第397回 KSP例会
日時：2013/01/12(土) 15:30-17:30
会場：基礎工学 国際棟 セミナー室

欠測値データ解析の意味と有効性

狩野　裕
大阪大学 基礎工学研究科
数理科学領域

1

基礎工学研究科 数理科学領域 統計数理講座

データ科学研究グループ

- スタッフ（3名）
- 2004年度～2013年度入学の院生
  - 修士課程学生の出身学部（34名）
    - 阪大基礎工学部：　15名
    - 阪大人間科学部：　6名
    - 他大学：　13名
  - 博士課程学生の出身大学院（12名，含社D）
    - 阪大基礎工修士：　6名
    - 阪大人科修士：　4名
    - 他大学修士：　2名
- 統計学に興味のある学生は，当研究グループへの進学もお考え下さい

3

Agenda

- 欠測値データ解析の基礎
- 分析事例1：経時データにおける欠測
- 分析事例2：大量欠測データ
- 分析事例3：相関分析における欠測
（付録，未完）

4

MCAR, MAR, NMAR
LD(CCA), PD
欠測値の適切な処理

欠測値データ解析の基礎

5

不完全データ

- Incomplete data
  - 計画通りの完全な形で得られないデータ
- 欠損値（欠測値）データ
  - 打ち切り，脱落，未回答，実験の失敗
- サンプルセレクション
  - 母集団から無作為標本抽出でない
  - 未回収データ
  - 標本抽出の計画そのものに問題がある場合も
  - 「サンプルの偏り」と呼ばれることもある
- 粗いデータ（coarse data)
  - （連続変数の）カテゴリ化，グループ化
- 秘匿，捏造

6

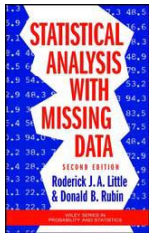
欠測値問題

- 欠損値と言うことも
  - Missing values, missing data
- 実際の応用場面でしばしば出会う重要な問題
- 統計理論としても興味ある話題
- Donald Rubinが1970年代に先駆的な仕事

7

Two Bibles

- Little and Rubin (2002)
  - Roderick Little
    - Biostatistics Department
      Statistics Department
      University of Michigan
- Allison (2001)
  - Paul Allison
    - Department of Sociology
      University of Pennsylvania



8

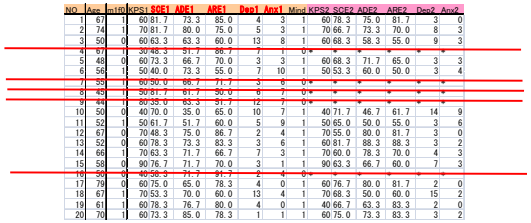
欠測値の発生要因が増加

- 個人情報，プライバシーの保護
  - 回答拒否
    - E.g. 国勢調査
- 実験協力者への配慮
  - 途中離脱（ドロップアウト）
    - 途中離脱したければいつでもそれを認める
- 欠測を出さないように強制することもある
  - 回答の精度が減少する
- 欠測データの対処法で多いのは
  - Listwise deletion (Complete Case Analysis)
  - Pairwise deletion

9

Listwise deletion(LD)

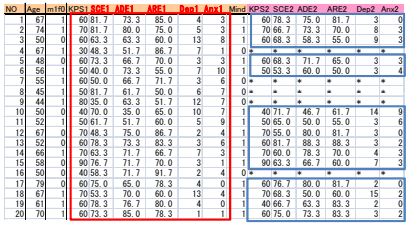
- Complete case analysis (CCA)
- 完全に揃っているケース（observation, record, unit）だけを統計解析の対象にする
  - 一つでも欠損したケースは分析から外す



10

Pairwise deletion(PD)

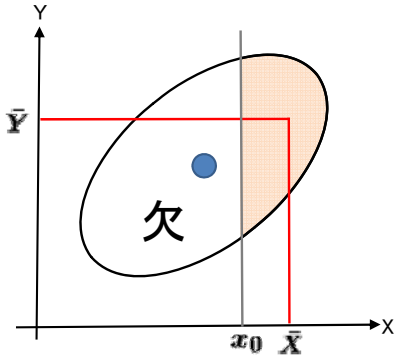
- 相関係数の推定に用いられることがある
- 相関を計算したい変数ペアだけをみて，使えるデータをすべて使う



11

LD(CCA)のバイアス

i	X	Y
1	X_1	Y_1
⋮	⋮	⋮
m	X_m	Y_m
$m+1$	X_{m+1}	欠
⋮	⋮	⋮
n	X_n	欠



欠測メカニズム：

Y is observed iff $X \geq x_0$

Y is missing iff $X < x_0$

for some $x_0 \in R^1$.

12

Statistical Methods in Psychology Journals: Guidelines and Explanation

- The two popular methods for dealing with missing data that are found in basic statistics packages -listwise and pairwise deletion of missing values- are among the worst methods available for practical applications(p.598)
  - L. Wilkinson and APA Task Force on Statistical Inference APA Board of Scientific Affairs
  - American Psychologist 1999

13

Publication Manual of the APA (2009, 6th edition)

- Similarly, missing data can have a detrimental effect on the legitimacy of the inferences drawn by statistical tests. For this reason, it is critical that the frequency or percentages of missing data be reported along with any empirical evidence and/or theoretical arguments for the causes of data that are missing. For example, data might be described as MCAR; MAR; or NMAR. It is also important to describe the methods for addressing missing data, if any were used (e.g., multiple imputation).

14

高名な心理学者より（H22.9.9）

- ミッシングバリューのあつかいですよね。これって、われわれ通常は、統計パッケージのオプションをそのまま使いますし、ミッシングバリューの扱いについて、論文で報告することはまずありません。
- 実際の査読のプロセスで、この論文を根拠に分析のしなおしを要求することが起こるかどうか、と言うことが問題ですよね、たぶん。

15

MCAR, MAR, NMAR

- Missing Completely At Random (MCAR)
  - 完全にランダムに欠測する
    - 各データの欠測確率が「どの」データにも依存しない
- Missing At Random (MAR)
  - 欠測が観測されたデータに依存してもよいが，欠測値には依存しない
    - 各データの欠測確率が「欠測」データに依存しない
- Not Missing At Random (NMAR, MNAR)
  - MARの否定
  - データの欠測確率が「欠測」データにも依存

16

MCAR, MAR, NMAR

	i	X	Y	R^X	R^Y
I_{11}	1	X_1	Y_1	1	1
	⋮	⋮	⋮	⋮	⋮
I_{10}	n_1	X_{n_1}	欠	1	0
	⋮	⋮	⋮	⋮	⋮
I_{01}	n_2	欠	Y_{n_2}	0	1
	⋮	⋮	⋮	⋮	⋮
I_{00}	n_3	欠	欠	0	0
	⋮	⋮	⋮	⋮	⋮
	n	欠	欠	0	0

確率変数(R^X, R^Y)を欠測指標(missing indicator; response indicator)と言う．
条件付き確率$P(R^X, R^Y|X, Y)$を欠測メカニズム(missing-data mechanism)と言う．

17

MCAR, MAR, NMAR

- Donald Rubin, Rod Little の貢献

観測変数ベクトル: $\boldsymbol{Y} = [Y_1, \ldots, Y_p]' = [\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}]'$

