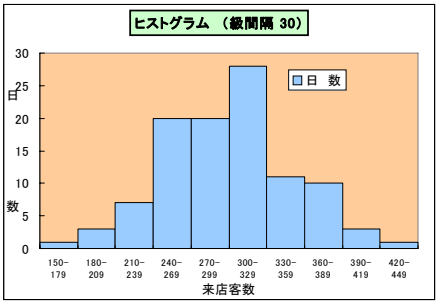
度数分布

- 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

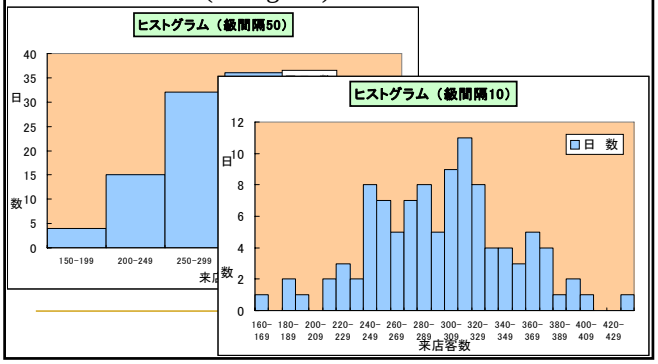
ヒストグラム

- ヒストグラム(histogram)・柱状グラフ



ヒストグラム

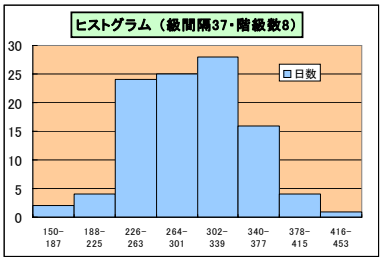
- ヒストグラム(histogram)・柱状グラフ



度数分布

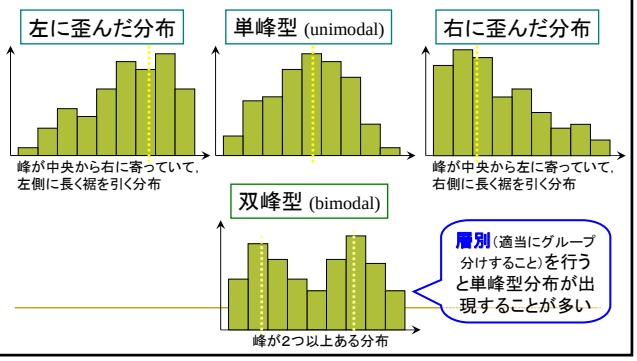
- 階級数8で書くと…

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

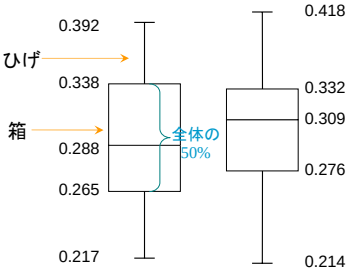
その他の手法2

- 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

（打率一覧は「その他の手法1」と同じ）

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

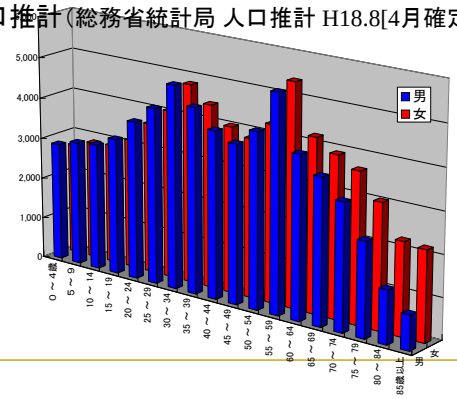
〔Bチーム〕
0.418 max.
0.332 Q_3
0.309 med.
0.276 Q_1
0.214 min.



