データサイエンス　B
第13回　質的データの解析

髙田　美樹

目次

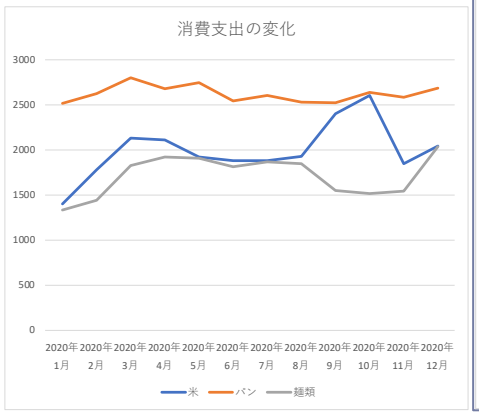
1. 質的データの度数分布とパレート図
2. クロス集計表
3. 行パーセントと列パーセント
4. オッズ比
5. 多重クロス表
6. シンプソンのパラドックス
7. 擬似的な関係
8. 交互作用

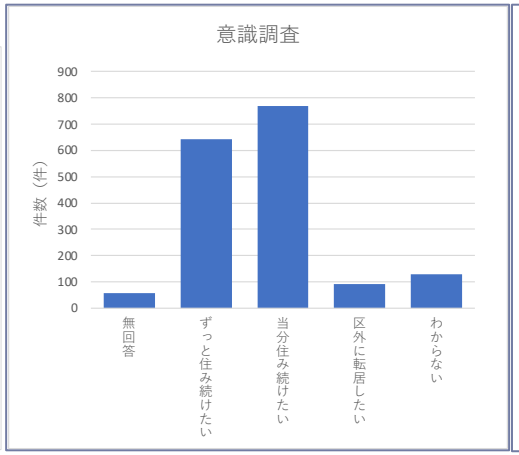
データの分類

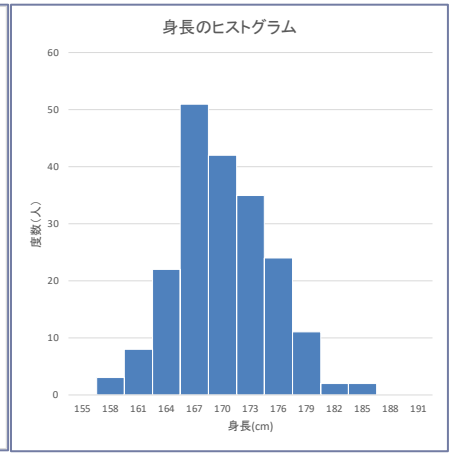
- データ
  - 定型データ
    - 時系列データ
    - クロスセクションデータ
      - 質的データ
      - 量的データ
    - パネルデータ
  - 非定型データ

画像・音・文書

時系列データ＋
クロスセクションデータ







質的データの度数分布

	A	B	C	D	E	F
1		個人情報漏えいの原因				
2						
3	日付	項目	分類		項目	件数
4	****	宛名間違い	誤送付		宛名間違い	314
5	****	紛失	紛失		封入ミス	323
6	****	メール誤送信	誤送付		配達ミス	137
7	****	配達ミス	誤送付		メール誤送信	764
8	****	メール誤送信	誤送付		FAX誤送信	110
9	****	その他	その他		紛失	394
10	****	メール誤送信	誤送付		作業ミス	232
11	****	封入ミス	誤送付		システムの誤り	102
12	****	作業ミス	その他漏えい		不正アクセス	54
13	****	宛名間違い	誤送付		口頭	37
14	****	紛失	紛失		ウイルス感染	29
15	****	FAX誤送信	誤送付		盗難	8
16	****	紛失	紛失		その他	140
17	****	システムの誤り	その他漏えい		合計	2644
18	****	宛名間違い	誤送付			

ファイル名：DSB13実習用データ.xlsx

=COUNTIF($B$4:$B$2647,E4)