欠測指標ベクトル: $\boldsymbol{R} = [R_1, \ldots, R_p]'$

MCAR: $P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) = P(\boldsymbol{R})$

MAR: $P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) = P(\boldsymbol{R}|\boldsymbol{Y}_{obs})$

NMAR: $P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \neq P(\boldsymbol{R}|\boldsymbol{Y}_{obs})$

欠測メカニズム: $P(\boldsymbol{R}|\boldsymbol{Y})$

18

推定方法

- MACRの場合
  - LD，PDは適用可能．推定精度が落ちる
  - （完全情報）最尤法(FIML)が勧められる
- MARの場合
  - LD，PDはバイアスが生じるので適用不可
  - FIMLを適用しなければならない
- NMARの場合
  - 欠測メカニズムを同定しFIMLを適用
    - 欠測メカニズムの同定は一般に困難
  - 目をつむってMAR
  - 補助変数法
    - 実際的な方法論．確立されてはいない

19

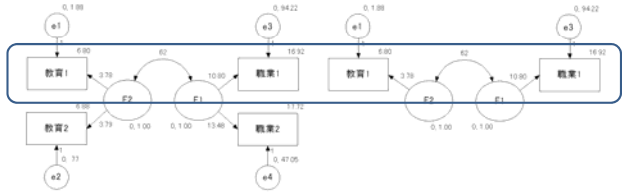
例

- MCARの例
  - 反復測定において，1回目の被調査者の中から無作為に選択し2回目の測定を実施．選択されなかった被調査者は2回目の測定が欠測
- MARの例
  - 入学試験と入学後の成績との関連を調べたいが，不合格者の入学後の成績は存在せず欠測
- NMAR
  - MAR以外

20

例：MCAR

第1回目
n=2020



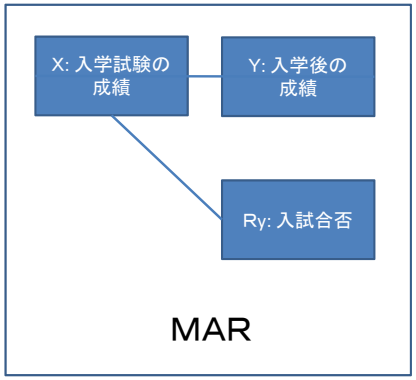
一回目と二回目の
両方測定
n=348

一回目のみ測定
（二回目は欠測）
n=1672

第二回目の測定は，n=2020の中から無作為にn=348を抽出
Allison(1987)，Wothke (2000) 計画による欠測

21

例：MAR



i	X	Y	R_y
1	X_1	Y_1	1
⋮	⋮	⋮	⋮
m	X_m	Y_m	1
$m+1$	X_{m+1}	欠	0
⋮	⋮	⋮	⋮
n	X_n	欠	0

欠測メカニズム

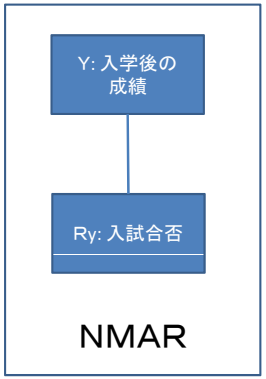
Y is observed iff $X \geq x_0$
Y is missing iff $X < x_0$
for some $x_0 \in R^1$.

MAR条件を満たす：

$$P(R_y = 0|X, Y) = P(R_y = 0|X)$$

22

例： NMAR



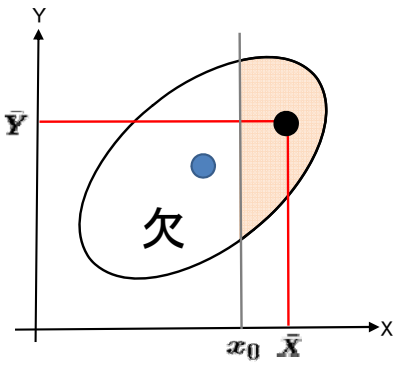
i	Y	R_y
1	Y_1	1
⋮	⋮	⋮
m	Y_m	1
$m+1$	欠	0
⋮	⋮	⋮
n	欠	0

$$P(R_y = 0|Y) \neq P(R_y = 0)$$

- 1次元データは
  - MCAR or NMAR

23

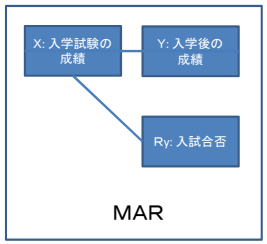
MARのときLD(CCA)は機能しない！



- LD(CCA)は大きなバイアスを生じる可能性

⇒

- 最尤推定(FIML)を用いる



24

尤度と最尤推定(FIML)

- Full-Information Maximum Likelihood
  - Direct Maximum Likelihood と呼ばれることも

$$L(\varphi|X, Y) = \prod_{i=1}^{m} N_2\left(\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \middle| \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}\right) \times \prod_{i=m+1}^{n} N(X_i|\mu_x, \sigma_{xx})$$

where $\varphi = (\mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})$

参考文献：岩崎(2001, § 5.2), Anderson (1957)

- AMOSの推定方法
- 線形混合モデル（SAS，SPSS…）

25

尤度関数（FIML）

$$L(\mu_x, \mu_y, \sigma_{xx}, \sigma_{xy}, \sigma_{yy} | X, Y)$$

$$= \prod_{i=1}^{m} \frac{1}{2\pi\sqrt{\sigma_{xx}\sigma_{yy}(1-\rho^2)}}$$

$$\times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{X_i-\mu_x}{\sqrt{\sigma_{xx}}}\right)^2 - 2\rho\left(\frac{X_i-\mu_x}{\sqrt{\sigma_{xx}}}\right)\left(\frac{Y_i-\mu_y}{\sqrt{\sigma_{yy}}}\right) + \left(\frac{Y_i-\mu_y}{\sqrt{\sigma_{yy}}}\right)^2\right\}\right]$$

$$\times \prod_{i=m+1}^{n} \frac{1}{\sqrt{2\pi\sigma_{xx}}} \exp\left[-\frac{1}{2}\left(\frac{X_i-\mu_x}{\sqrt{\sigma_{xx}}}\right)^2\right]$$

with $\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}$

- 多母集団の同時分析
  - 欠測パターンの数だけ母集団を考える
  - 対応する母数は母集団間で等値する
    - 平均も推定し等値する

26

完全データにおける諸統計量

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

$$s_{xx} = \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$s_{yy} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

$$s_{xy} = \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})$$

i	X	Y	R_y
1	X_1	Y_1	1
⋮	⋮	⋮	⋮
m	X_m	Y_m	1
$m+1$	X_{m+1}	欠	0
⋮	⋮	⋮	⋮
n	X_n	欠	0

27

最尤推定量（FIMLE)

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}_{xx} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