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．

例題1

- 人口推計（総務省統計局 人口推計 H18.8[4月確定値]）



演習1

- 度数分布・ヒストグラムを作成しよう
  - 以下に20個のデータがある．これより，度数分布を作成せよ．
  - 作成した度数分布に相対度数・累積度数を付加せよ．
  - 作成した度数分布をもとにヒストグラムを作成せよ．
- 幹葉プロットを作成しよう
  - 以下の20個のデータから幹葉プロット（ステムプロット）を作成せよ．

17.63	13.78	17.16	13.71	20.96	12.23	7.75	11.26	9.40	6.60
5.48	13.23	9.72	2.17	8.50	8.47	9.62	9.96	8.68	11.80

1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

データの代表値を考える

- 例:16個のデータ

$$\boldsymbol{x} = (\underbrace{x_1,\ x_2,\ \cdots,\ x_{16}}_{16\text{個}})$$

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?

1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
x	11	9	-3	14	5	23

$(n = 6)$

代表値 averages

- 平均(算術平均, 相加平均) arithmetic mean

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

(n個の観測値 $x_1, \cdots, x_n$ に対して)

例:

10	7	3	5	7	5	10	9
6	7	50	7	5	7	6	10

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

例2:

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

$$\bar{x} = \frac{1}{104}(168.5\times 2 + 206.5\times 4 + \cdots + 434.5\times 1) \approx 296.4$$

代表値 averages

参:p.28

例:

10	7	3	5	7	5	10	9
6	7	50	7	5	7	6	10

- 幾何平均 geometric mean

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

$$x_G = \sqrt[16]{\prod_{i=1}^{16} x_i} = \sqrt[16]{10\times 7\times \cdots \times 10} \approx 7.51$$

(n個の観測値 $x_1, \cdots, x_n$ に対して)

参考:対数を利用すると積を和で計算できる!

$$\log x_G = \log \sqrt[n]{x_1 \times \cdots \times x_n} = \frac{1}{n}\{\log x_1 + \cdots + \log x_n\}$$

平均地価上昇率を求めてみよう

地価上昇率	
'83-'84年	21.8%
'84-'85年	30.5%
'85-'86年	53.6%
'86-'87年	50.0%
'87-'88年	12.9%

$$x_G = \sqrt[5]{1.218\times 1.305\times 1.536\times 1.5\times 1.129} \approx 1.328 \rightarrow 32.8\%$$

代表値 averages

例: 10 7 3 5 7 5 10 9 / 6 7 50 7 5 7 6 10

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

- 調和平均 harmonic mean

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

（n個の観測値 $x_1,\cdots,x_n$ に対して）

調和平均は，逆数の算術平均の逆数

バスの往復時平均時速を求めてみよう！
[行き：時速25km，帰り：時速15km]⇒平均時速は？

$$x_H = \frac{1}{\frac{1}{2y}\left(\frac{y}{15}+\frac{y}{25}\right)} = 18.75$$

代表値 averages

- 中央値 median
  - データをソートしたとき，ちょうど真ん中に来る値

(sort, 値の小さい(大きい)順に並べること)

- 最頻値 mode
  - データの中で最も頻繁に出てくる値

例: 10 7 3 5 7 5 10 9 / 6 7 50 7 5 7 6 10 → 3 5 5 5 6 6 7 7 / 7 7 7 9 10 10 10 50

元のデータ　ソート後のデータ

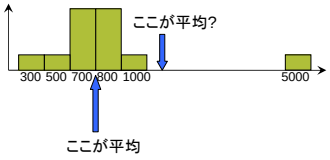
あれれ？
データ数が偶数のときはどうするの？

median 7
mode 7

代表値 averages

- 算術平均，中央値，最頻値の関係
  - 例：年収(単位：万円)の代表値は？

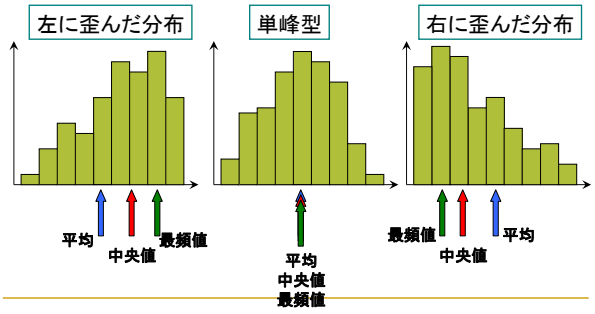
700 500 1000 800 5000 700 300 800 700 800



