

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

行ラベル	列ラベル	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

行ラベル	列ラベル	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

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

	農協B	農協C	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
 - リサージュ曲線
 - 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

	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

	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

	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) \times 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 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	

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) \times y (質的) 式

	紅茶	緑茶	珈琲	計	連関係数	紅茶	緑茶	珈琲	計
男	1	2	3	6	クロス集計から理論度数を求める	1.8	1.2	3.0	6
女	2	0	2	4		1.2	0.8	2.0	4
計	3	2	5	10		3	2	5	10

連関係数 $1.8 = \frac{3 \cdot 6}{10}$, $2.0 = \frac{5 \cdot 4}{10}$

□ クラメル の 連関係数 *Cramer's coefficient of association*

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

$(0 \leq V \leq 1)$

$$\chi^2 = \frac{(1-1.8)^2}{1.8} + \frac{(2-1.2)^2}{1.2} + \dots + \frac{(0-0.8)^2}{0.8} + \frac{(2-2.0)^2}{2.0}$$

$n = 10$
 $m = \min\{2-1, 3-1\}$
 (行数-1)と(列数-1)の小さい方

ピアソンの統計量

2変数の関係

□ 2変数の関係1: x (質的) \times y (質的) 式

□ クラメル の 連関係数 *Cramer's coefficient of association*

	紅	緑	珈	計		紅	緑	珈	計	
男	0	3	9	12	完全相関	3	1	8	12	無相関
女	6	0	0	6		3	2	1	6	
計	6	3	9	18		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		紅	緑	珈	計	
男	3	1	8	12	多少相関	4	2	6	12	無相関
女	3	2	1	6		2	1	3	6	
計	6	3	9	18		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		紅	緑	珈	計	
男	4	2	6	12	無相関	4	2	6	12	無相関
女	2	1	3	6		2	1	3	6	
計	6	3	9	18		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) \times 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 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	

点グラフ

2変数の関係

□ 2変数の関係2: x (量的) \times 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*

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

2変数の関係

□ 2変数の関係2: x (量的) \times y (質的)式

□ 相関比 *correlation ratio*

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

	紅茶	緑茶	珈琲	全平均
個数	3	2	5	23
平均	16	41	20	23
偏差平方	49	324	9	840

$S_T = 840 = 49 \times 3 + 324 \times 2 + 9 \times 5$
 級間変動 = 級平均と全平均との偏差平方の加重和

	紅茶	緑茶	珈琲
偏差平方	4	81	64
1	1	81	16
9			4
			16
			100
合計	14	162	200

$S_B = 376 = 14 + 162 + 200$
 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) \times y (質的)式

□ 相関比 *correlation ratio*

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

嗜好と飲量は**完全相関** 嗜好と飲量は**多少相関** 嗜好と飲量は**無相関**

	紅茶	緑茶	珈琲	全平均
個数	3	2	5	23
平均	16	41	20	23
偏差平方和	49	324	9	840

偏差平方和: 0, 0, 0 合計: 0 級内変動: 0

	紅茶	緑茶	珈琲	全平均
個数	3	2	5	23
平均	16	41	20	23
偏差平方和	49	324	9	840

偏差平方和: 4, 81, 64 合計: 376 級内変動: 376

	紅茶	緑茶	珈琲	全平均
個数	3	2	5	23
平均	23	23	23	23
偏差平方和	0	0	0	0

偏差平方和: 16, 64, 64 合計: 314 級内変動: 314

2変数の関係

□ 2変数の関係3: x (量的) \times 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) \times 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*

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

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

$$\approx \frac{46}{5.848 \cdot 9.706} \approx 0.81$$

$(-1 \leq r_{xy} \leq 1)$

2変数の関係

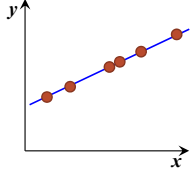
□ 2変数の関係3: x (量的) \times y (量的)式

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

$$r_{xy} = \frac{\text{cov}_{xy}}{S_x \cdot S_y} \quad (-1 \leq r_{xy} \leq 1)$$

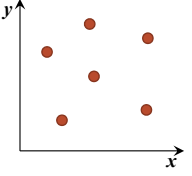
$r_{xy} = \frac{\text{cov}_{xy}}{S_x \cdot S_y} = 1$

身長と体重は**正の相関**



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

身長と体重は**無相関**



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

身長と体重は**負の相関**

