2017/11/21 Tue.

問題解決技法入門

4. Data Analysis
1. Cross Tabulation

堀田 敬介

クロス集計とは

- クロス集計(表) cross tabulation
  - 2つ以上の**属性間**の**関係**を知りたい時に使う集計方法のひとつ．分割表ともよぶ

元データ

id	性別	年齢	嗜好1	嗜好2
1	女性	34	猫	紅茶
2	女性	21	犬	紅茶
3	男性	29	猫	紅茶
4	男性	69	猫	珈琲
5	女性	38	猫	紅茶
6	男性	64	猫	紅茶
7	男性	38	犬	珈琲
8	女性	37	猫	珈琲
9	男性	16	犬	珈琲
10	女性	25	犬	珈琲
11	女性	21	犬	紅茶
12	女性	17	猫	紅茶
13	男性	20	猫	珈琲
14	男性	16	犬	珈琲
15	女性	18	犬	紅茶

⋮

加工データ

ここを加工した

id	性別	年代	嗜好1	嗜好2
1	女性	30	猫	紅茶
2	女性	20	犬	紅茶
3	男性	20	猫	紅茶
4	男性	60	猫	珈琲
5	女性	30	猫	紅茶
6	男性	60	猫	紅茶
7	男性	30	犬	珈琲
8	女性	30	猫	珈琲
9	男性	10	犬	珈琲
10	女性	20	犬	珈琲
11	女性	20	犬	紅茶
12	女性	10	猫	紅茶
13	男性	20	猫	珈琲
14	男性	10	犬	珈琲
15	女性	10	犬	紅茶

⋮

クロス集計（例1）

「年代」と「嗜好1」の人数をクロス集計

	列ラベル		
行ラベル	犬	猫	総計
10	13	7	20
20	16	16	32
30	16	23	39
40	16	25	41
50	13	16	29
60	19	17	36
70	3		3
総計	96	104	200