  - 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
    - 第1四分位点 Q_1
    - 第2四分位点 Q_2 ＝中央値(Median)
    - 第3四分位点 Q_3
  - 四分位数の定義はいくつかある
    - n個のデータ$(x_1,x_2,\ldots,x_n)$について，
      - $k := p\times(n\text{-}1)$ とし，
        $Q_i = x_{\lfloor k\rfloor+1} + (k-\lfloor k\rfloor)\times(x_{\lfloor k\rfloor+2} - x_{\lfloor k\rfloor+1})$
        （第1四分位の時 p=0.25，第3四分位の時 p=0.75）
      - $Q_1 = x_{\lfloor 0.25\times n\rfloor}$，$Q_3 = x_{n+1-\lfloor 0.25\times n\rfloor}$
      - ..., etc.

※quartile：四分位数
quantile：分位数

例 10 7 3 5 7 5 10 9 / 6 7 50 7 5 7 6 10

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中間点

$$x_{MR} = \frac{1}{2}\{\max(x_1,\cdots,x_n)+\min(x_1,\cdots,x_n)\}$$

（n個の観測値 $x_1,\cdots,x_n$ に対して）

例: 10 7 3 5 7 5 10 9 / 6 7 50 7 5 7 6 10

$$x_{MR} = \frac{1}{2}\{\max(10,7,\cdots,10)+\min(10,7,\cdots,10)\}$$
$$= \frac{1}{2}(50+3) = 26.5$$

演習2

- 代表値を計算しよう
  - 総務省統計局（http://www.stat.go.jp）の [家計調査]−[貯蓄・負債編]−[調査結果]−[詳細結果表：年度平均]−[表8-11：貯蓄・純貯蓄・負債現在高階級別]から，貯蓄額を世帯ごとの度数で表したデータを取得し，グラフ化せよ．グラフの形状はどのようになるか？
  - 次に，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．ただし，100万未満，及び4000万以上の階級値はそれぞれ50万，5000万とする．
  - 中央値以外の四分位点と，ミッドレンジを計算せよ．
- 以下の10個のデータについて「算術平均」「中央値」「最頻値」「第1四分位数」「第3四分位数」「ミッドレンジ」を求めよ．