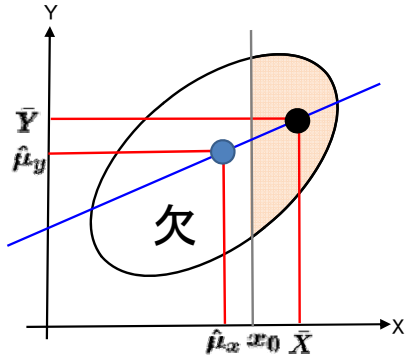
$$\hat{\mu}_y = \bar{Y} + \frac{s_{xy}}{s_{xx}}(\hat{\mu}_x - \bar{X})$$

$$\hat{\sigma}_{yy} = s_{yy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2(\hat{\sigma}_{xx} - s_{xx})$$

$$\hat{\sigma}_{xy} = s_{xy}\frac{\hat{\sigma}_{xx}}{s_{xx}} \quad \left[\text{or} \quad \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_{xx}} = \frac{s_{xy}}{s_{xx}}\right]$$

28

FIMLEの図的解釈



$$y = \bar{Y} + \frac{s_{xy}}{s_{xx}}(x - \bar{X})$$

$$y = \hat{\mu}_y + \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_{xx}}(x - \hat{\mu}_x)$$

これらは同値．よって

$$\begin{cases} \hat{\mu}_y = \bar{Y} + \dfrac{s_{xy}}{s_{xx}}(\hat{\mu}_x - \bar{X}) \\ \dfrac{\hat{\sigma}_{xy}}{\hat{\sigma}_{xx}} = \dfrac{s_{xy}}{s_{xx}} \end{cases}$$

29

FIMLEの解釈(Yの分散)

$$\hat{\sigma}_{yy} = s_{yy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2 (\hat{\sigma}_{xx} - s_{xx})$$

$$\Longleftrightarrow$$

$$\hat{\sigma}_{yy} - \left(\frac{s_{xy}}{s_{xx}}\right)^2 \hat{\sigma}_{xx} = s_{yy} - \left(\frac{s_{xy}}{s_{xx}}\right)^2 s_{xx}$$

$$\Longleftrightarrow$$

$$\hat{\sigma}_{yy} - \frac{\hat{\sigma}_{xy}^2}{\hat{\sigma}_{xx}} = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = \hat{\sigma}_e^2$$

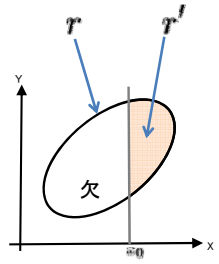
30

合格者の相関 r'から全員の相関 r を推計

$$r = \frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}} = \frac{s_{xy}\frac{\hat{\sigma}_x^2}{s_x^2}}{\sqrt{\hat{\sigma}_x^2\left(s_y^2 + \left(\frac{s_{xy}}{s_x^2}\right)^2 (\hat{\sigma}_x^2 - s_x^2)\right)}}$$

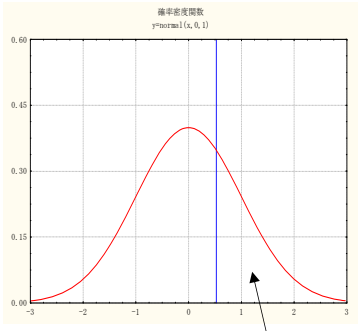
$$= \frac{r'}{\sqrt{(1-k^2)(r')^2 + k^2}}$$

where $r' = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$, $k^2 = \frac{s_x^2}{\hat{\sigma}_x^2}$



31

相関係数の修正公式



入学試験Xの分布

30%

$$r = \frac{r'}{\sqrt{(1-k^2)r'^2 + k^2}}$$

ここで

$$k^2 = \frac{\text{合格者の } X\text{の分散}}{\text{受験者全員の } X\text{の分散}}$$

	合格率		
	30%	20%	10%
k^2 =	0.27	0.22	0.17
r' =	0.30	0.30	0.30
r =	0.51	0.54	0.59

32

この修正公式は well-known

- Pearson, K. (1903)
  - Philos. Trans. Royal Soc. London A200.
- Lord and Novick (1968)
  - Statistical Theories of Mental Test Scores.
- Rubin, D. B. (1976)
  - Inference and Missing Data. Biometrika.
  - MAR+FIML

- 芝＋渡部(1988). 入試データの解析. 新曜社
- 岡田謙介・繁桝算男(2010)
  - 「小標本における選抜効果を補正する相関係数の推定について」日本テスト学会誌.
- 岡田謙介(2010) personal communication

33

まとめ

- 相当割合の欠測値を含むときは，どのような理由で欠測したのか，すなわち，欠測メカニズムに思いを馳せることが大事
  - MARを仮定したFIMLで分析してみる
  - NMARを想定した分析をせざるを得ないかもしれない
- 欠測値を含むオブザベーションが少なければ，CCA（LD）で大きな問題にならないだろう

34

末期がん患者の心理的適応に関する研究
多くの脱落（ドロップアウト）を含むデータの分析

分析事例１：経時データにおける欠測

35

末期がん患者の心理的適応に関する研究

- 平井　啓・鈴木要子・恒藤　暁・池永昌之・柏木哲夫
  - 末期がん患者のセルフ・エフィカシーと心理的適応の時系列変化に関する研究
  - 心身医学, 2002, 42, 111-118
- 平井 啓
  - 末期がん患者の心理的適応に関する研究
  - 学位論文，2002, 大阪大学

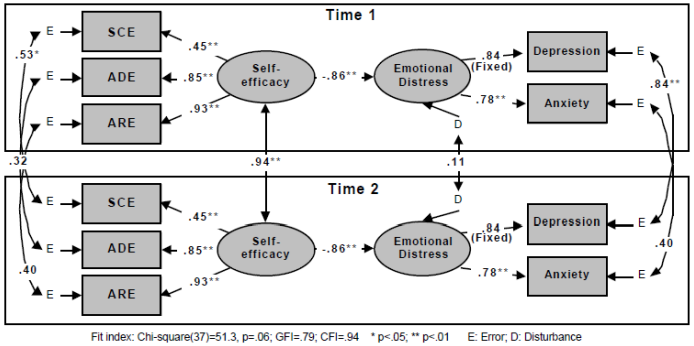
36

末期がん患者のセルフ・エフィカシー

- 末期がん患者用セルフエフィカシー尺度
  - SCE: 身体症状に対する効力感（Symptom Coping Efficacy）
  - ADE: 日常生活動作に対する効力感（ADL Efficacy)
  - ARE: 情動統制に対する効力感（Affect Regulation Efficacy）
- 抑うつ・不安の尺度
  （HAD: Hospital Anxiety and Depression Scale）
  - Dep，Anx
- 標本
  - 末期がん患者 n=42
  - 原則として二週間おいて反復測定（測定回数：2回）
  - 完全データ n=31
    - 11名は二回目ドロップアウト

37

平井氏の分析：LD(n=31)



共分散行列の推定

Listwise Deletion (n=31) 標本分散

	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	179.9									
ADE1	39.3	249.2								
ARE1	66.4	199.0	272.9							
DEP1	−13.2	−43.2	−46.1	16.9						
ANX1	−13.5	−31.6	−31.4	8.5	10.0					
SCE2	100.4	67.0	114.5	−13.2	−10.7	146.9				
ADE2	48.3	166.7	177.1	−38.0	−22.5	89.1	215.7			
ARE2	64.4	149.5	206.0	−38.1	−27.5	112.9	172.6	198.9		
DEP2	−13.1	−43.3	−51.4	14.9	7.1	−18.4	−45.6	−42.1	20.2	
ANX2	0.9	−26.7	−33.6	6.7	6.3	−7.5	−25.1	−28.2	9.2	9.2
Mean	70.8	72.6	72.2	5.7	3.8	69.2	72.7	72.2	5.5	3.1

