第9回 統計教育の方法論ワークショップ
主催：日本統計学会・統計数理研究所
日時：平成２５年３月１,２日
会場：学習院大学

欠測とサンプリングの相関分析・関連性分析への影響：
関西にタコ焼き器は多いのか？

狩野 裕・森川耕輔・磯崎郷平
（大阪大学 大学院基礎工学研究科）

本日の内容

- 教養統計学での講義例×２
  - 質的データの関連性分析
    - たこ焼き，焼いたことがありますか？
  - 量的データの相関分析
    - 女性は体型に気を使う？

関西にたこ焼き器は多いのか？
and
シンプソンのパラドックス

質的データの関連性分析

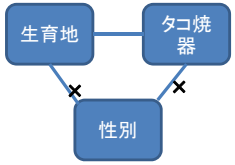
たこ焼器の有無と生育地，性別（O大学H学部2009年度）

性別	生育地	たこ焼器 あり	たこ焼器 なし	合計
男	関西	16	3	19
	非関西	6	11	17
女	関西	16	1	17
	非関西	7	10	17
合計		45	25	70

生育地	たこ焼器 あり	たこ焼器 なし	合計		
関西	32	4	36	オッズ比=	12.92
非関西	13	21	34	クラメールのV=	0.53
合計	45	25	70	カイ2乗値=	19.54

生育地	性別 男	性別 女	合計		
関西	19	17	36	オッズ比=	1.12
非関西	17	17	34	クラメールのV=	0.03
合計	36	34	70	カイ2乗値=	0.054

性別	たこ焼器 あり	たこ焼器 なし	合計		
男	22	14	36	オッズ比=	0.75
女	23	11	34	クラメールのV=	−0.07
合計	45	25	70	カイ2乗値=	0.33



$$\text{オッズ比}=\frac{a/b}{c/d}=\frac{ad}{bc},$$

$$\text{クラメールの}V=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

過去10年の履歴

2002年度

生育地	たこ焼器		計
	あり	なし	
関西	32	7	39
非関西	10	18	28
計	42	25	67

カイ2乗値= 14.96
V= 0.47
オッズ比 8.23

2003年度

生育地	たこ焼器		計
	あり	なし	
関西	32	4	36
非関西	11	13	24
計	43	17	60

カイ2乗値= 13.15
V= 0.47
オッズ比 9.45

2004年度

生育地	たこ焼器		計
	あり	なし	
関西	37	13	50
非関西	11	18	29
計	48	31	79

カイ2乗値= 10.02
V= 0.36
オッズ比 4.66

2006年度

生育地	たこ焼器		計
	あり	なし	
関西	34	5	39
非関西	14	22	36
計	48	27	75

カイ2乗値= 18.95
V= 0.50
オッズ比 10.69

2007年度

生育地	たこ焼器		計
	あり	なし	
関西	38	0	38
非関西	8	19	27
計	46	19	65

カイ2乗値= 37.79
V= 0.76
オッズ比

2008年度

生育地	たこ焼器		計
	あり	なし	
関西	32	6	38
非関西	9	12	21
計	41	18	59

カイ2乗値= 10.91
V= 0.43
オッズ比 7.11

2009年度

生育地	たこ焼器		計
	あり	なし	
関西	32	4	36
非関西	13	21	34
計	45	25	70

カイ2乗値= 19.54
V= 0.53
オッズ比 12.92

2010年度

生育地	たこ焼器		計
	あり	なし	
関西	25	1	26
非関西	10	15	25
計	35	16	51

カイ2乗値= 18.67
V= 0.60
オッズ比 37.50

2011年度

生育地	たこ焼器		計
	あり	なし	
関西	27	8	35
非関西	9	17	26
計	36	25	61

カイ2乗値= 11.15
V= 0.43
オッズ比 6.38

2012年度

生育地	たこ焼器		計
	あり	なし	
関西	31	7	38
非関西	12	13	25
計	43	20	63

カイ2乗値= 7.85
V= 0.35
オッズ比 4.80

色々な確率

- 周辺確率(marginal probability)
  - P(関西) = 36/70, P(非関西) = 34/70
    P(あり) = 45/70, P(なし) = 25/70
- 同時確率(joint probability)
  - P(関西∩あり)=32/70 … 合計4つ
- 条件付確率(conditional probability)