1 20 20 22 23 24 25 26 26 50

1次元のデータ

n個

$$\boldsymbol{x} = (x_1,\ x_2,\ \cdots,\ x_n)$$

- 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

16個

$$\boldsymbol{x} = (x_1,\ x_2,\ \cdots,\ x_{16})$$

- 例：16個のデータ

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 dispersion

- 範囲 range
  - データの最大値と最小値の差

$$R = \max(x_1, \cdots, x_n) - \min(x_1, \cdots, x_n)$$

（n個の観測値 $x_1, \cdots, x_n$ に対して）

- 四分位偏差 quartile deviation
  - 第3四分位点と第1四分位点の隔たりの半分

$$Q = \frac{1}{2}(Q_3 - Q_1)$$

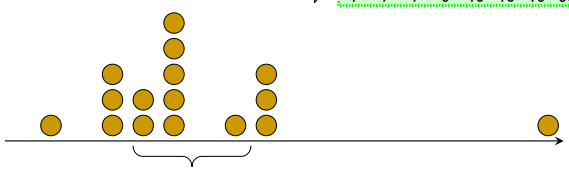
（n個の観測値 $x_1, \cdots, x_n$ に対して）

散らばり dispersion

- 例：16個のデータ

10 7 3 5 7 5 10 9
6 7 50 7 5 7 6 10

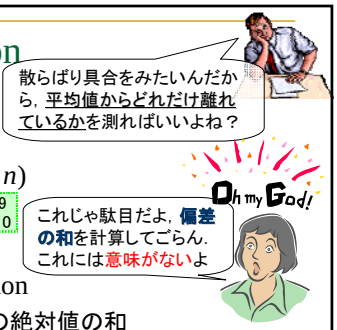
⇒ 3 5 5 5 6 6 7 7
7 7 7 9 10 10 10 50



四分位偏差 (9.75 – 5.25) / 2 = 2.25

範囲 50 – 3 = 47

散らばり dispersion



- 偏差 deviation
  - 各データと平均との差

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

例：10 7 3 5 7 5 10 9
6 7 50 7 5 7 6 10

- 平均偏差 mean deviation
  - 各データと平均との差の絶対値の和

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

（n個の観測値 $x_1, \cdots, x_n$ に対して）

例：10 7 3 5 7 5 10 9
6 7 50 7 5 7 6 10

散らばり dispersion

注：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

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

- 分散 variance
  - 各データと平均との差の2乗和

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (n\text{個の観測値 } x_1, \cdots, x_n \text{ に対して})$$

例：

10	7	3	5	7	5	10	9
6	7	50	7	5	7	6	10

$$S^2 = \frac{1}{16}\sum_{i=1}^{16}(x_i - \bar{x})^2 = \frac{1}{16}\{(10-9.625)^2 + \cdots\} \approx 112.48$$

- 標準偏差 standard deviation
  - 分散の平方根

絶対値はめんどくさいから嫌だ！

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (n\text{個の観測値 } x_1, \cdots, x_n \text{ に対して})$$

例：

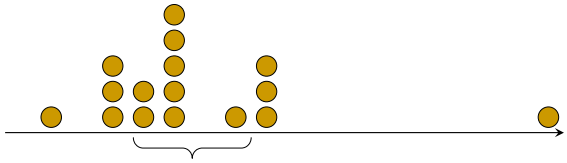
10	7	3	5	7	5	10	9
6	7	50	7	5	7	6	10

$$S \approx \sqrt{112.48} \approx 10.61$$

散らばり dispersion

- 例：16個のデータ

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	
	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
	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



四分位偏差 (9.75 – 5.25) / 2 = 2.25

範囲 50 – 3 = 47

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．

1次元のデータ

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

- データの変換
  - 標準化（正規化）

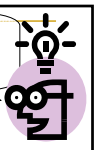
Cf. 偏差値

	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！

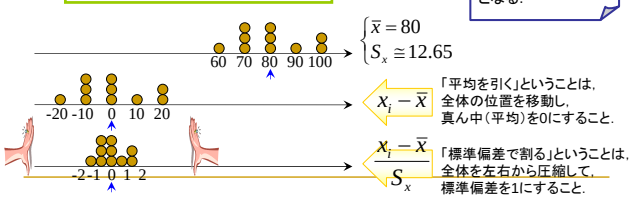


- 標準化 standardization
  - 各データについて，平均を引き標準偏差で割る

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

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．



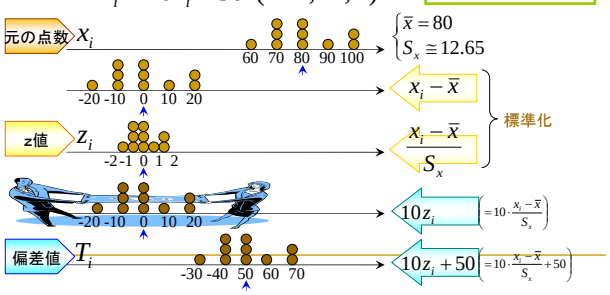
データの一次変換

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

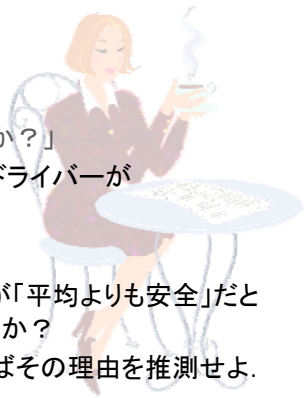
期末試験											
	得点	40	20	60	20	40	10	50	45	25	15
	z得点	0.5	−1	1.7	−1	0.5	−1	1.1	0.8	−0	−1
	偏差値	55	42	67	42	55	36	61	58	45	39

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

Coffee Break !



