統計の分析と利用

1. データとその扱い Part Ⅱ

堀田 敬介

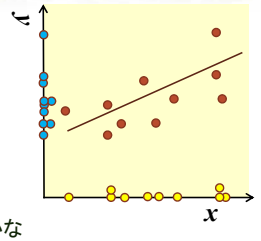
1-1. 一次元のデータ
- 1変数の図化：度数分布とヒストグラム，幹葉プロット，箱ひげ図
- 1変数の数表現：代表値と散らばり，データの標準化
- データの尺度

1-2. 二次元のデータ
- 2変数の関係1：散布図，共分散・相関係数
- 2変数の関係2：クロス集計，クラメルの連関係数
- 2変数の関係3：点グラフ，相関比

2013/10/11, Fri.～

2次元のデータ

- データが2つになるコトの意味
  - 1次元のデータ：1変数x
    - データの分布はどうなっているかな
    - 代表値は？ 散らばり具合は？
  - 2次元のデータ：2変数x,y
    - 変数xのデータの分布はどうなっているかな
    - 変数xの代表値は？ 散らばり具合は？
    - 変数yのデータの分布はどうなっているかな
    - 変数yの代表値は？ 散らばり具合は？
    - 変数xとyの関係（相関？因果？）はどうかな？

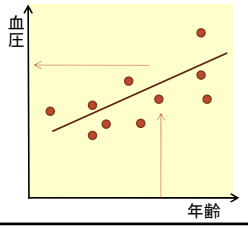


2次元のデータ

- 2変数x,yの相関関係
  - 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，単なる関係
      - 例：数学の成績と英語の成績

	A	B	C	D	E	F	G	H	I
数学 x	65	75	84	72	69	70	72	68	78
英語 y	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



One Point:
相関関係
常に成り立つ
成り立つとは限らない
因果関係

2次元のデータ

- 2変数x,yの相関関係を調べる方法（図と式）

尺度

- 例1

	A	B	C	D	E	F	G	H	I	J	
性別 x	男	男	女	男	男	男	女	女	男	女	質的
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的

クロス集計 — 連関係数

- 例2

	A	B	C	D	E	F	G	H	I	J	
飲量 x	15	32	16	30	50	12	14	24	18	19	量的
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的

点グラフ — 相関比

- 例3

	A	B	C	D	E	F	G	H	I	J	
身長 x	176	170	163	173	170	171	165	170	176	156	量的
体重 y	61	73	54	65	67	62	51	57	77	43	量的

散布図 — 相関係数

2変数の関係

- 2変数の関係1：x(質的)×y(質的)図

	A	B	C	D	E	F	G	H	I	J	
性別 x	男	男	女	男	男	男	女	女	男	女	質的
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的

クロス集計

	紅茶	緑茶	珈琲	計	
男	1	2	3	6	← 周辺度数
女	2	0	2	4	
計	3	2	5	10	← 総度数
	↑	↑	↑		
	周辺度数				

2変数の関係

- 2変数の関係1：x(質的)×y(質的)式
  - クラメルの連関係数 *Cramer's coefficient of association*

連関係数

	紅茶	緑茶	珈琲	計
男	1	2	3	6
女	2	0	2	4
計	3	2	5	10

クロス集計から理論度数を求める

	紅茶	緑茶	珈琲	計
男	1.8	1.2	3.0	6
女	1.2	0.8	2.0	4
計	3	2	5	10

$1.8=\frac{3\cdot 6}{10}$　$2.0=\frac{5\cdot 4}{10}$

$$V=\sqrt{\frac{\chi^2}{n\cdot m}} \qquad (0\le V\le 1)$$

$$\chi^2=\frac{(1-1.8)^2}{1.8}+\frac{(2-1.2)^2}{1.2}+\cdots+\frac{(0-0.8)^2}{0.8}+\frac{(2-2.0)^2}{2.0}$$
$$n=10$$
$$m=\min\{2-1,3-1\}$$

ピアソンの χ^2 統計量

(行数-1)と(列数-1)の小さい方

2変数の関係

- 2変数の関係1：x(質的)×y(質的)式
  - クラメルの連関係数 *Cramer's coefficient of association*

	紅	緑	珈	計
男	0	3	9	12
女	6	0	0	6
計	6	3	9	18

$$\chi^2=\frac{(0-4)^2}{4}+\frac{(3-2)^2}{2}+\frac{(9-6)^2}{6}+\frac{(6-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(0-3)^2}{3}=18$$
$$n=18$$
$$m=\min\{2-1,3-1\}=1$$
$$\rightarrow V=\sqrt{\frac{18}{18\cdot 1}}=1$$ 嗜好と性別は**完全相関**