MAR FIML (n=42)

	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	230.7									
ADE1	73.4	228.4								
ARE1	67.4	170.2	254.0							
DEP1	−17.1	−37.8	−42.1	15.4						
ANX1	−18.6	−31.6	−33.0	8.1	10.8					
SCE2	119.0	72.7	108.1	−14.4	−13.6	150.2				
ADE2	63.6	148.5	157.4	−34.3	−22.6	89.4	201.1			
ARE2	69.5	131.9	194.8	−35.5	−29.3	111.0	160.3	192.9		
DEP2	−15.4	−36.9	−46.5	13.5	6.8	−18.3	−41.1	−38.8	18.8	
ANX2	3.0	−21.8	−33.2	6.0	6.5	−6.8	−22.5	−28.0	8.5	9.6
Mean	67.8	70.5	71.8	5.9	4.3	68.0	71.7	71.7	5.6	3.1

LD−FIML

	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	−50.7									
ADE1	−34.1	20.8								
ARE1	−1.0	28.8	18.9							
DEP1	3.9	−5.4	−3.9	1.6						
ANX1	5.1	0.0	1.6	0.4	−0.8					
SCE2	−18.6	−5.7	6.4	1.2	3.0	−3.3				
ADE2	−15.3	18.2	19.8	−3.8	0.1	−0.3	14.5			
ARE2	−5.2	17.6	11.2	−2.7	1.7	1.9	12.3	6.0		
DEP2	2.3	−6.4	−4.9	1.4	0.3	−0.2	−4.5	−3.3	1.4	
ANX2	−2.1	−5.0	−0.5	0.7	−0.2	−0.8	−2.7	−0.2	0.7	−0.3
Mean	3.0	2.2	0.4	−0.1	−0.4	1.2	0.9	0.6	−0.1	0.0

平均の差異：
効力感：完全データ(LD)が高い
抑うつ：完全データ(LD)が低い

相関行列の推定

Listwise Deletion (n=31)

	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	1.00									
ADE1	0.19	1.00								
ARE1	0.30	0.76	1.00							
DEP1	−0.24	−0.67	−0.68	1.00						
ANX1	−0.32	−0.63	−0.60	0.65	1.00					
SCE2	0.62	0.35	0.57	−0.27	−0.28	1.00				
ADE2	0.25	0.72	0.73	−0.63	−0.48	0.50	1.00			
ARE2	0.34	0.67	0.88	−0.66	−0.62	0.66	0.83	1.00		
DEP2	−0.22	−0.61	−0.69	0.61	0.50	−0.34	−0.69	−0.67	1.00	
ANX2	0.02	−0.56	−0.67	0.54	0.66	−0.20	−0.56	−0.66	0.67	1.00
MEAN	70.80	72.64	72.20	5.71	3.84	69.19	72.68	72.20	5.48	3.10
VARIANCE	179.92	249.17	272.88	16.92	10.01	146.85	215.65	198.90	20.19	9.25

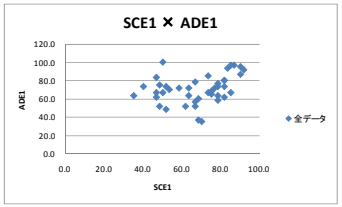
MAR FIML (n=42)

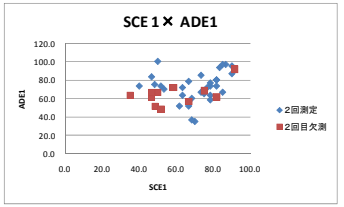
	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	1.00									
ADE1	0.32	1.00								
ARE1	0.28	0.71	1.00							
DEP1	−0.29	−0.64	−0.68	1.00						
ANX1	−0.37	−0.64	−0.63	0.63	1.00					
SCE2	0.64	0.39	0.55	−0.30	−0.34	1.00				
ADE2	0.30	0.69	0.70	−0.62	−0.48	0.51	1.00			
ARE2	0.33	0.63	0.88	−0.65	−0.64	0.65	0.81	1.00		
DEP2	−0.24	−0.56	−0.67	0.80	0.48	−0.34	−0.67	−0.65	1.00	
ANX2	0.06	−0.47	−0.67	0.49	0.64	−0.18	−0.51	−0.65	0.64	1.00
MEAN	67.78	70.48	71.83	5.86	4.26	67.95	71.75	71.65	5.56	3.06
VARIANCE	230.65	228.42	254.03	15.36	10.81	150.19	201.11	192.93	18.76	9.59

	LD-FIML									
	SCE1	ADE1	ARE1	DEP1	ANX1	SCE2	ADE2	ARE2	DEP2	ANX2
SCE1	0.00									
ADE1	−0.14	0.00								
ARE1	0.02	0.06	0.00							
DEP1	0.05	−0.03	0.00	0.00						
ANX1	0.05	0.00	0.03	0.02	0.00					
SCE2	−0.02	−0.04	0.02	0.04	0.06	0.00				
ADE2	−0.05	0.03	0.03	−0.01	0.00	−0.01	0.00			
ARE2	0.01	0.04	0.00	−0.01	0.02	0.01	0.02	0.00		
DEP2	0.02	−0.05	−0.02	0.01	0.02	0.00	−0.02	−0.02	0.00	
ANX2	−0.04	−0.09	0.00	0.05	0.02	−0.03	−0.05	−0.01	0.04	0.00
MEAN	3.02	2.15	0.37	−0.15	−0.42	1.24	0.93	0.55	−0.07	0.03
VARIANCE	−50.73	20.74	18.85	1.56	−0.81	−3.34	14.54	5.97	1.42	−0.34

42

欠測パターンと基本統計量





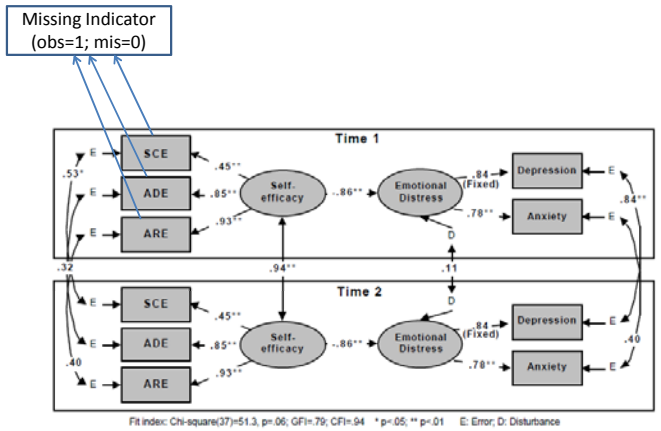
第一回目の測定の平均

	SCE1	ADE1	ARE1	Dep1	Anx1
完全データ(n=31)	70.8	72.6	72.2	5.7	3.8
欠測データ(n=11)	59.2	64.4	70.8	6.3	5.5
差(上−下)	11.6	8.2	1.4	−0.6	−1.8