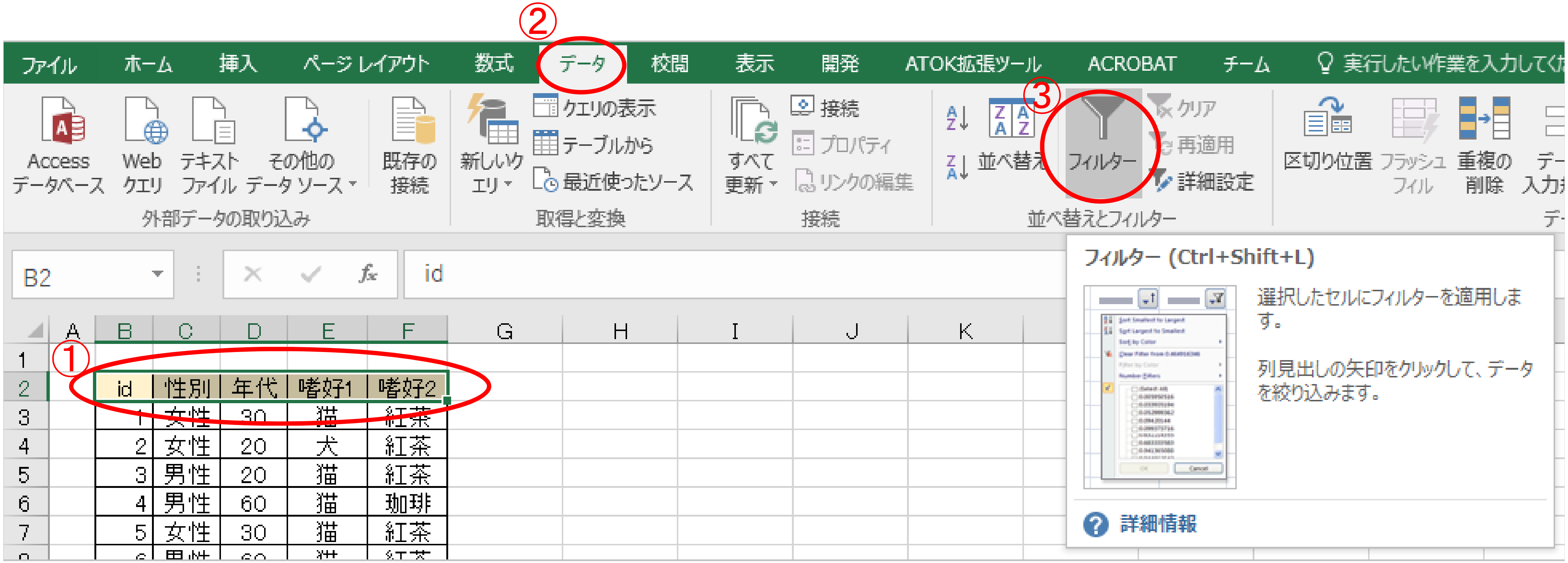
クロス集計（例2）

「性別」と「嗜好2」の人数をクロス集計

	列ラベル		
行ラベル	紅茶	珈琲	総計
女性	53	41	94
男性	57	49	106
総計	110	90	200

クロス集計の前に：フィルタを使おう

- 集計したいデータ項目を選択①し [データ②]-[フィルタ③]