2020年度「個人情報の取扱いにおける事故報告集計結果」より作成

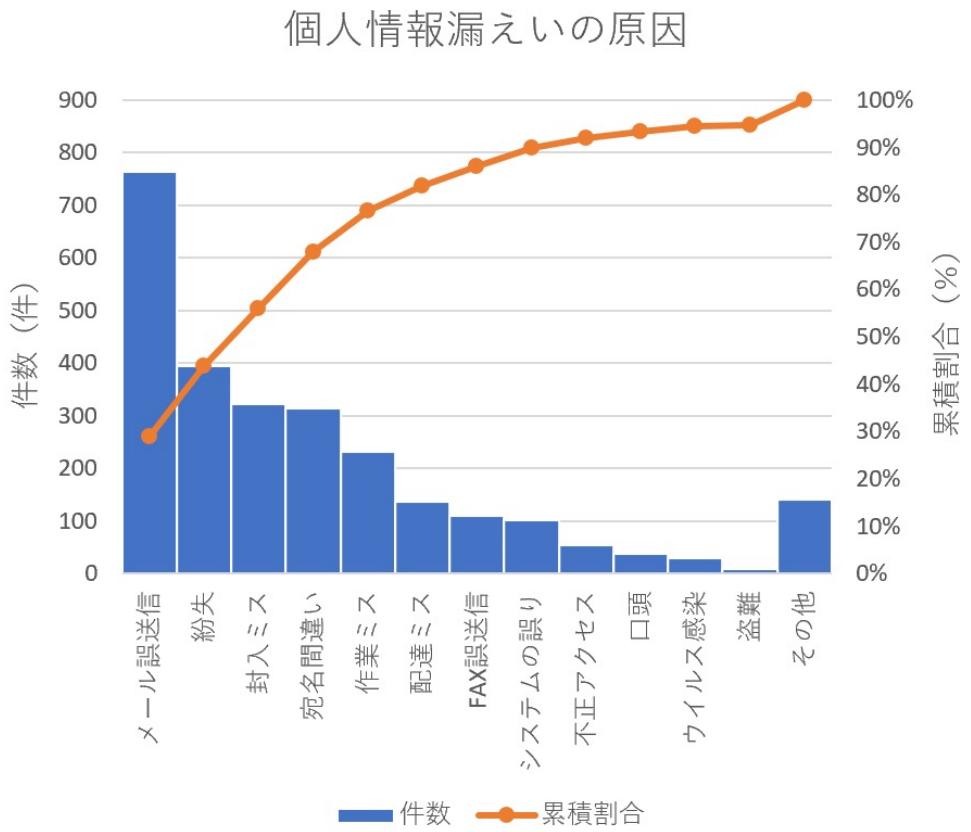
https://privacymark.jp/system/reference/pdf/2020JikoHoukoku_211005.pdf

ABC分析

件数の多い順に対策を立てる

項目	件数	累積	累積割合	ランク
メール誤送信	764	764	28.9%	A
紛失	394	1158	43.8%	A
封入ミス	323	1481	56.0%	A
宛名間違い	314	1795	67.9%	B
作業ミス	232	2027	76.7%	B
配達ミス	137	2164	81.8%	C
FAX誤送信	110	2274	86.0%	C
システムの誤り	102	2376	89.9%	C
不正アクセス	54	2430	91.9%	C
口頭	37	2467	93.3%	C
ウイルス感染	29	2496	94.4%	C
盗難	8	2504	94.7%	C
その他	140	2644	100.0%	C
合計	2644			

パレート図



2変数データの集計

	A	B	C	D	E
1	高齢者の口腔状態と将来の閉じこもり				
2	id	歯	咀嚼困難	むせ	閉じこもり
3	1	あり	あり	あり	あり
4	2	あり	あり	あり	あり
5	3	あり	あり	あり	あり
6	4	あり	あり	あり	あり
7	5	あり	あり	あり	あり
8	6	あり	あり	あり	あり
9	1	あり	あり	あり	なし
10	2	あり	あり	あり	なし
11	3	あり	あり	あり	なし
12	4	あり	あり	あり	なし
13	5	あり	あり	あり	なし
14	6	あり	あり	あり	なし
15	7	あり	あり	あり	なし
16	8	あり	あり	あり	なし
17	9	あり	あり	あり	なし
18	10	あり	あり	あり	なし
19	11	あり	あり	あり	なし
20	12	あり	あり	あり	なし
21	13	あり	あり	あり	なし
22	14	あり	あり	あり	なし

N = 26,579

高齢者の口腔状態と閉じこもりの関連調査（東北大学）から作成
http://www.tohoku.ac.jp/japanese/newimg/pressimg/tohokuuniv-press20211213_03web_oral.pdf

クロス集計表

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	10336	254	10590
20本未満	15234	755	15989
総計	25570	1009	26579