	紅	緑	珈	計
男	3	1	8	12
女	3	2	1	6
計	6	3	9	18

$$\chi^2=\frac{(3-4)^2}{4}+\frac{(1-2)^2}{2}+\frac{(8-6)^2}{6}+\frac{(3-2)^2}{2}+\frac{(2-1)^2}{1}+\frac{(1-3)^2}{3}=17/4$$
$$n=18$$
$$m=\min\{2-1,3-1\}=1$$
$$\rightarrow V=\sqrt{\frac{17/4}{18\cdot 1}}\approx 0.49$$ 嗜好と性別は**多少相関**

	紅	緑	珈	計
男	4	2	6	12
女	2	1	3	6
計	6	3	9	18

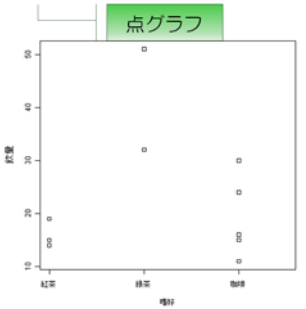
$$\chi^2=\frac{(4-4)^2}{4}+\frac{(2-2)^2}{2}+\frac{(6-6)^2}{6}+\frac{(2-2)^2}{2}+\frac{(1-1)^2}{1}+\frac{(3-3)^2}{3}=0$$
$$n=18$$
$$m=\min\{2-1,3-1\}=1$$
$$\rightarrow V=\sqrt{\frac{0}{18\cdot 1}}=0$$ 嗜好と性別は**無相関**

2変数の関係

- 2変数の関係2：x(量的)×y(質的)図

	A	B	C	D	E	F	G	H	I	J	
飲量 x	15	32	16	30	50	12	14	24	18	19	量的
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的

点グラフ



2変数の関係

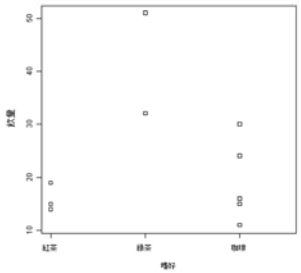
- 2変数の関係2：x(量的)×y(質的)式

	A	B	C	D	E	F	G	H	I	J	
飲量 x	15	32	16	30	50	12	14	24	18	19	量的
嗜好 y	紅茶	緑茶	珈琲	珈琲	緑茶	珈琲	紅茶	珈琲	珈琲	紅茶	質的

相関比

- 相関比 ***correlation ratio***

$$\eta^2 = \frac{S_T}{S_B + S_T} \qquad (0 \le \eta^2 \le 1)$$



2変数の関係

- 2変数の関係2：x(量的)×y(質的)式
  - 相関比 ***correlation ratio***

$$\eta^2 = \frac{S_T}{S_B + S_T} \quad (0 \le \eta^2 \le 1) \qquad \eta^2 = \frac{840}{376 + 840} \approx 0.691$$

	紅茶	緑茶	珈琲	
	14	32	12	
	15	50	16	
	19		18	
			24	
			30	
個数	3	2	5	全平均
平均	16	41	20	23
偏差平方	49	324	9	840 S_T
				級間変動

$49 = (16-23)^2$
$324 = (41-23)^2$
$9 = (20-23)^2$
$S_T = \underline{840} = 49 \times 3 + 324 \times 2 + 9 \times 5$
級間変動
＝級平均と全平均との偏差平方の加重和

偏差平方	4	81	64	
	1	81	16	
	9		4	
			16	
			100	合計
計	14	162	200	376 S_B
				級内変動

$14 = (14-16)^2 + (15-16)^2 + (19-16)^2$
$162 = (32-41)^2 + (50-41)^2$
$200 = (12-20)^2 + (16-20)^2 + \ldots + (30-20)^2$
$S_B = \underline{376} = 14 + 162 + 200$
級内変動
＝級内データと級平均との偏差平方の和

2変数の関係

- 2変数の関係2：x(量的)×y(質的)式
  - 相関比 ***correlation ratio***

$$\eta^2 = \frac{840}{0 + 840} = 1 \qquad \eta^2 = \frac{840}{376 + 840} \approx 0.691 \qquad \eta^2 = \frac{0}{314 + 0} = 0$$