クロス集計の前に：フィルタを使おう

- フィルタをかけ，欲しいデータだけを抽出
  - 例：「犬」好きで「紅茶」が好きな「女性」を抽出

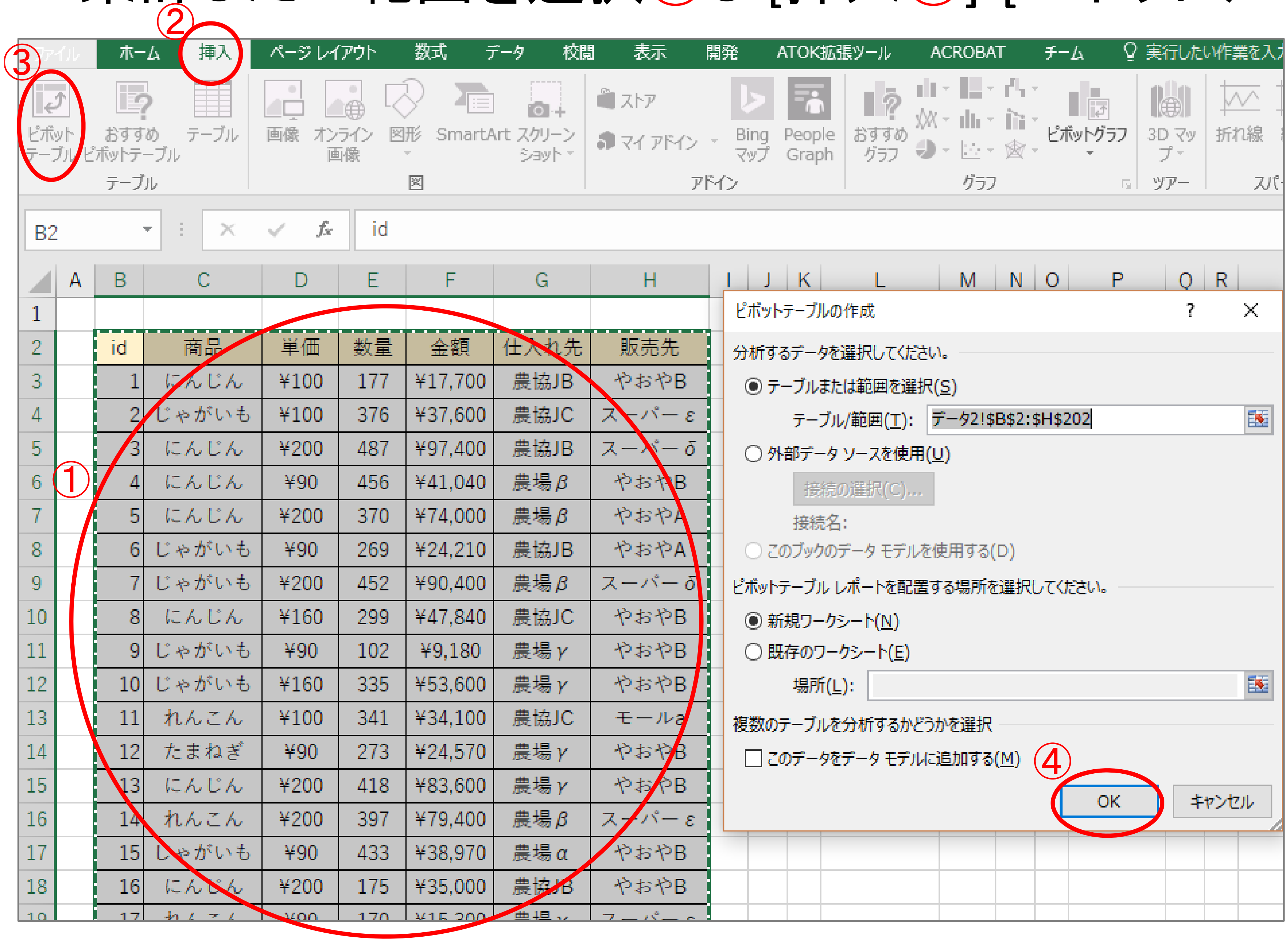
フィルタで選択

	A	B	C	D	E	F
1						
2		i	性	年	嗜好	嗜好
4		2	女性	20	犬	紅茶
13		11	女性	20	犬	紅茶
17		15	女性	10	犬	紅茶
28		26	女性	40	犬	紅茶
29		27	女性	30	犬	紅茶
31		29	女性	60	犬	紅茶
32		30	女性	60	犬	紅茶
38		36	女性	50	犬	紅茶
41		39	女性	20	犬	紅茶
53		51	女性	40	犬	紅茶
54		52	女性	50	犬	紅茶
55		53	女性	40	犬	紅茶
61		59	女性	60	犬	紅茶
66		64	女性	60	犬	紅茶
74		72	女性	60	犬	紅茶
82		80	女性	20	犬	紅茶
83		81	女性	10	犬	紅茶
92		90	女性	60	犬	紅茶
106		104	女性	30	犬	紅茶
112		110	女性	40	犬	紅茶
113		111	女性	10	犬	紅茶
117		115	女性	50	犬	紅茶
118		116	女性	10	犬	紅茶
119		117	女性	50	犬	紅茶
162		160	女性	20	犬	紅茶
171		169	女性	20	犬	紅茶
186		184	女性	30	犬	紅茶
197		195	女性	10	犬	紅茶
203						

データが選択（抽出）されたものだけだとわかるように，行番号が「青色」になっている

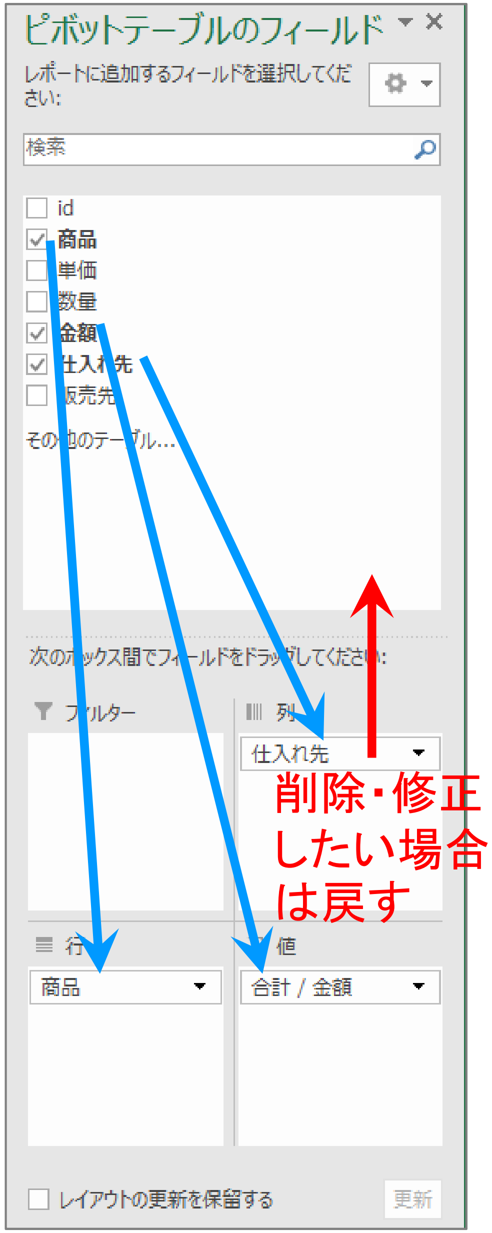
Excelでクロス集計

- 集計したい範囲を選択①し [挿入②]-[ピボットテーブル③] [OK④]