- 車のドライバーに
  「あなたは安全運転か？」
  と尋ねたところ，大半のドライバーが
  「平均以上です」
  と答えた．
- さて，大半のドライバーが「平均よりも安全」だということがあり得るだろうか？
- もしあり得るのだとすればその理由を推測せよ．

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

2次元のデータ

- 相関と回帰
- 共分散
- 相関係数

2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
  - 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

相関の**度合い**を**数値化**することで，比較・分析できないか？

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次元のデータ $x_1,\cdots,x_n$，$y_1,\cdots,y_n$ 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

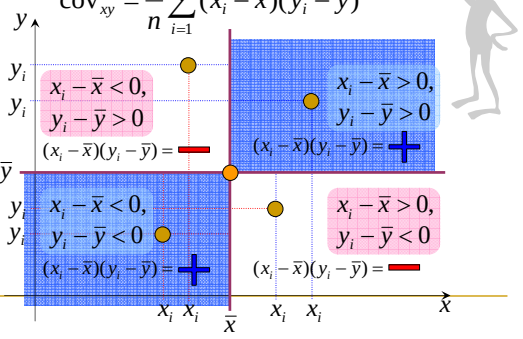
太郎君がリッチな食事をとるとき，花子さんは貧乏な食事で我慢してるの？

相関関係

- 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？



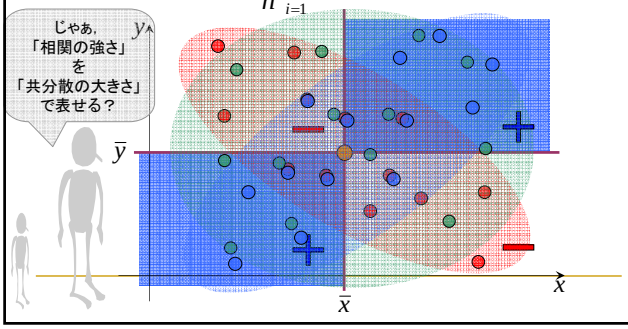
相関関係

- 共分散と相関

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

cov_{xy} ＝ ＋ → 正の相関，0 → 無相関，− → 負の相関

じゃあ，「相関の強さ」を「共分散の大きさ」で表せる？



相関関係

- 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？

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

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？

  - 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

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！

06dist1_資料.xls

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

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

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

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，測定単位を気にしなくても，2種類のデータの**関係の強さを表している**．

- 参考：その他の相関係数
  - 偏相関係数 partial correlation coefficient
  - （スピアマンの）順位相関係数 rank correlation coefficient
  - （ケンドールの）順位相関係数 rank correlation coefficient
  - 時系列データに対する自己相関係数 auto-correlation coefficient

演習5

- 相関を計算しよう
  - 以下のデータはある試験の結果である.
  - 英語と国語，英語と数学，国語と数学の結果について，それぞれ共分散を計算せよ.
  - 英語と国語，英語と数学，国語と数学の結果について，それぞれ(ピアソンの積率)相関係数を計算せよ.
  - 各教科のテスト結果に相関はみられるか？
    (例えば，国語ができる学生は英語もできるとか，数学ができる学生は国語ができない，など)

	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

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

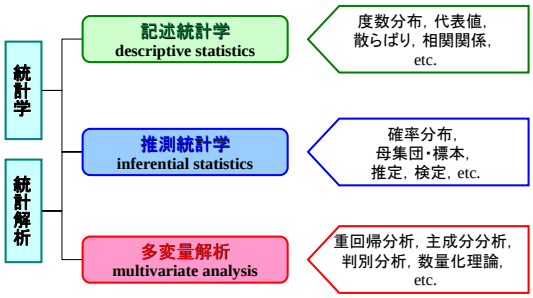
	県民所得	
	平均	標準偏差
1965年	26.6	7.5
1975年	117.5	23.8

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



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)
- 宮川公男他 「入門経営科学」 実教出版(1999)
- 荒木勉他「Excelで学ぶ統計解析」実教出版(2000)
- 荒木勉「Excelで学ぶ需要予測」実教出版(2000)
- 多田実他「Excelで学ぶ経営科学」オーム社(2003)
- 村上雅人「なるほど統計学」海鳴社(2002)
- 丹慶勝市「図解雑学 統計解析」ナツメ社(2003)
- 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，加藤貴昭訳「メジャーリーグの数理科学 上・下」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)