平均の差異:
効力感: 完全データが高い
抑うつ: 完全データが低い

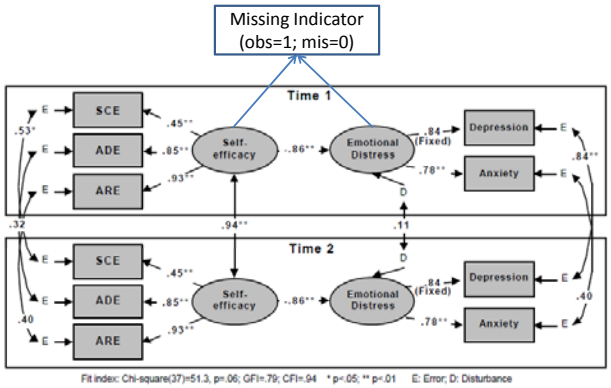
43

MAR



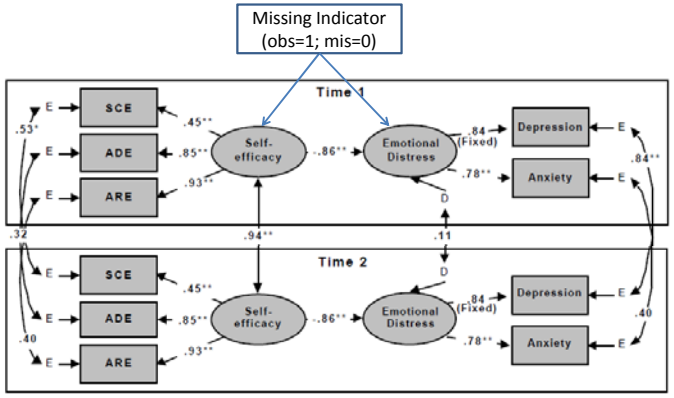
44

NMAR: shared-parameter model



45

NMAR: pattern mixture model



Fit index: Chi-square(37)=51.3, p=.06; GFI=.79; CFI=.94 * p<.05; ** p<.01 E: Error; D: Disturbance

分析結果

時点間で等値制約あり

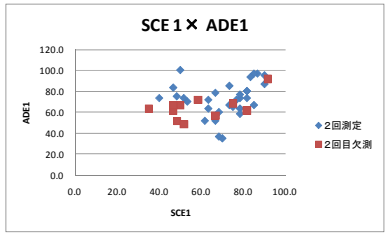
		LD (CCA)	MAR (DML)	MAR (full)		Selection Model (Shared-parameter model)			Pattern Mixture Model		
						SE	ED	SE+ED	SE	ED	SE+ED
推定値 (SE→ED)	time1	−0.86	−0.92	−0.92	time1	−0.92	−0.93	−0.92	−0.95	−0.90	−0.97
	time2	−0.89	−0.93	−0.93	time2	−0.94	−0.94	−0.94	−0.93	−0.93	−0.89
欠測指標に関する検定統計量 (est/s.e.)	SCE1			1.926	SE	1.009		0.031	0.904		0.999
	ADE1			1.642	ED		−1.020	−0.180		−0.417	−0.187
	ARE1			−1.238							
対数尤度		−973.64	−1162.71	−1162.97		−1186.33	−1186.31	−1186.31	−1185.74	−1186.17	−1185.72
自由母数の数		19	29	33		31	31	32	32	32	33
Akaike (AIC)		1985.27	2383.42	2431.93		2434.65	2434.63	2436.62	2435.48	2436.34	2437.45
Bayesian (BIC)		2012.52	2433.81	2489.28		2488.52	2488.50	2492.23	2491.09	2491.95	2494.79
Sample-Size Adjusted BIC		1953.32	2343.00	2385.94		2391.44	2391.42	2392.02	2390.88	2391.74	2391.45

時点間で等値制約なし

		LD (CCA)	MAR (DML)	MAR (full)		Selection Model (Shared-parameter model)			Pattern Mixture Model		
						SE	ED	SE+ED	SE	ED	SE+ED
推定値 (SE→ED)	time1	−0.81	−0.93	−0.93	time1	−0.94	−0.93	−0.93	−0.95	−0.86	−0.85
	time2	−0.79	−0.82	−0.82	time2	−0.83	−0.83	−0.87	−0.82	−0.82	−0.82
欠測指標に関する検定統計量 (est/s.e.)	SCE1			1.926	SE	0.921		−0.642	0.879		0.713
	ADE1			1.642	ED		−0.959	−0.732		−0.683	−0.399
	ARE1			−1.238							
対数尤度		−986.18	−1154.26	−1174.53		−1177.99	−1177.94	−1176.28	−1177.98	−1178.20	−1177.90
自由母数の数		29	39	43		41	41	42	41	41	42
Akaike (AIC)		1990.36	2386.55	2435.06		2437.97	2437.88	2436.56	2437.96	2438.39	2439.80
Bayesian (BIC)		2031.94	2454.32	2509.76		2509.22	2509.12	2509.54	2509.20	2509.64	2512.78
Sample-Size Adjusted BIC		1941.59	2332.19	2375.13		2380.83	2380.73	2378.02	2380.81	2381.24	2381.25

考察：欠測メカニズム

- 欠測パターンによって分布が異なる
  - MCARではない
  - e.g. SCE1, ADE1
  - この分布の差異について何らかの説明が要る



考察：標準誤差

等値制約あり

		Parameters	Estimates	S.E.	Est./S.E.	StdYX
FIML	n=42	SE1→ED1	−2.96	0.53	−5.63	−0.92
		SE2→ED2	−3.27	0.63	−5.17	−0.93
LD (CCA)	n=31	SE1→ED1	−2.91	0.64	−4.56	−0.86
		SE2→ED2	−3.36	0.68	−4.97	−0.89

等値制約なし

		Parameters	Estimates	S.E.	Est./S.E.	StdYX
FIML	n=42	SE1→ED1	−2.82	0.54	−5.23	−0.93
		SE2→ED2	−2.87	0.68	−4.25	−0.82
LD (CCA)	n=31	SE1→ED1	−2.65	0.63	−4.24	−0.81
		SE2→ED2	−2.80	0.69	−4.03	−0.79

結論

- LD(CCA)による分析結果は他の分析と異なるようである
  - 欠測メカニズムはMCARとは考え難い
  - LD(CCA)以外はかなり似通っている
- 欠測指標へのパス
  - NMARのモデルでは有意ではない
  - MARのモデルでは有意に近い
- このデータについてはMARを仮定してもよいのではないか
  - 経時データの欠測はMARが妥当であることが多い
  - FIMLが勧められる
  - DLはバイアスを生じている．小さめに推定
- たとえMCARが妥当と考える場合でも
  - LDはFIMLに比して標準誤差が大きい
    - LDはサンプルを捨てるので非効率
    - n=42は小標本．貴重なデータを活かしたい
- この実証分析の解釈は変わらない
  - このモデルによると、第1面接時のセルフ・エフィカシーと抑うつ・不安の関係と第2面接時のそれとでは、統計的に違いがあることは示されなかった。（平井　2002，p.82）

50

釈明の例

The data contain substantial missing values at the second measurement (t=2). We have assumed that the missing-data mechanism was MAR (e.g., Little and Rubin 2002) and that the data were analyzed with FIML (Graham, 2009) under normality assumption. It will not be unrealistic that the first measurement can predict the missingness at the second. It should be noted that similar estimates were obtained when analyzed after listwise deletions.