Excelでクロス集計

- [ピボットテーブルのフィールド]



- ➢ 上半分にデータの「項目（属性，フィールド）」名が並んでいる
- ➢ 下半分の「行」「列」「値」の（最低）3つを指定する
- ➢ 「行」「列」にクロスさせたい項目を，「値」に集計したい項目を，それぞれ該当の場所にドラッグ&ドロップ　→クロス集計表がExcelシート内に完成
- ➢ 修正・編集も同様（ドラッグ&ドロップ）

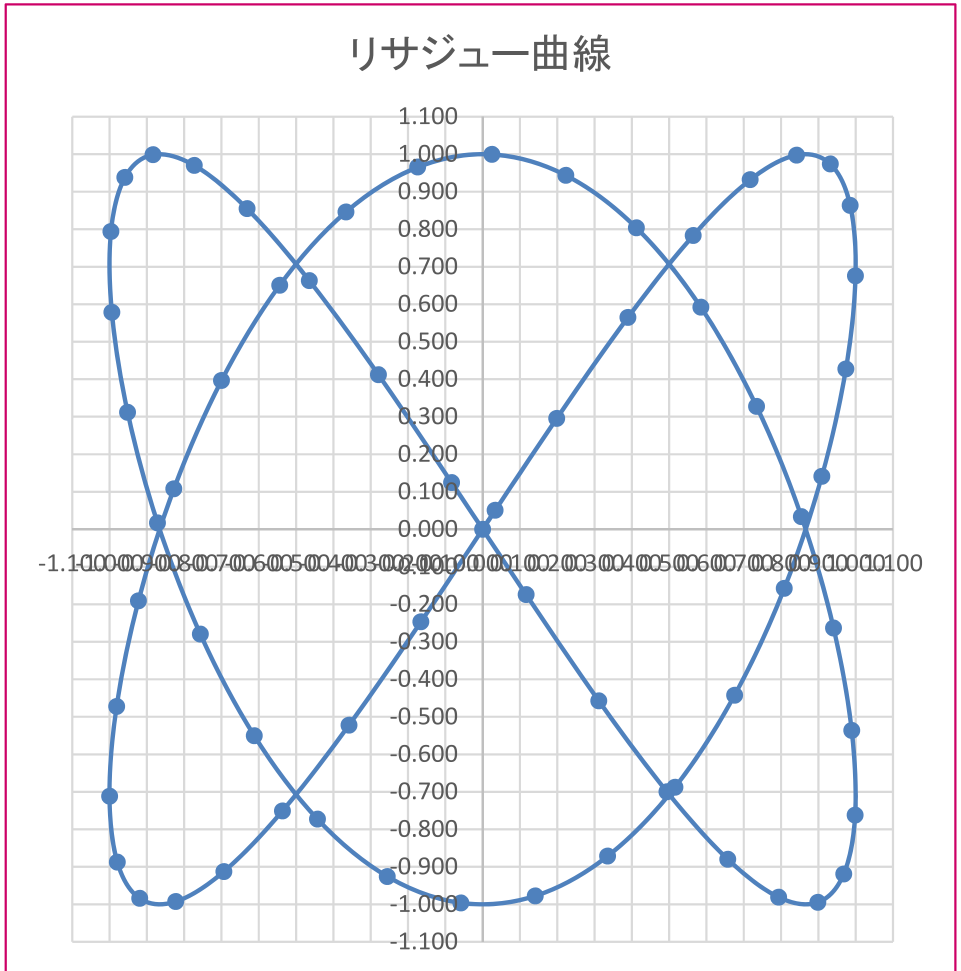
例）左の設定でできたクロス集計表
「行」=「商品」，「列」=「仕入れ先」，「値」=「金額」合計