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	19828	671	20499
あり	5742	338	6080
総計	25570	1009	26579

むせ	非閉じこもり	閉じこもり	総計
なし	22141	844	22985
あり	3429	165	3594
総計	25570	1009	26579

行パーセント

行パーセント

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	97.6%	2.4%	100.0%
20本未満	95.3%	4.7%	100.0%
総計	96.2%	3.8%	100.0%

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	96.7%	3.3%	100.0%
あり	94.4%	5.6%	100.0%
総計	96.2%	3.8%	100.0%

むせ	非閉じこもり	閉じこもり	総計
なし	96.3%	3.7%	100.0%
あり	95.4%	4.6%	100.0%
総計	96.2%	3.8%	100.0%

クロス集計表

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	10336	254	10590
20本未満	15234	755	15989
総計	25570	1009	26579

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	19828	671	20499
あり	5742	338	6080
総計	25570	1009	26579

むせ	非閉じこもり	閉じこもり	総計
なし	22141	844	22985
あり	3429	165	3594
総計	25570	1009	26579

行パーセント

行パーセント

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	97.6%	2.4%	100.0%
20本未満	95.3%	4.7%	100.0%
総計	96.2%	3.8%	100.0%

咀嚼困難	非閉じこもり	閉じこもり	総計
なし			
あり			
総計			

20本以上の方で閉じこもってしまったのは2.4%
20本未満の方は、4.7%

むせ	非閉じこもり	閉じこもり	総計
なし	96.3%	3.7%	100.0%
あり	95.4%	4.6%	100.0%
総計	96.2%	3.8%	100.0%

クロス集計表

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	10336	254	10590
20本未満	15234	755	15989
総計	25570	1009	26579

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	19828	671	20499
あり	5742	338	6080
総計	25570	1009	26579

むせ	非閉じこもり	閉じこもり	総計
なし	22141	844	22985
あり	3429	165	3594
総計	25570	1009	26579

列パーセント

列パーセント

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	40.4%	25.2%	39.8%
20本未満	59.6%	74.8%	60.2%
総計	100.0%	100.0%	100.0%

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	77.5%	66.5%	77.1%
あり	22.5%	33.5%	22.9%
総計	100.0%	100.0%	100.0%

むせ	非閉じこもり	閉じこもり	総計
なし	86.6%	83.6%	86.5%
あり	13.4%	16.4%	13.5%
総計	100.0%	100.0%	100.0%

クロス集計表

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	10336	254	10590
20本未満	15234	755	15989
総計	25570	1009	26579

咀嚼困難	非閉じこもり	閉じこもり	総計
なし	19828	671	20499
あり	5742	338	6080
総計	25570	1009	26579

むせ	非閉じこもり	閉じこもり	総計
なし	22141	844	22985
あり	3429	165	3594
総計	25570	1009	26579

列パーセント

列パーセント

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	40.4%	25.2%	39.8%
20本未満	59.6%	74.8%	60.2%
総計	100.0%	100.0%	100.0%

咀嚼困難			
なし			%
あり			%
総計			%

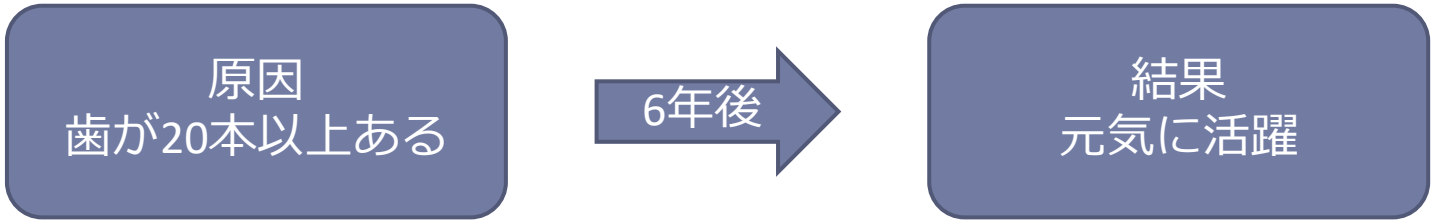
むせ	非閉じ		
なし			
あり			
総計	100.0%	100.0%	100.0%

閉じこもってしまった方のうち、3/4が20本未満だった

~~歯が20本以上あっても、1/4が閉じこもってしまうのか~~

因果関係

歯が20本以上あったから、元気に活躍できている？



結果（応答変数）

原因
（説明変数）

歯の本数	非閉じこもり	閉じこもり	総計
20本以上	10336	254	10590
20本未満	15234	755	15989
総計	25570	1009	26579