$$P(\text{あり}|\text{関西}) = \frac{32}{36} = \frac{P(\text{あり}\cap\text{関西})}{P(\text{関西})} = \frac{32/70}{36/70}$$

$$P(\text{関西}|\text{あり}) = \frac{32}{45} = \frac{P(\text{あり}\cap\text{関西})}{P(\text{関西})} = \frac{32/70}{45/70}$$

生育地	たこ焼器		合計
	あり	なし	
関西	32	4	36
非関西	13	21	34
合計	45	25	70

O大学E学部での調査

- O大学E学部から無作為に選ばれた60名についての調査
- 非関西の標本サイズが小さいので，正確な推測ができない
- そこで，その集団から関西=非関西=50となるように標本抽出する
  - 理由を考える

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	2	8	10
計	37	23	60

表A

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	10	40	50
計	45	55	100

表B

分割表データから安易に確率を計算できない！

- 表Bのデータから以下を考察する
- NG
  - P(あり)=37/60≠45/100
  - P(非関西|なし)=8/23≠40/55
- GOOD
  - P(あり|非関西)=2/10=10/50
  - P(あり|関西)=35/50=35/50
- 標本抽出の方法が異なる
  - 表A:単純無作為抽出法
  - 表B:層別抽出法

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	2	8	10
計	37	23	60

表A

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	10	40	50
計	45	55	100

表B

「表B」の情報から「表A」に基づく正しい確率を求められるか？

- そのためにはどのような情報が必要か？
- 周辺確率が必要
  - P(関西)=5/6
  - P(非関西)=1/6
- ベイズの定理が有用

表A

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	2	8	10
計	37	23	60

表B

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	10	40	50
計	45	55	100

表A

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	2	8	10
計	37	23	60

表B

生育地	たこ焼器		計
	あり	なし	
関西	35	15	50
非関西	10	40	50
計	45	55	100

- 事前確率
  - P(関西)=50/60=5/6
  - P(非関西)=10/60=1/6

$$P(\text{あり}) = P(\text{あり}|\text{関西})P(\text{関西}) + P(\text{あり}|\text{非関西})P(\text{非関西})$$
$$= \frac{35}{50} \times \frac{5}{6} + \frac{10}{50} \times \frac{1}{6} = \frac{37}{60}$$

$$P(\text{非関西}|\text{なし}) = \frac{P(\text{なし} \cap \text{非関西})}{P(\text{なし})}$$
$$= \frac{P(\text{なし}|\text{非関西})P(\text{非関西})}{P(\text{なし})} = \frac{40/50 \times 1/6}{23/60} = \frac{8}{23}$$

シンプソンのパラドックス

- LINEを使うと夫婦別姓に賛成するのか？
- 中学生が理解できるように説明せよ

LINE(スマホ)	夫婦別姓	
	賛成	反対
使用	82	28
未使用	28	82

LINE(スマホ)	若年層		中高年層	
	夫婦別姓		夫婦別姓	
	賛成	反対	賛成	反対
使用	80	20	2	8
未使用	8	2	20	80

まとめ：クロス集計表とその見方

- 2×2分割表による関連性の検討
  - カイ2乗検定，クラメールのV，オッズ比
- 各種の確率とその運用
  - 言葉の定義
  - 加法定理，乗法定理，ベイズの定理の活用
- データの見方
  - データの発生機構を見破る眼力
    - 層別抽出法と単純無作為抽出法
    - 二項分布とたこう分布
  - シンプソンのパラドックス
- サイエンス・コミュニケーション
  - シンプソンのパラドックス