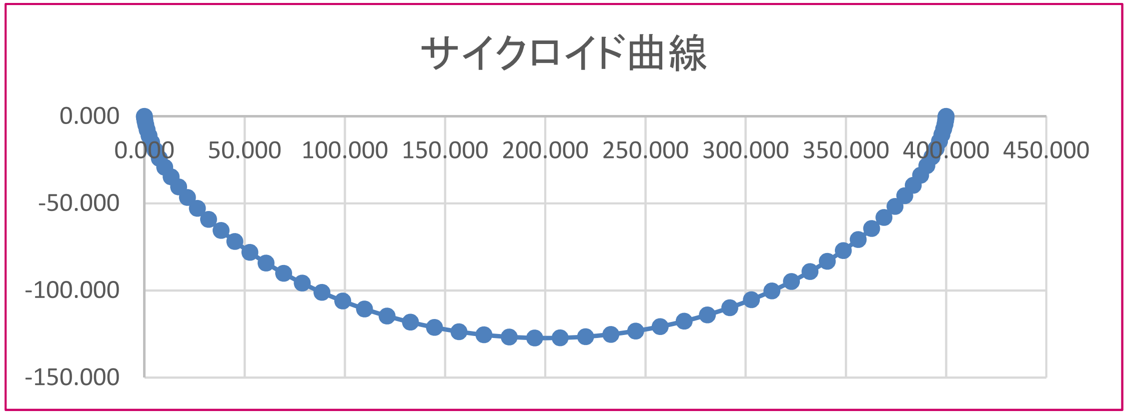
合計 / 金額	列ラベル					
行ラベル	農協JB	農協JC	農場α	農場β	農場γ	総計
じゃがいも	¥227,990	¥209,870	¥416,070	¥181,680	¥301,090	¥1,336,700
たまねぎ	¥632,040	¥209,560	¥223,420	¥208,200	¥316,740	¥1,589,960
にんじん	¥455,810	¥291,930	¥492,780	¥443,520	¥208,200	¥1,892,240
はくさい	¥359,360	¥435,900	¥61,860	¥398,340	¥172,360	¥1,427,820
れんこん	¥425,520	¥88,400	¥285,020	¥437,740	¥617,690	¥1,854,370
総計	¥2,100,720	¥1,235,660	¥1,479,150	¥1,669,480	¥1,616,080	¥8,101,090

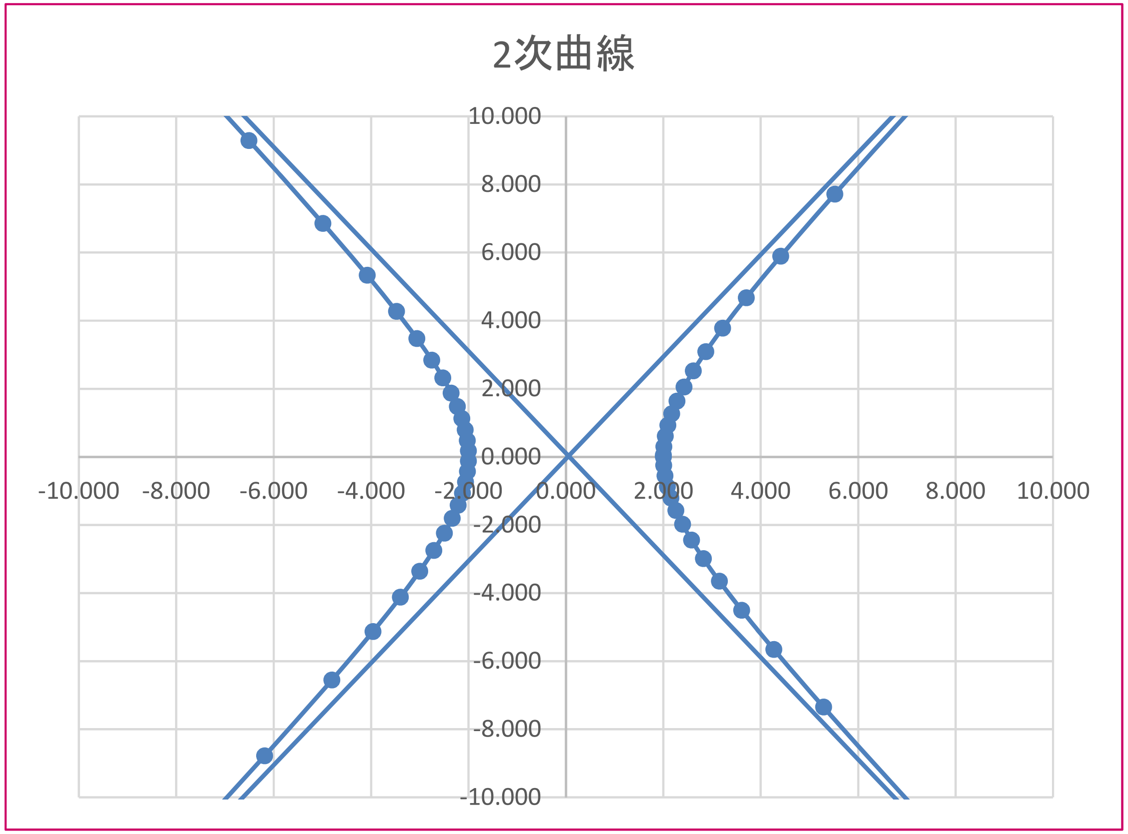
Coffee Break!