51

NTTみらいネット研究所＋廣瀬慧助教との共同研究
第一印象の評価

分析事例2：大量欠測データ

52

第一印象の構造を探る

- 実験協力者（被験者，40歳未満）に
  人物刺激を与え，第一印象を評価させる
  - 人物刺激：　4人
  - 評価項目：94個
  - その他の項目あり
- 本調査研究の（中間）ゴール
  - 適切な評価項目の選択
    - 因子を共有する
    - しばしば用いられる評価項目の同定
  - 評価項目の特徴を捉える
    - 潜在因子の同定

53

多くの評価項目（94個）を用意する

- 長所
  - 実験協力者の評価軸の多様性に対応
  - 似た項目を複数個用意し，それらの回答を平均することで，測定誤差を減少させる
    - 尺度や多重指標の考え方
- 短所
  - 実験協力者に過大な負担がかかる
  - 回答の精度が下がる
    - 最後まで答えてくれない（回答を中断する）
    - 後半では回答が適当になり精度が落ちる

56

回答方法に工夫

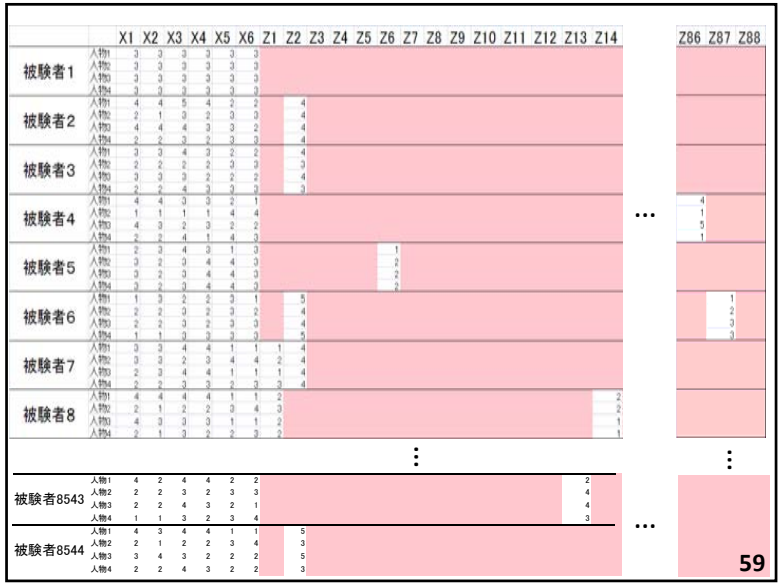
- 典型的な評価項目を6個を事前選定
  - 共通項目とし，実験協力者全員が回答
- 88個の選択項目を用意
  - 88項目から4項目を選択回答
- 4人の人物刺激
- 各実験協力者は
  (6+4)×4=40
  の評価項目に回答する
  - 実行可能な回答数
  - cf. (6+88)×4=376
- インターネット調査
  - n= 8,542

57

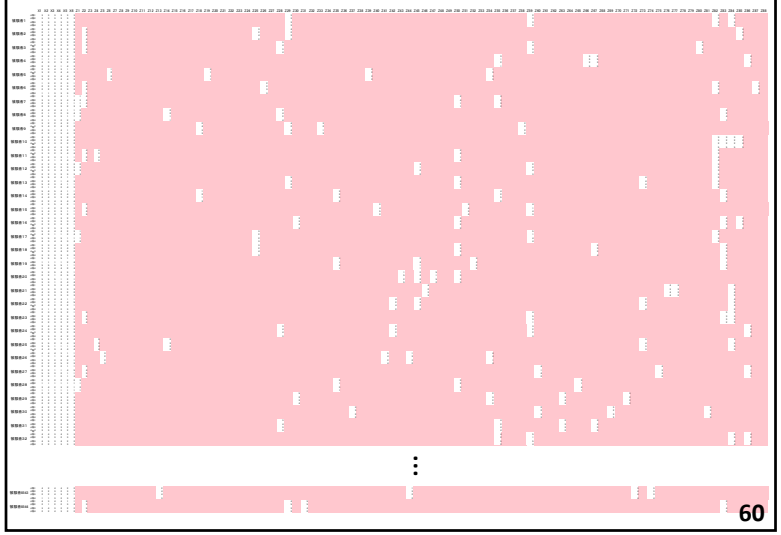
評価項目（必答と選択）

- 必ず回答する評価項目　（共通特性項目，6項目）
  1. 感じのよい　—　感じの悪い
  2. 親しみやすい　—　親しみにくい
  3. 慎重な　—　軽率な
  4. 分別のある　—　分別のない
  5. 積極的な　—　消極的な
  6. 自信のある　—　自信のない
- 選択して回答する評価項目　（個別特性項目，88項目）
  - きちんとしている　— だらしない
  - 不潔な　— 清潔な
  - ・・・・

58



59



60

評価項目の選択回答

- 長所
  - 実験協力者の多様な評価軸に対応
    - 個々人の評価軸は多くない
  - 実験協力者の負担を減少させ，回答の精度を向上させる
- 短所
  - データに大量の欠測（未回答項目）が生じ，通常の分析は不可能となる
- 選択回答は分析に不適切な回答様式であるとされてきた
  - 最近の分析技法の発展で分析可能となった

61

簡便法

- 個別特性項目Y_k に回答したデータを選択し$(Y_1,\ldots,Y_6,Y_k)$を因子分析する
  - 7変数，3因子の因子分析
  - 完全データ
- 上記を88回繰り返す(k=7,8,…,94)

→ 解釈できる分析結果とはならなかった

62

分析モデル

- 想定分布
  - 多変量正規分布
- 欠測にMARを仮定
- 尤度

$$\boldsymbol{Y}_{[i]} = [Y_1,\ldots,Y_6,Y_{i_1},\ldots,Y_{i_4}]^T$$
$$E(\boldsymbol{Y}_{[i]}) = \boldsymbol{\mu}_{[i]}, \quad \mathrm{Var}(\boldsymbol{Y}_{[i]}) = \Sigma_{[i]}$$
$$L(\boldsymbol{\mu},\Sigma) = \prod_{i=1}^{n} N_{10}(\boldsymbol{\mu}_{[i]},\Sigma_{[i]})$$

- （探索的）因子分析
  - 評価項目の潜在構造を同定する

63

標本と分析結果

- 標本
  - インターネット調査
  - n=8,544
- 3個の因子を抽出（評価項目数：94）
  - 第1因子：人　柄（評価項目数：27）
  - 第2因子：知　性（評価項目数：14）
  - 第3因子：積極性（評価項目数：19）
- 上記因子と直接関係しないと判断した項目
  - 94－(27＋14＋19)＝34

64

第１因子：人 柄 (評価項目数: 27)