散布図と相関係数

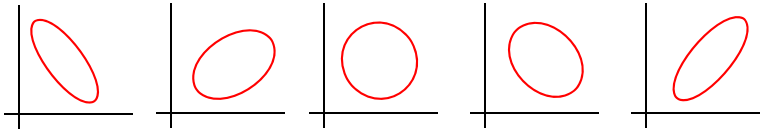
量的データの相関分析

相関関係と関数関係

- 関数関係 $y = f(x)$
  - x を定めるとy が一意的に決まる
- 相関関係 $y = f(x) + e$
  - xを定めてもyは分布する
  - xを定めるとyの分布が(一意的に)決まる
    - 一般に，yの分布はxと関係する

散布状況と相関係数

- 散布図の形と(ピアソンの)相関係数の対応を理解する
  - 形の数量化は難問
    - 散布図の形と相関係数は一対一に対応しない
    - 相関係数は変数間の直線関係を評価する指標
  - 相関の強さは相関係数だけで判断しない
    - 必ず散布図を描く

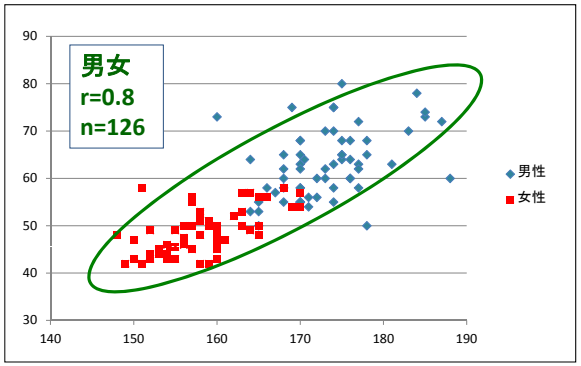


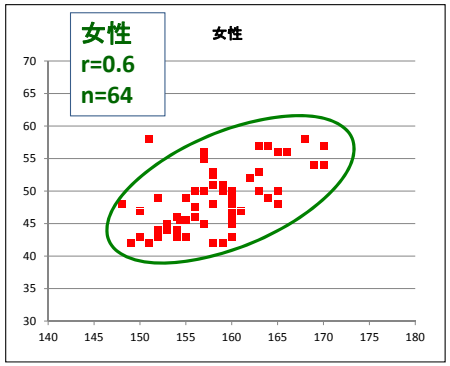
相関分析：身長と体重

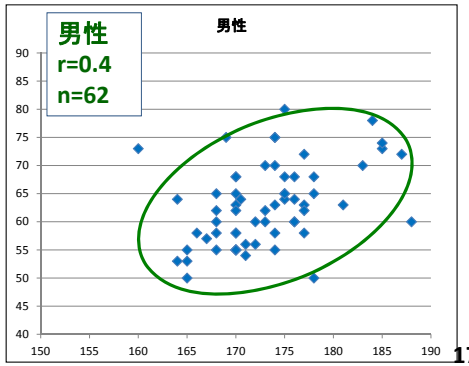
- 調査時期：1998.4
- 調査対象：O大学H学部1年次生(クラス内調査，無記名，欠席なし)