オッズ比

$$\text{オッズ} = \frac{\text{結果が生じる}}{\text{結果が生じない}}$$

歯の本数	非閉じこもり		閉じこもり		活躍できる比率
20本以上	a	10336	b	254	40.693
20本未満	c	15234	d	755	20.177
			オッズ比		2.017

$$\text{オッズ比} = \frac{a \times d}{b \times c}$$

オッズ比

歯の本数	非閉じこもり	閉じこもり	オッズ
20本以上	10336	254	40.693
20本未満	15234	755	20.177
		オッズ比	2.017

咀嚼困難	非閉じこもり	閉じこもり	オッズ
なし	19828	671	29.550
あり	5742	338	16.988
		オッズ比	1.739

むせ	非閉じこもり	閉じこもり	オッズ
なし	22141	844	26.233
あり	3429	165	20.782
		オッズ比	1.262

３つの変数

年収	保有	保有していない	合計
年収400万以上	639	415	1054
年収400万未満	1192	2561	3753
合計	1831	2976	4807

性別	保有	保有していない	合計
男性	1052	769	1821
女性	779	2207	2986
合計	1831	2976	4807

年収が多いことによる影響は

$$\text{オッズ比} = \frac{639 \times 2561}{1192 \times 415} = 3.38$$

男性であることの影響は

$$\text{オッズ比} = \frac{1052 \times 2207}{769 \times 779} = 3.88$$

全国家計構造調査（旧全国消費実態調査） 平成16年全国消費実態調査 全国 家計収支編より作成

https://www.e-stat.go.jp/dbview?sid=0000111360

多重クロス表

層化

性別	年収	保有	保有していない	合計
男性	年収400万以上	514	252	766
	年収400万未満	538	517	1055
	合計	1052	769	1821
女性	年収400万以上	125	163	288
	年収400万未満	654	2044	2698
	合計	779	2207	2986

年収が多いことにより影響は

男性：オッズ比＝$\frac{514 \times 517}{538 \times 252}$ = 1.96

女性：オッズ比＝$\frac{125 \times 2044}{654 \times 163}$ = 2.40

シンプソンのパラドックス（イスラエルのワクチン有効性）

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax	Fully Vax	vs. severe disease
All ages			214	301	**Vax don't work!**

2021年8月15日現在

この時点でのイスラエルの状況は、
1. 12歳以上の全居住者のほぼ80%が接種済み
2. 年齢によるワクチン接種の差が大きい
3. ほぼすべての高齢者がワクチン接種済み（居住者の90%以上が50歳以上）
4. ワクチン未接種の大多数は若い人たち（ワクチン未接種の85%未満50歳未満）
5. 高齢者は若い人より呼吸器ウイルスで重症化する可能性が桁違いに高い

