Handling Incomplete Data in Longitudinal Surveys

Joop Hox

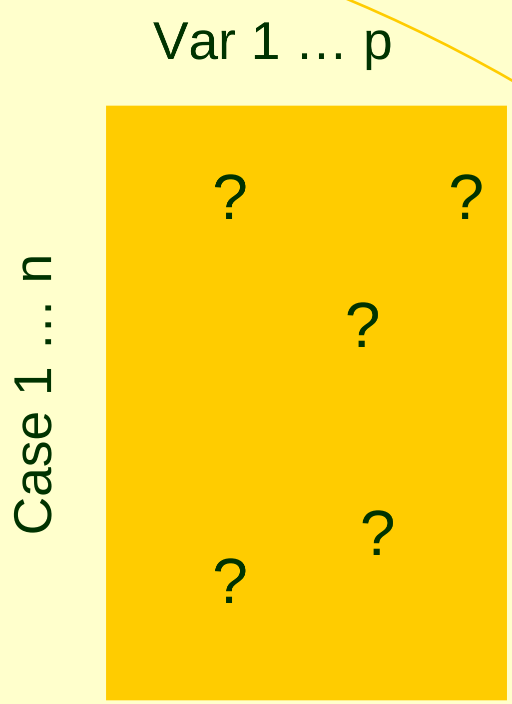
Edith de Leeuw

University of Essex

Universiteit Utrecht

Methodology of Longitudinal Surveys (MOLS) Short Course July 2006

Part I: Missing Data Introduction



Edith de Leeuw

Universiteit Utrecht

University of Essex

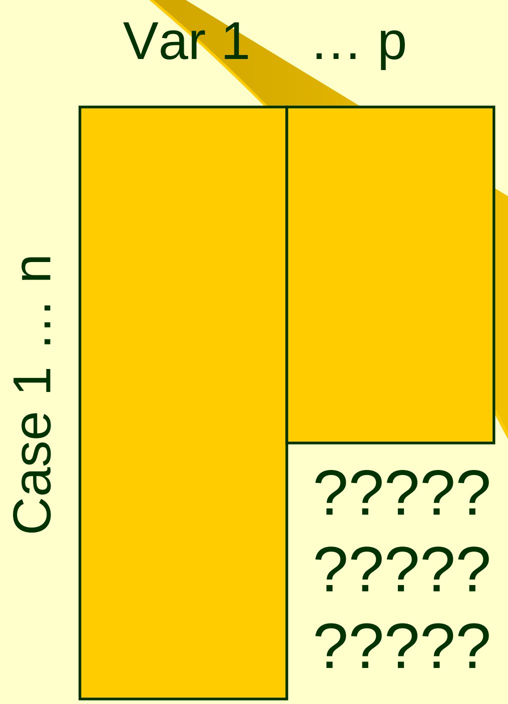
Mols 2006

Terminology

- Unit nonresponse
  - Failure to obtain ***any*** information from an eligible sample unit
    - Business, household, person
- Item nonresponse
  - Aka ‘missing data’
  - Unit participates
  - Failure to obtain ***information*** for one or more questions, given that the other questions are completed

Pattern Unit Nonresponse

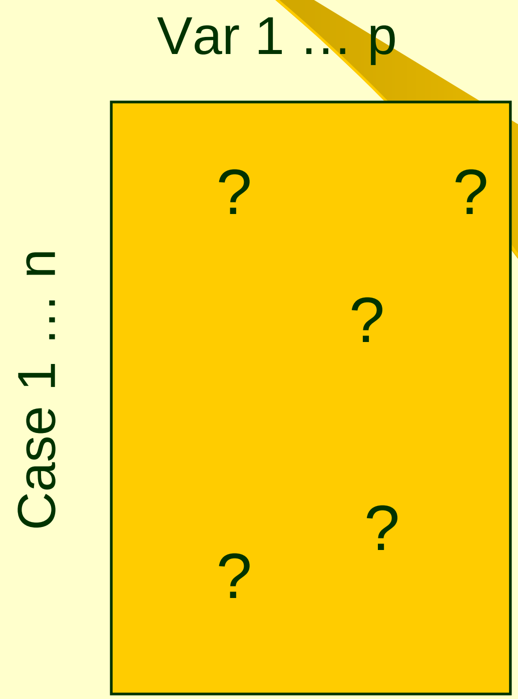
- Unit Nonresponse: All variables missing for some cases
  – But we may have some background variables
- **?** = missing
- Example: nonresponse in surveys
- Example: double sampling designs