性別	身長	体重	性別	身長	体重	性別	身長	体重	性別	身長	体重	性別	身長	体重
男01	176	60	男32	178	50	女01	150	47	女32	157	55	女63	152	43
男02	177	63	男33	173	60	女02	152		女33	158	42	女64	164	
男03	168	58	男34	170.5	64	女03	160	50	女34	158		女65	151.5	
男04	178	68	男35	168	62	女04	165	48	女35	157	55	女66	162	52
男05	170	55	男36	168	60	女05			女36			女67	158	
男06	174	58	男37	177	72	女06	149	42	女37	150	43	女68	160	
男07	175	68	男38	175	64	女07	160	45	女38	148	48	女69		
男08	170	55	男39	170	65	女08	160	43	女39	158	51	女70	165	
男09	177	62	男40	169	75	女09	154.4	45.5	女40	160	50	女71	156	46
男10	170	58	男41	171	56	女10	159		女41	151	58	女72	156	46
男11	173	70	男42	170	68	女11	157	56	女42	158		女73	161	47
男12	164	53	男43	187	72	女12	159		女43	160	47	女74	151	42
男13	177	58	男44	174	63	女13	156	46	女44	164	57	女75	152	49
男14	184	78	男45	167	57	女14	155	49	女45	162		女76	153	45
男15	170	68	男46	168	55	女15	154	44	女46	163	50	女77	158	
男16	166	58	男47	165	53	女16	154	43	女47	158		女78	160	
男17	174	75	男48	165	50	女17	156	47.5	女48	160		女79	150	
男18	170	55	男49	188	60	女18	160	49	女49	156	50	女80	170	54
男19	174	55	男50	171	54	女19	153	44	女50	152	44	女81	159	51
男20	168	65	男51	185	73	女20	160	46	女51	164	49	女82	170	57
男21	160	73	男52	181	63	女21	159	50	女52	166.8		女83	157	45
男22	176	68	男53	178	65	女22	166	56	女53	165	56	女84	159	42
男23	170	63	男54	174	75	女23	158	48	女54	157	50	女85	158	53
男24	183	70	男55	165	55	女24			女55	154	46	女86	169	54
男25	172	56	男56	185	74	女25			女56	153		女87	163	57
男26	164	64	男57	176	64	女26			女57	169	54	女88	168	58
男27	172	60	男58	173	62	女27	160	48	女58	163		女89	155	45.5
男28	170	55	男59	170	62	女28	154	46	女59			女90	156	
男29	170	58	男60	176	60	女29			女60	165	50	女91	160	48
男30	175	80	男61	175	65	女30	163	53	女61	160	48	女92	164	57
男31	175	65	男62	174	70	女31	158	52.5	女62	155	43	女93		

身長と体重のデータ (CCA,LD)





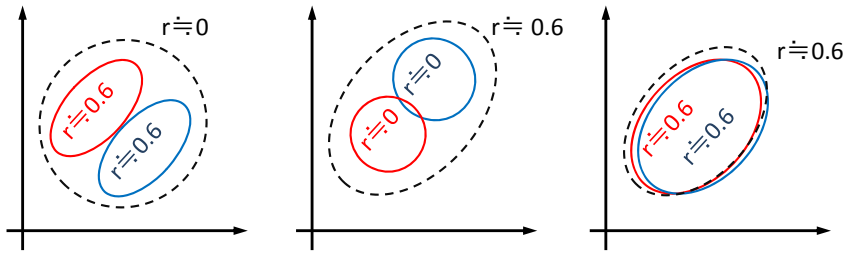


ピアソンの標本相関係数を用いる時の注意

- 形の数量化は難問
  - 散布図の形と相関係数は一対一に対応しない
- いくつかの観点
  - 標本サイズを考慮すべし
    - 検定，区間推定
  - 直線的な関係を測る尺度
    - 非線形の関係の評価は不適切
  - 外れ値
  - 複数個の集団が混在
  - 標本の偏り
    - サンプルセレクション（切断された集団，選抜効果）
    - 欠測
  - 擬相関
    - 因果と相関

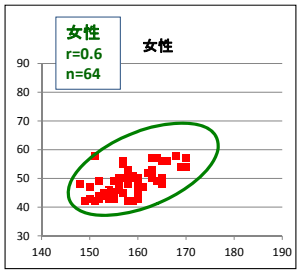
データの合併

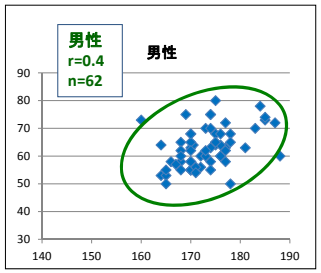
- 等質なデータは合併した方がよい
- 異質なデータの合併は真実を見えにくくする
- データを合併するのは，データの特徴に違いが無い場合に限られる
  - 分析の精度が向上する
- 一般には，一組のデータを，異なった特徴をもつ複数個のグループに分ける作業が重要である
  - 層別という