嗜好と飲量は**完全相関**

	紅茶	緑茶	珈琲	
	16	41	20	
	16	41	20	
	16		20	
			20	
			20	
個数	3	2	5	全平均
平均	16	41	20	23
偏差平方和	49	324	9	840
				級間変動
偏差平方和	0	0	0	
	0	0	0	
	0		0	
			0	
			0	合計
計	0	0	0	0
				級内変動

嗜好と飲量は**多少相関**

	紅茶	緑茶	珈琲	
	14	32	12	
	15	50	16	
	19		18	
			24	
			30	
個数	3	2	5	全平均
平均	16	41	20	23
偏差平方和	49	324	9	840
				級間変動
偏差平方和	4	81	64	
	1	81	16	
	9		4	
			16	
			100	合計
計	14	162	200	376
				級内変動

嗜好と飲量は**無相関**

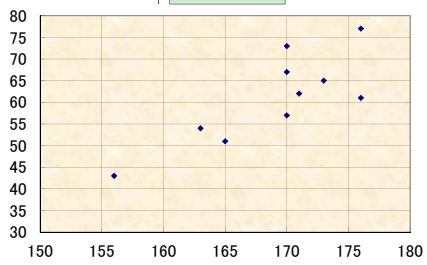
	紅茶	緑茶	珈琲	
	19	15	15	
	21	31	20	
	29		25	
			25	
			30	
個数	3	2	5	全平均
平均	23	23	23	23
偏差平方和	0	0	0	0
				級間変動
偏差平方和	16	64	64	
	4	64	9	
	36		4	
			4	
			49	合計
計	56	128	130	314
				級内変動

2変数の関係

- 2変数の関係3：x(量的)×y(量的)図

	A	B	C	D	E	F	G	H	I	J	
身長 x	176	170	163	173	170	171	165	170	176	156	量的
体重 y	61	73	54	65	67	62	51	57	77	43	量的

散布図



2変数の関係

- 2変数の関係3：x(量的)×y(量的)式

	A	B	C	D	E	F	G	H	I	J	平均
身長 x	176	170	163	173	170	171	165	170	176	156	169
体重 y	61	73	54	65	67	62	51	57	77	43	61

相関係数

- ピアソンの積率相関係数 *Pearson's product-moment correlation coefficient*

$$r_{xy}=\frac{\mathrm{cov}_{xy}}{S_x\cdot S_y}\approx\frac{46}{5.848\cdot 9.706}\approx 0.81 \qquad (-1\le r_{xy}\le 1)$$

$$\mathrm{cov}_{xy}=\frac{(176-169)(61-61)+\cdots+(156-169)(43-61)}{10}=46 \quad (x,y\text{の共分散})$$

$$S_x=\sqrt{\frac{(176-169)^2+\cdots+(156-169)^2}{10}}\approx 5.848 \quad (x\text{の標準偏差})$$

$$S_y=\sqrt{\frac{(61-61)^2+\cdots+(43-61)^2}{10}}\approx 9.706 \quad (y\text{の標準偏差})$$

2変数の関係

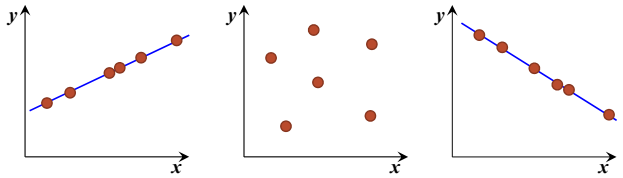
- 2変数の関係3：x(量的)×y(量的)式
  - ピアソンの積率相関係数 *Pearson's product-moment correlation coefficient*