えくせるであそぶ

- 媒介変数表記とExcel散布図(平滑線)による関数の描画
  - リサジュー曲線
  - サイクロイド曲線
  - 2次曲線(円・楕円・双曲線)







【参考】

2変数間の分析法

尺度によって分析法が変わることに注意

- 2変数*x*,*y*の相関関係を調べる方法（図表と式）
  - 例1

例1	A	B	C	D	E	F	G	H	I	J	尺度
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

例2	A	B	C	D	E	F	G	H	I	J	尺度
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

例3	A	B	C	D	E	F	G	H	I	J	尺度
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
女	2	0	2	4
計	3	2	5	10

周辺度数

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

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

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

連関係数

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

- クラメルの連関係数 ***Cramer's coefficient of association***

$$\boxed{V=\sqrt{\frac{\chi^2}{n\cdot m}}}$$

$$(0\le V\le 1)$$

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

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

$$\begin{cases} \chi^2 = \dfrac{(0-4)^2}{4} + \dfrac{(3-2)^2}{2} + \dfrac{(9-6)^2}{6} + \dfrac{(6-2)^2}{2} + \dfrac{(0-1)^2}{1} + \dfrac{(0-3)^2}{3} \\ = 18 \\ n = 18 \\ m = \min\{2-1, 3-1\} = 1 \end{cases}$$

$$\rightarrow V = \sqrt{\frac{18}{18 \cdot 1}} = 1$$

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

$$\begin{cases} \chi^2 = \dfrac{(3-4)^2}{4} + \dfrac{(1-2)^2}{2} + \dfrac{(8-6)^2}{6} + \dfrac{(3-2)^2}{2} + \dfrac{(2-1)^2}{1} + \dfrac{(1-3)^2}{3} \\ = 17/4 \\ n = 18 \\ m = \min\{2-1, 3-1\} = 1 \end{cases}$$

$$\rightarrow V = \sqrt{\frac{17/4}{18 \cdot 1}} \approx 0.49$$

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

$$\begin{cases} \chi^2 = \dfrac{(4-4)^2}{4} + \dfrac{(2-2)^2}{2} + \dfrac{(6-6)^2}{6} + \dfrac{(2-2)^2}{2} + \dfrac{(1-1)^2}{1} + \dfrac{(3-3)^2}{3} \\ = 0 \\ n = 18 \\ m = \min\{2-1, 3-1\} = 1 \end{cases}$$