https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

シンプソンのパラドックス（イスラエルのワクチン有効性）

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**

$$\frac{214}{1{,}302{,}912} \times 1000{,}000 = 16.4$$

$$\frac{301}{5{,}634{,}634} \times 1000{,}000 = 5.3$$

$$\frac{16.4}{5.3} = 3.07$$

ワクチン効果：

$$\frac{1 - \text{未接種者率}}{\text{接種者率}}$$

$$= \frac{1 - 5.3}{16.4} = 67.5\%$$

スライド17訂正事項

シンプソンのパラドックス（イスラエルのワクチン有効性）

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**

$$\frac{214}{1{,}302{,}912} \times 1000{,}000 = 16.4$$

$$\frac{301}{5{,}634{,}634} \times 1000{,}000 = 5.3$$

$$\frac{16.4}{5.3} = 3.07$$

ワクチン効果：

$$= \frac{1 - 5.3}{16.4} = 67.5\%$$

ワクチン効果：

$$1 - \frac{\textit{未接種者率}}{\textit{接種者率}} = 1 - \frac{5.3}{16.4} = 67.5\%$$

シンプソンのパラドックス（イスラエルのワクチン有効性）

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

未接種者における若年の割合

$$\frac{1{,}116{,}834}{1{,}302{,}912} = 85.7\%$$

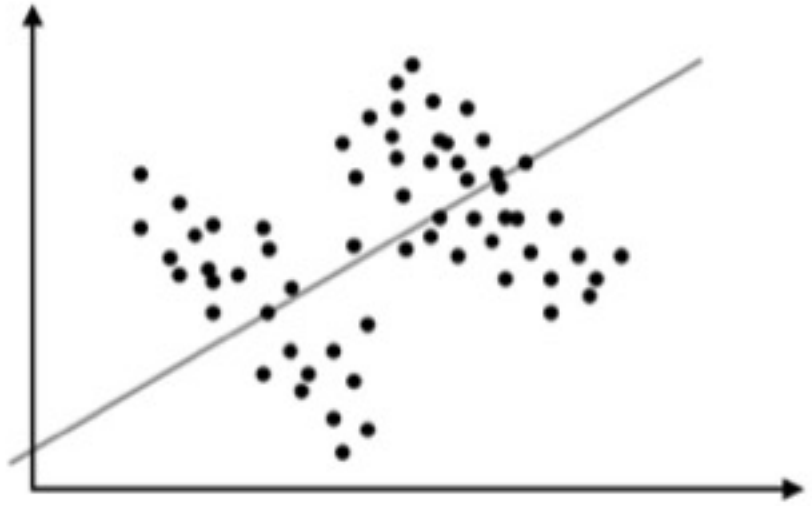
年齢によるリスク（未接種）

$$\frac{91.9}{3.9} = 23.6\text{倍}$$

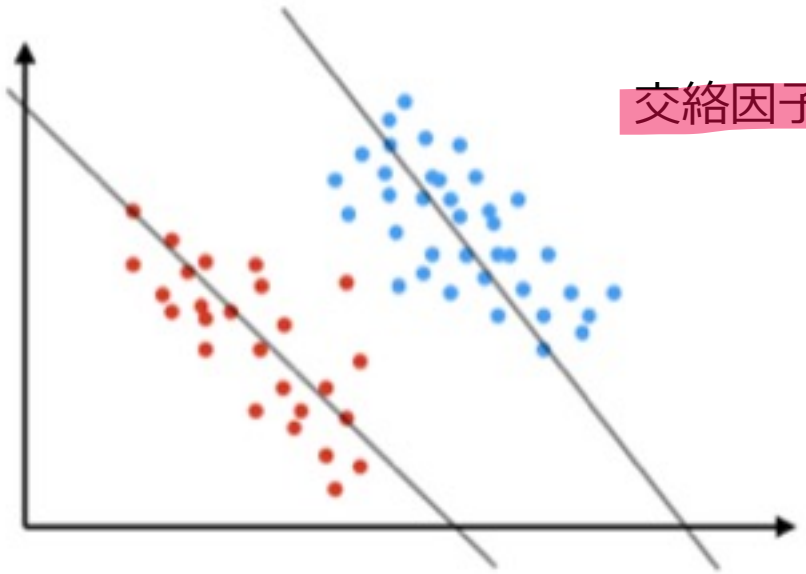
年齢によるリスク（接種）

$$\frac{13.6}{0.3} = 43.2\text{倍}$$

シンプソンのパラドックス



全体では正の相関



層化すると、それぞれ負の相関

シンプソンのパラドックス（イスラエルのワクチン有効性）

Age	Population (%)		Severe cases/100k		Severe Case Risk	Efficacy
	% Not Vax	% Fully Vax	Not Vax	Fully Vax	Ratio w/ 30-39 UnVax	vs. severe disease
12-15	62.1%	29.9%	0.30	0.00	1/20x	100%
16-19	21.9%	73.5%	1.60	0.00	1/4x	100%
20-29	20.5%	76.2%	1.50	0.00	1/4x	100%
30-39	16.2%	80.9%	6.20	0.20	1	96.8%
40-49	13.2%	84.4%	16.50	1.00	2.7x	93.9%
50-59	10.0%	88.0%	40.20	2.90	6.5x	92.8%
60-69	8.8%	89.8%	76.60	8.70	12.4x	88.7%
70-79	4.2%	94.6%	190.10	19.80	30.7x	89.6%
80-89	5.6%	92.6%	252.30	47.90	40.7x	81.1%
90+	6.1%	90.5%	510.9	38.60	82.4x	92.4%