考察：女性の相関が少し高い？

- データ採取方法
  - ある授業でのアンケート調査（無記名）
    - 対象：O大学H学部1年次生
  - 実測でない





考察：女性の相関が少し高い？

- 女性は体型に気を配っている
  - 食生活に注意し，スタンダードな体型を保つように努力
- 女性の回答に偏りの可能性がある
  - 理想体型により近い値を回答
- 男性は体型に興味がない
  - 記憶があいまい
  - 回答に誤差 → 相関が低下
- 女性に回答拒否が多い
  - どういった女性が回答拒否しやすいか？
- 外れ値っぽい個体がある
- 本当に差があるのか
- 体重の分布は正規でない

真の推定値！

- 学校保健統計調査(2010)
  - 統計法第33条に基づく調査票情報提供の申出 →文科省
  - 身長・体重の2次元データ
  - 17歳，男21,108，女21,115
    - 欠測obs（男1127; 女1007)
    - MCARとみてよい（身体測定に欠席のようである）
- 真の推定値
  - 17歳男：r=0.412
  - 17歳女：r=0.428

女性の相関が少し高い？

- 女性は体型に気を配っている
  - 食生活に注意し，スタンダードな体型を保つように努力
- 女性の回答に偏りの可能性がある
  - 理想体型により近い値を回答
- 男性は体型に興味がない
  - 記憶があいまい
  - 回答に誤差 → 相関が低下
- 女性に回答拒否が多い
  - どういった女性が回答拒否しやすいか？
- 外れ値っぽい個体がある
- 本当に差があるのか
  - 有意性検定
- 体重は正規分布しない

本格的な統計分析へ

- 回答拒否
  - 男性（n=62)
    - 拒否なし
  - 女性（n=93）
    - 完全データ：64
    - 欠測データ：29
      - 身長と体重欠測： 9
      - 体重のみ欠測： 20
      - 身長のみ欠測： 0
- 欠測やサンプルセレクションのモデリング
  - 近々，どこかで発表
- 記憶があいまい．回答に誤差
  - 変量内誤差モデルによる分析
  - 近々，どこかで発表
- 他

まとめ

- 身長＋体重なるシンプルなデータから様々な統計キーワードを学ぶことができる
  - 身長と体重の分布
  - 散布図と相関係数の対応 (r = 0.4, 0.6, 0.8)
  - データの合併・層別
  - 女性の相関がより高い？
    - 外れ値，欠測値
      - データ採取状況をよく検討する
      - 教員は男性？無記名？
    - 測定誤差
    - 有意性検定

次回予告：教養統計学での講義例

- データの見方の教授法
- 標本分布
- 不偏推定
- 第9回統計教育方法論WS
  - 相関分析・関連性分析
    - 欠測とサンプリングの相関分析・関連性分析への影響：関西にタコ焼き器は多いのか？
- 第7回統計教育方法論WS
  - 条件付き確率
    - 統計学者が講義する条件付き確率：モンティホール問題など

次回予告: 不偏推定

打率の推定

プロ野球において，選手 A は前半戦 200 打数 60 安打，後半戦 100 打数 10 安打であった．通算成績は，普通の平均（単純平均）

$$\hat{p} = \frac{\frac{60}{200} + \frac{10}{100}}{2} = \frac{80}{400}$$

では算出されず，

$$\hat{p}' = \frac{60 + 10}{200 + 100} = \frac{70}{300}$$

と計算される．この理由を，(1) 小学生，(2) 中学生，(3) 高校生，(4) 大学生，のそれぞれが理解できるように説明せよ．

Any questions are welcome!