$$\rightarrow V = \sqrt{\frac{0}{18 \cdot 1}} = 0$$

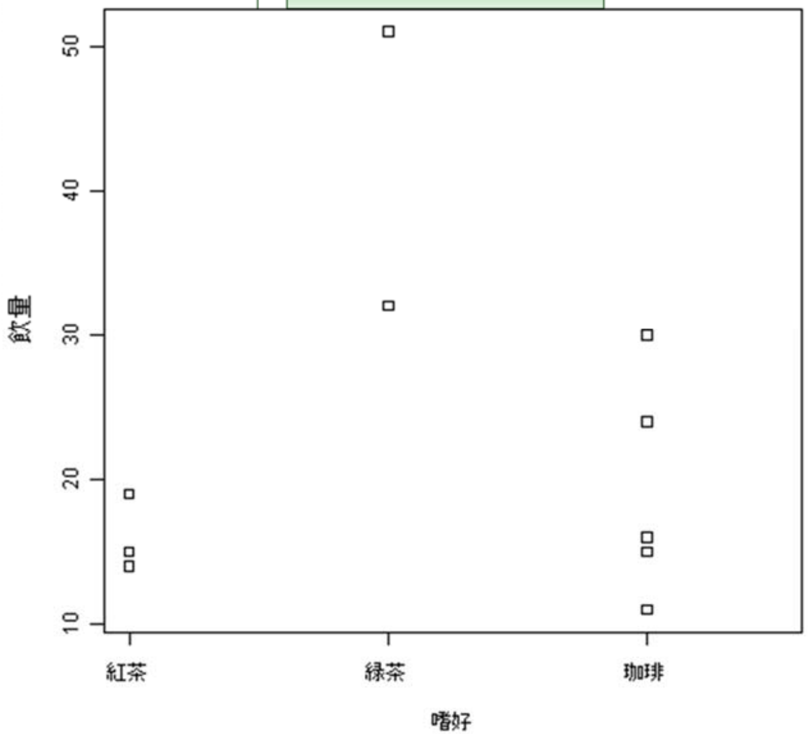
嗜好と性別は**無相関**

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

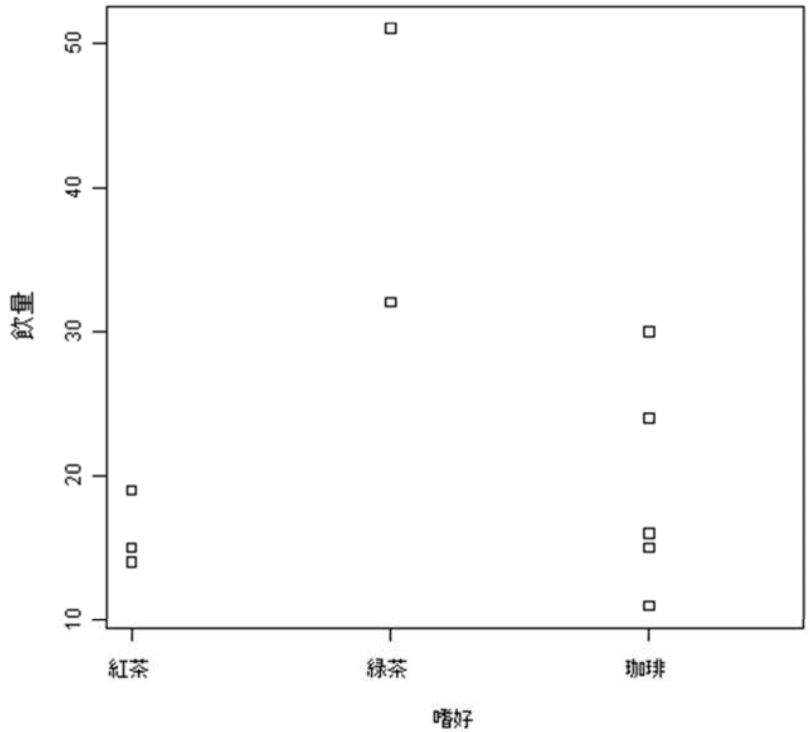
- 2変数の関係2 : x(量的) × y (質的)式

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

量的

質的

相関比



- 相関比 ***correlation ratio***

$$\eta^2 = \frac{S_T}{S_B + S_T}$$

$$\left(0 \le \eta^2 \le 1\right)$$

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 = **840** = 49 × 3 + 324 × 2 + 9 × 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 = **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*