項目番号	評価項目	第1因子（人柄）	第2因子（知性）	第3因子（積極性）
66	あたたかい―つめたい	0.81	0.09	−0.04
71	やわらかい―かたい	0.81	−0.08	0.00
2	親しみやすい―親しみにくい	0.80	−0.02	0.15
41	素直な―強情な	0.73	0.10	−0.16
74	のんびりした―せっかちな	0.72	−0.06	−0.34
91	安らげる―いらいらする	0.72	0.15	−0.07
1	感じのよい―感じの悪い	0.71	0.16	0.06
14	カジュアルな―フォーマルな	0.69	−0.22	0.17
65	思いやりがある―思いやりがない	0.68	0.25	−0.14
61	穏やかな―激しい	0.68	0.24	−0.40

65

第2因子：知性（評価項目数: 14）

項目番号	評価項目	第1因子（人柄）	第2因子（知性）	第3因子（積極性）
62	理知的な―感覚的な	0.00	0.64	0.04
35	責任感の強い―無責任な	0.17	0.62	0.03
31	きちんとしている―だらしない	0.19	0.61	0.04
63	大人っぽい―子供っぽい	0.06	0.58	0.05
46	いい加減な―几帳面な	−0.07	−0.55	0.18
68	冷静な―情熱的な	0.08	0.55	−0.23
3	慎重な―軽率な	0.19	0.53	−0.17
4	分別のある―分別のない	0.31	0.52	−0.02
33	真面目な―不真面目な	0.40	0.52	−0.13
29	安定した―不安定な	0.14	0.52	0.10

66

第３因子：積極性(評価項目数: 19)

項目番号	評価項目	第1因子（人柄）	第2因子（知性）	第3因子（積極性）
5	積極的な―消極的な	−0.17	0.10	0.78
50	外向的な―内向的な	−0.04	−0.03	0.77
25	大胆な―小心な	−0.02	−0.06	0.73
49	にぎやかな―静かな	0.11	−0.14	0.72
6	自信のある―自信のない	−0.22	0.21	0.72
51	おしゃべりな―無口な	0.08	−0.16	0.71
44	はっきりした―ぼんやりした	−0.19	0.24	0.66
20	元気な―病弱な	0.15	−0.03	0.64
45	新しい―古い	0.02	−0.03	0.57
78	個性的な―無個性な	0.02	0.10	0.56

67

分析結果のまとめ

- 第一印象は3個の要素に分類できる
  - 人柄(27)，知性(14)，積極性(19)
- 対応する評価項目を同定した

- 評価項目確定
  - 実際に選択された割合（使用頻度）
  - コンテキストに合う項目

  を勘案し最終決定する
- 次のステージへ進む

68

結論と課題

- 大量欠測(90%)データに対してMARに基づくFIMLは，予想以上に機能した
  - ポイント
    - FIML with MAR は強力であった
    - 共通項目を設けたこと
      - 各想定因子に2つずつ
      - できれば3つ以上が望ましかった
    - 大標本であること
      - n= 8,542

69

結論と課題

- MARの仮定について
  - MARの仮定はやや強いものである
  - MARの無批判な適用は慎むべきである
  - 他の簡便な方法よりベター
    - より強い仮定(MCAR)が必要
  - データ採取不可能よりベター
- 因子分析モデルを適用する状況では，MARは近似的に成立していると期待できる
  - 似た項目が複数個存在する

70

結論と課題

- 評価項目の選択（非選択）はMCARと考えてもよい？
  - 第一印象の測定データの測定手順
    - 評価項目の選択 → 人物刺激→ 評価実施
    - 評価項目の選択は人物刺激によらない

  - 選択のモデルを検討する必要がある
    - Discrete choice model (method)

71

結論と課題

- この方法が妥当であることを論証するには
  - 検討事項
    - 必要なサンプルサイズ
    - 共通項目の有効性
    - 欠測がNMARのときにFIMLがどの程度機能するか

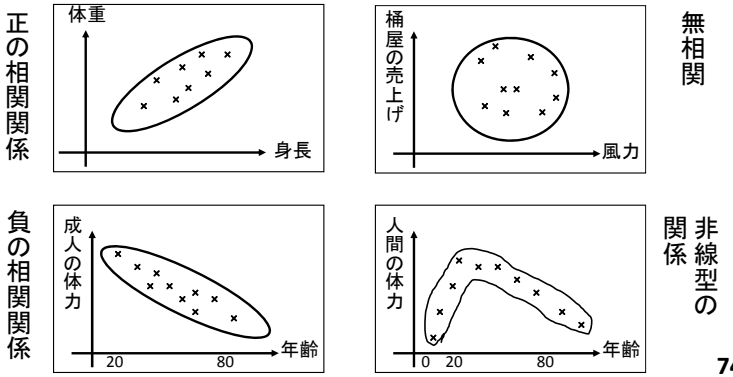
72

MARでない欠測
FIMLが機能せず
研究途上

分析事例3：相関分析における欠測

講義ノートより：相関関係とは

- 2つの事柄の間に確定的ではないが何らかの関係（傾向）が見出されることがある



74

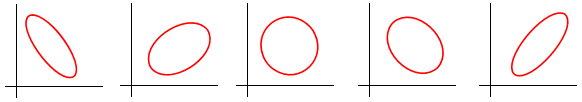
講義ノートより：相関関係と関数関係

- 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
  - 一般に，yの分布はxと関係する

75

講義ノートより：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



77

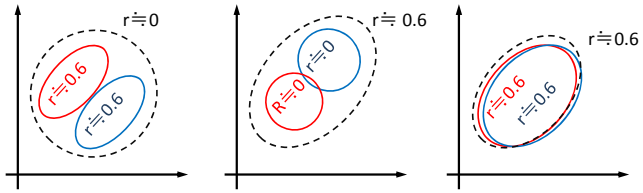
講義ノートより：ピアソンの標本相関係数を用いる時の注意

- 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
- 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

78

講義ノートより：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



79

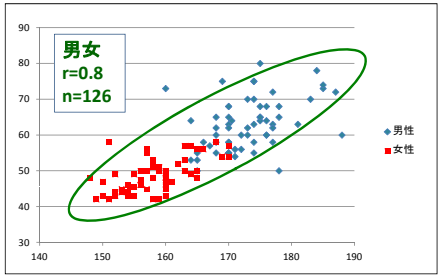
相関分析：身長と体重

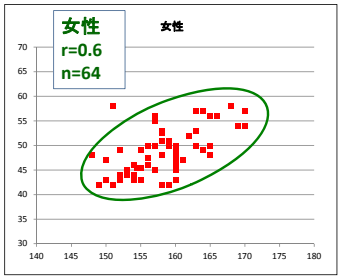
80

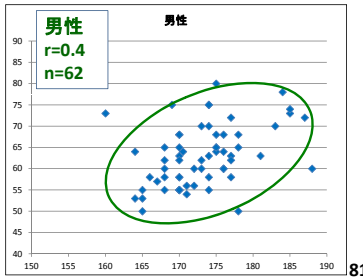
- 調査時期：1998.4
- 調査対象：O大学H学部1年次生（クラス内調査，欠席なし）

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)



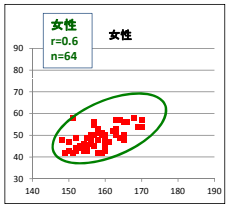


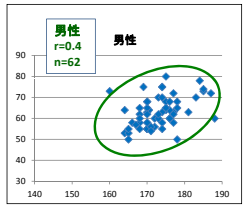


81

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
    - 対象：O大学H学部1年次生
  - 実測でない