(from: Little & Rubin, 1987, p57)

Item-Nonresponse Pattern

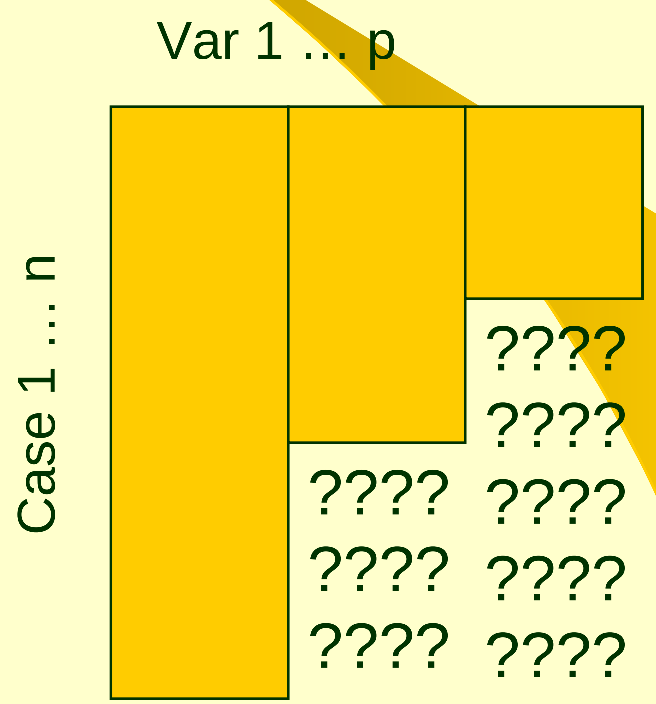
- General pattern: various variables missing
- **?** = missing



Special Nonresponse Pattern: Monotone Missing

- Monotone missing: Blocks of missing variables
- Monotonically increasing number of missings
- **?** = missing
- Prime Example:

  Panel Attrition



Many Manifestations of Missing Data in LS: Time 0

- Longitudinal Studies
  – Measurement over time
  – More than 1 measurement occasion or wave
- First Manifestation of Missing Data in L.S.
  – Time 0: Initial Recruitment or Panel Formation
    - Unit Nonresponse
      – Non-contact
      – Refusal
      – Others
    - Item Nonresponse
      – Do-not-know
      – Refusal
      – Others

Many Manifestations Missing Data in LS: Time 1,..,p

- Next Manifestation of Missing Data
  - Time 1, 2, 3,...
    - Unit Nonresponse
      - Drop-out or wave nonresponse: Participant in study does not produce a completed questionnaire or interview at a specific time point
      - Attrition or panel mortality: Participant stops to respond to all subsequent questionnaires or interviews
    - Item Nonresponse
      - Topic this course

Suggested Readings

- De Leeuw, Edith (2005), Dropout in Longitudinal Surveys: Strategies to limit the problem (course pack). A later version of this paper appeared in B. S. Everitt and D. C. Howell (Eds). Encyclopedia of Statistics in Behavioral Science, 2005. Volume 1, pp.515-518. Chichester: Wiley.
- Hox, Joop and De Leeuw, Edith (1999). Handling of Incomplete Multivariate Data, Glossary of Important Terms, K.M, 20, 62, 139-140 (course pack)