$$\eta^2 = \frac{840}{0+840} = 1$$

$$\eta^2 = \frac{840}{376+840} \approx 0.691$$

$$\eta^2 = \frac{0}{314+0} = 0$$

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**

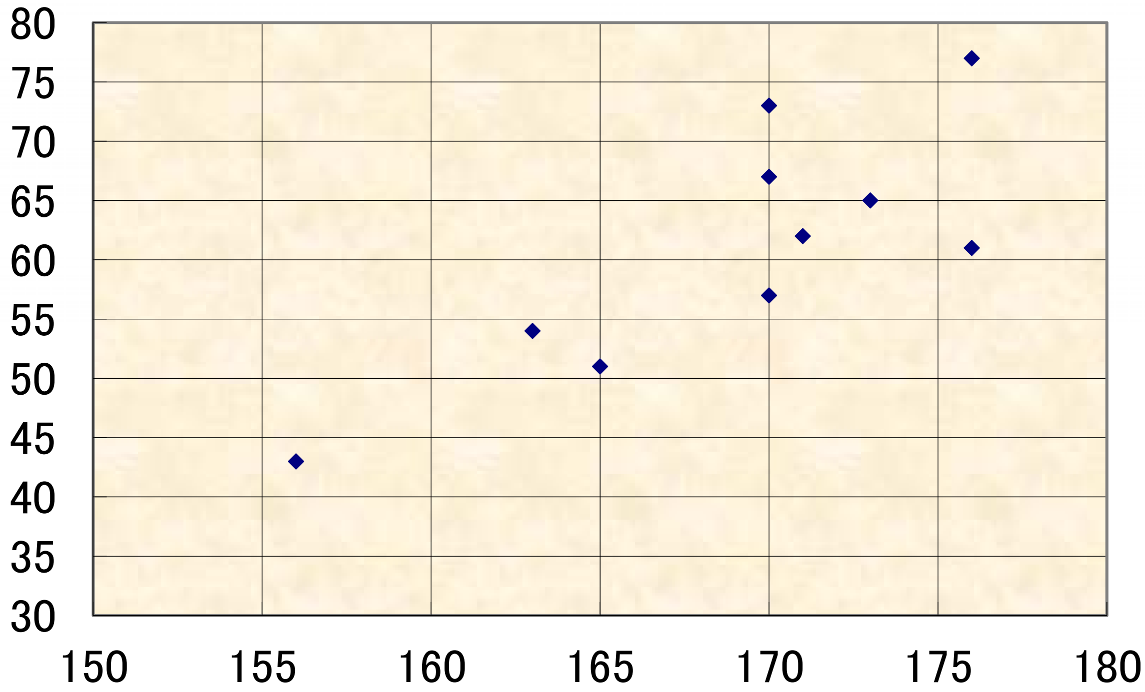
	紅茶	緑茶	珈琲	
	19	15	15	
	21	31	20	
	29		25	
			25	
			30	
個数	3	2	5	全平均
平均	23	23	23	23
偏差平方和	0	0	0	0
				級間変動
偏差平方和	16	64	64	
	4	64	9	
	36		4	
			4	
			49	合計
計	56	128	130	314
				級内変動

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***

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y} \approx \frac{46}{5.848 \cdot 9.706} \approx 0.81$$

$$\left\{
\begin{array}{l}
\mathrm{cov}_{xy} = \dfrac{(176-169)(61-61)+\cdots+(156-169)(43-61)}{10} = 46 \quad \text{(}x,y\text{の共分散)} \\
S_x = \sqrt{\dfrac{(176-169)^2+\cdots+(156-169)^2}{10}} \approx 5.848 \quad \text{(}x\text{の標準偏差)} \\
S_y = \sqrt{\dfrac{(61-61)^2+\cdots+(43-61)^2}{10}} \approx 9.706 \quad \text{(}y\text{の標準偏差)}
\end{array}
\right.$$

$$\left(-1 \le r_{xy} \le 1\right)$$

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***

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

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y} = 1$$

身長と体重は<u>正の相関</u>

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y} = 0$$

身長と体重は<u>無相関</u>

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

身長と体重は<u>負の相関</u>