82

考察：女性の相関が少し高い？

- 女性の回答に偏りの可能性
  - 理想体型により近い値を回答
- 男性は記憶があいまい
  - 体型に興味がない
  - 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？

83

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

84

回答拒否とFIML

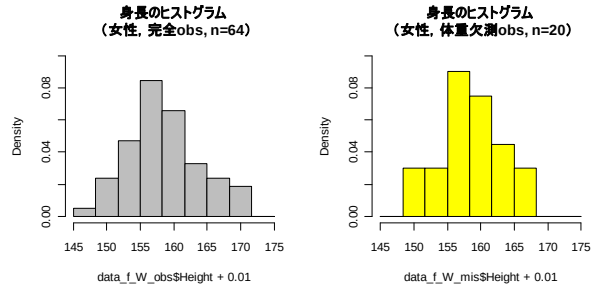
- 男性（n=62)
  - 拒否なし
- 女性（n=93）
  - 完全データ：64
  - 欠測データ：29
    - 身長と体重欠測： 9
    - 体重のみ欠測： 20
    - 身長のみ欠測： 0

- CCA(LD)による相関分析
  - 男性：r=0.414 (n=62)
  - 女性：r=0.600 (n=64)
- FIML(MAR)による相関分析
  - AMOS，Mplus
  - 男性：r=0.414 (n=62)
  - 女性：r=0.585 (n=64+20)

  → うまくいかない！

85

身長のデータと欠測

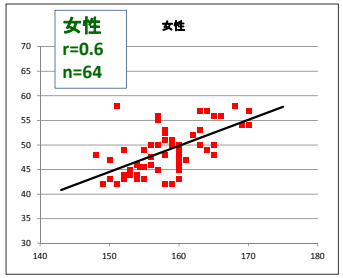


- 欠測・観測は身長と関係なさそう
- 欠測はNMARが疑われる
  → FIMLが機能しない理由

86

NMARの下での推測と課題

- 欠測メカニズムをモデル化して尤度に組み込む必要がある
  - 標準体重からずれるほど欠測確率が増大するモデル？



87

NMARの下での推測と課題

- 身長と体重ともに欠測するのはどのような場合か検討する
- 身長は正規分布と考えてよいが，体重は？

- 研究継続中．．．

88

まとめ

まとめ

- 欠測値を適切に処理しない分析は，推定結果に無視できないバイアスが生じる
- 今までは，重要性に比して，その扱いはズサンであった
- 応用研究において欠測値の扱いと報告の方法は近々ドラスティックに変化する可能性がある
  - APAの方針
  - 査読のプロセスで、この論文を根拠に分析のしなおしを要求することが起こるかどうか
  - LDやPDは避けるべき
    - 欠測値がごく少数の場合はOK
- まず，欠測メカニズムに思いを馳せることが重要
  - MCAR, MAR, NMAR

まとめ

- 現状では，MARを仮定した下でのFIMLが基本
  - FIMLは，NMARの場合でも従来法(LD, PD,代入法)よりベターだろう
  - NMARのとき欠測メカニズムのモデリングが必要になる場合もあるだろう
- NMARに対するFIML（MAR)の適用可能性について，理論研究が必要
- 欠測データ解析に関する統計的理論研究も発展させる必要がある
  - 欠測メカニズムは様々であり，実質科学の研究者とのコラボレーションが求められる

文献

- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Little and Rubin (2002). Statistical Analysis with Missing Data 2nd Ed. Wiley.
- Kano and Takai (2011). Analysis of NMAR missing data without specifying missing-data mechanisms in a linear latent variate model. Journal of Multivariate Analysis. 102, 1241-1255.
- 岩崎（2002）．不完全データの統計解析．サイエンティスト社．
- 狩野 (2009). 不完全データの解析．生産と技術．61(1)，生産技術振興協会．71-76.
- 金, 廣瀬, 今田, 吉田, 松尾, 藤井 (2012/5)．個人属性が対人認知構造に及ぼす影響について～Webアンケートによる大規模調 査の解析結果から～．電子情報通信学会技術研究報告, HCS, vol.112, no.45, pp.97-102.
- 廣瀬，金，狩野，吉田，今田，松尾 (2012/8) 初対面の第一印象を「プロデュース」する～L1型正則化法と因子分析 の新たな応用．2012年度統計関連学会連合大会．北海道大学．

- Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.
- Peugh, J.L. and Enders, C.K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525-556.
- Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1-26.
- Schafer, J.L., & Graham, J.W. (2002) Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Wilkinson, L. and APA Task Force on Statistical Inference APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals: Guidelines and Explanation. *American Psychologist*, 54, 594-604.

93

ご清聴有難うございました

94

受講生募集「データ科学特論I」

本講義では，実質科学においてデータ解析を研究の道具として実際に使う学生（＋研究者）を対象に，具体的な分析の手順・方法とそれらの数理的基礎，そしてそれらを適用する際の注意事項等を講述します．講義に出席し必要な課題を提出し合格した受講生へは単位認定を行います．

科目名称：データ科学特論I　　単位数：2
受講対象：全国の大学院生 他
担当教員：狩野 裕，内田雅之（世話教員）
開講学期：第1学期（2013年8月26日(月)～8月31日(土)　集中）予定
会　　場：大阪大学 基礎工学研究科（豊中キャンパス）
受講要件：学部1年次レベルの統計学を履修または自習した者
単位認定：出席とクラス内活動，レポート課題により総合評価
参　　考：　文科省大学間連携共同教育推進事業

95

講義計画

第1日
　講義題目：　データ科学の基礎（2コマ）
　講義担当者：　狩野　裕（大阪大学 基礎工学研究科）
　講義題目：　データ解析環境Rの基礎（1コマ）
　講義担当者：　熊谷悦生（大阪大学 基礎工学研究科）
第2日
　講義題目：　一般化線形（混合）モデル（3コマ）
　講義担当者：　久保拓弥（北海道大学 地球環境科学研究院）
第3日
　講義題目：　ノンパラメトリック回帰（3コマ）
　講義担当者：　坂本　亘（大阪大学 基礎工学研究科）
第4日
　講義題目：　判別分析，クラスター分析，MDS（3コマ）
　講義担当者：　今泉　忠（多摩大学 大学院経営情報学研究科）
第5日
　講義題目：　正定値カーネルによるデータ解析（3コマ）
　講義担当者：　金森敬文（名古屋大学 情報科学研究科）

96

講義内容

第１日　データ科学の基礎・データ解析環境Rの基礎（狩野裕＋熊谷悦生）

科学的な研究において何故データを分析する必要があるのか，データ分析で何が分かるのか，何故確率分布を考える必要があるのか等，データ科学の根源的な内容を解説する．統計的推測に関する知識を整理すると共に正確な理解を促す．また，確率的な現象の意味，推測の過誤，検出力とサンプルサイズの設計，そして，適切な統計分析のための３つのS (statistical significance, practical significance, theoretical significance)等についても講述する．

データ解析環境Rはフリーのソフトであり，実際のデータ解析に用いられるだけでなく，統計手法や確率的現象の理解を助けるためにも用いられる．本講義全体においてRが利用されるため，基本的な利用方法を解説する．

97