Handling Incomplete Data in Longitudinal Surveys

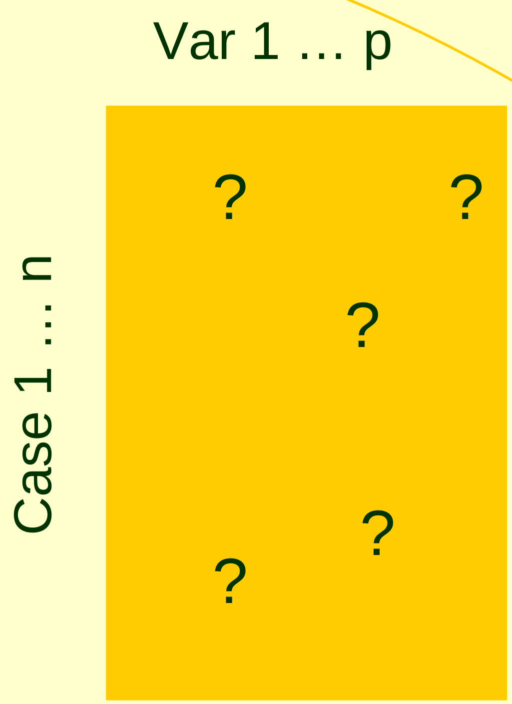
Joop Hox

Edith de Leeuw

University of Essex

Universiteit Utrecht

Part II: Diagnosing Missing Data



Edith de Leeuw

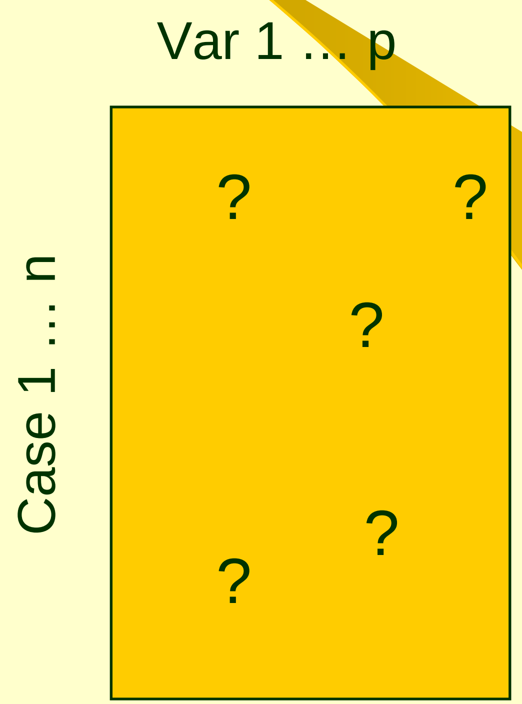
Universiteit Utrecht

University of Essex

Mols 2006

Item-Nonresponse Pattern

- General pattern: various variables missing
- **?** = missing



Why a Problem?

- Gaps in data matrix
- Loss of information
- Bad image (quality criterion)

- Ignoring *(deletion of missing cases)* has problems:
  - Analyses are performed on different (sub) data sets
  - Analyses can be inconsistent with each other
  - Difficult to present results consistently over analyses
  - Potential for bias
  - Strong assumption (MCAR)

Why a Problem continued

- Potential for *biased* results
  - Univariate analysis and (general) low item-nonresponse: bias is generally small
  - Multivariate analysis, even with low item nonresponse for each question, cumulates to a substantial proportion of records that are missing