$$r_{xy}=\frac{\mathrm{cov}_{xy}}{S_x\cdot S_y} \qquad (-1\le r_{xy}\le 1)$$

$r_{xy}=\frac{\mathrm{cov}_{xy}}{S_x\cdot S_y}=1$ 身長と体重は**正の相関**

$r_{xy}=\frac{\mathrm{cov}_{xy}}{S_x\cdot S_y}=0$ 身長と体重は**無相関**

$r_{xy}=\frac{\mathrm{cov}_{xy}}{S_x\cdot S_y}=-1$ 身長と体重は**負の相関**



2変数の相関 x(量的)×y(量的) についての補足

ピアソンの積率相関係数に関する補足と注意点

2変数の相関に関する補足

2変数の相関

- 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？

2変数の相関

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

- 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？

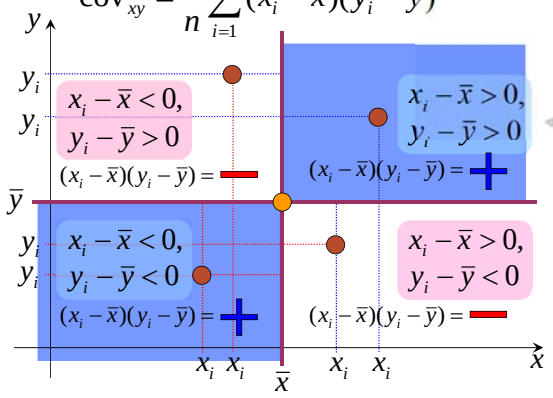
	月	火	水	木	金	
太郎	¥400	¥300	¥100	¥200	¥200	¥240 ← 平均
偏差	160	60	−140	−40	−40	
花子	¥100	¥200	¥300	¥400	¥200	¥240 ← 平均
偏差	−140	−40	60	160	−40	
積	−22,400	−2,400	−8,400	−6,400	1,600	−7,600 共分散

2変数の相関

- 共分散 *covariance*

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

共分散って，一体何を測ってるの？



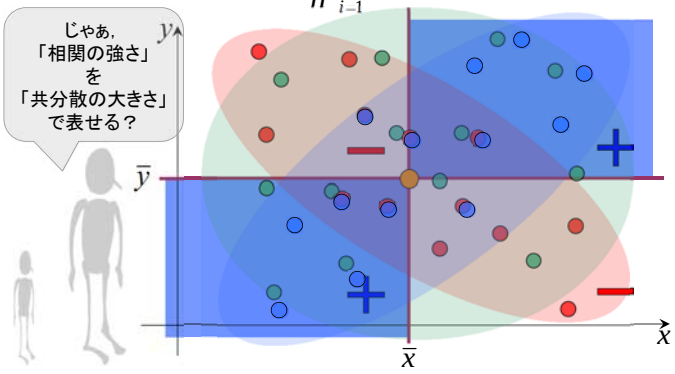
2変数の相関

- 共分散 *covariance*

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

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

じゃあ，「相関の強さ」を「共分散の大きさ」で表せる？



2変数の相関

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

- 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？

2変数の相関

- 共分散 *covariance*

測定単位が変わると，相関の度合いが変わってしまう！

	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

	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

共分散

2変数の相関

- 相関係数 *correlation*

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y} \qquad (-1 \le r_{xy} \le 1)$$

ピアソンの積率相関係数
Pearson's product-moment correlation coefficient

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

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．

2変数の相関

- 相関係数 *correlation*

測定単位が変わっても，相関の度合いは変わらない

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

	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	Ave.	St.Dev.
花子	¥100	¥200	¥300	¥400	¥200	¥240	101.98
偏差	−140	−40	60	160	−40	Cov.	Corr.
積	−2,240万	−240万	−840万	−640万	160万	−760万	−0.731

2変数の相関

A	$R_1, R_2, \ldots, R_n$	(R_i : Aがiを好きな順番)
B	$Q_1, Q_2, \ldots, Q_n$	(Q_j : Bがjを好きな順番)

- 順序尺度に対する相関係数
  - スピアマンの**順位相関係数** *Spearman 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** *Kendall tau rank correlation coefficient*

$$r_K = \frac{G - H}{G + H} \qquad (-1 \le r_K \le 1)$$

G：正順の数
H：逆順の数

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 $r_K = -1$

$$G + H = (n-1) + (n-2) + \cdots + 2 + 1 = \frac{(n-1)n}{2}$$ ←全対数

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（$G+H$個）について，正順と逆順の個数の差を比較したもの

$$r_K = \frac{G-H}{G+H} = \frac{24-4}{24+4} = \frac{20}{28} = \frac{5}{7} \approx 0.714$$

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

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)
- 村上雅人「なるほど統計学」海鳴社(2002)
- 金子治平ほか「よくわかる統計学Ⅰ」ミネルヴァ書房(2007)
- 大村平「改訂版 統計解析のはなし」日科技連(2006,1980)
- 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)

- 大村平「QC数学のはなし」日科技連(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「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)