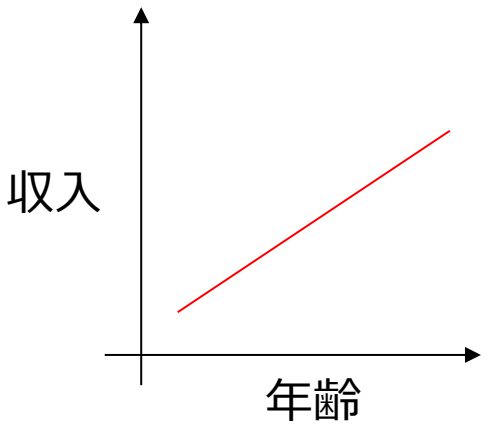
擬似的な関係

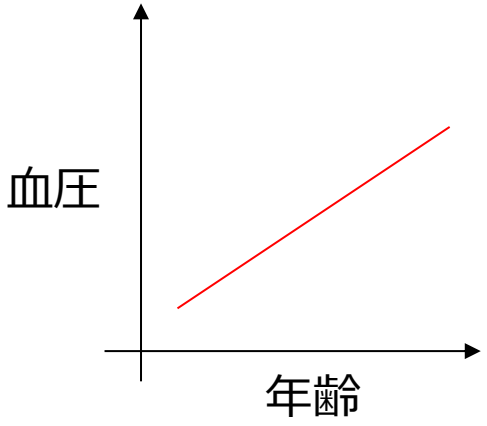
血圧	年収700万円		合計
	以上	未満	
高い	385	2115	2500
正常	240	2260	2500
合計	625	4375	5000

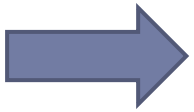
行パーセント

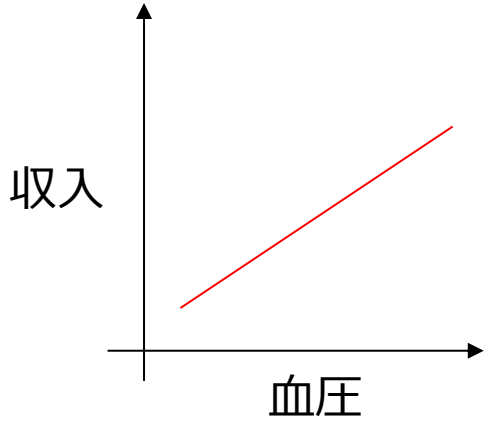
血圧	年収700万円		合計
	以上	未満	
高い	0.154	0.846	1.000
正常	0.096	0.904	1.000
合計	0.125	0.875	1.000

高血圧の人は年収が高い？









擬似的な関係　年齢で層化

年齢層	血圧	年収700万円		合計
		以上	未満	
20-29歳	高い	15	285	300
	正常	60	1140	1200
	合計	75	1425	1500
30-39歳	高い	70	630	700
	正常	80	720	800
	合計	150	1350	1500
40-49歳	高い	300	1200	1500
	正常	100	400	500
	合計	400	1600	2000

合計	高い	385	2115
	正常	240	2260
	合計	625	4375

行パーセント

年齢層	血圧	年収700万円		合計
		以上	未満	
20-29歳	高い	0.05	0.95	1
	正常	0.05	0.95	1
	合計	0.05	0.95	1
30-39歳	高い	0.1	0.9	1
	正常	0.1	0.9	1
	合計	0.1	0.9	1
40-49歳	高い	0.2	0.8	1
	正常	0.2	0.8	1
	合計	0.2	0.8	1

交互作用

性別	映画	面白い	面白くない	合計
男性	映画A	342	132	474
	映画B	239	253	492
	合計	581	385	966
女性	映画A	215	223	438
	映画B	382	167	549
	合計	597	390	987

行パーセント

性別	映画	面白い	面白くない	合計
男性	映画A	0.722	0.278	1.000
	映画B	0.486	0.514	1.000
	合計	0.601	0.399	1.000
女性	映画A	0.491	0.509	1.000
	映画B	0.696	0.304	1.000
	合計	0.605	0.395	1.000

満足度

	映画A	映画B
男性	72.2%	48.6%
女性	49.1%	69.6%

まとめ

1. 質的データの度数分布とパレート図
2. クロス集計表
3. 行パーセントと列パーセント
4. オッズ比
5. 多重クロス表
6. シンプソンのパラドックス
7. 擬似的な関係
8. 交互作用

参考文献

- 身近な統計　石橋克也　渡辺美智子　NHK出版
- 統計学Ｉオフィシャル　スタディ　ノート　日本統計協会