- So: do something

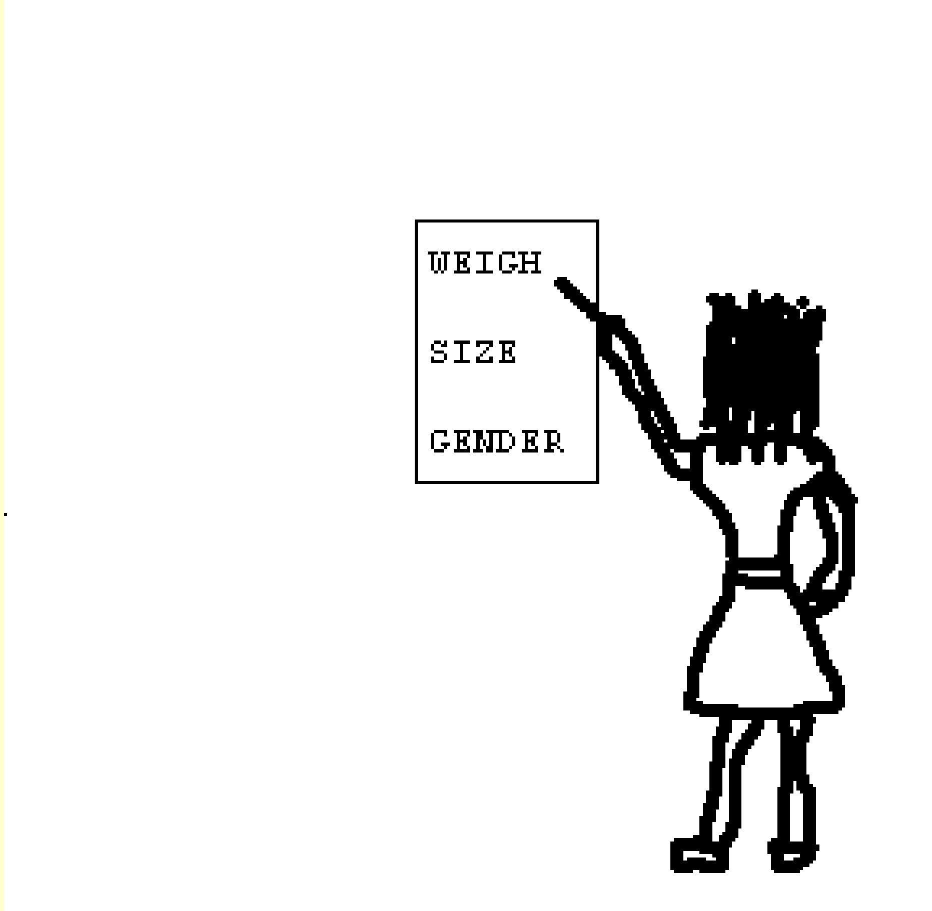
- Simply ignoring (standard option in SPSS and other packages) ***not wise***

Important Distinctions

- Missing Completely At Random (MCAR)
  - Missing values random sample of all values
- Missing At Random (MAR)
  - Missing values random sample of all values within classes defined by covariates (conditional)
- Not Missing At Random (NMAR)
  - Missingness is related to unobserved (missing) value

(Little & Rubin, 1987, p14)

A Silly Example



Hoytink, 2004

Illustration: Survey Research

- Interviewer overlooks a question by accident
  - Turns two pages in one
- Elderly person has difficulty remembering event
- Participant refuses to answer

Sources Item-Nonresponse

- Researcher (by design)
- Interviewer
- Respondent
- Questionnaire
- Method of Data Collection

- Interaction between sources, e.g, respondent and questionnaire

What Can Be Done

- Missing by Design
  - Special analyses (e.g., multi-level analysis)
- Partial Non-Response (e.g., break-of)
  - Prevent
  - Adjust:
    - Delete cases and treat as unit-nonresponse (weighting)
    - Keep cases and impute missing answers
- Item Non-Response
  - Prevent (see extra slides at end + De Leeuw et al 2003)
  - Adjust (impute!, see lecture this afternoon)

What is Known

- Respondents: Age and Education
- Interviewer: Training and Supervision
- Topic: Sensitive Questions
- Questionnaire: Lay-out, Do-not-know category, Number of categories, graphical tools
- Mode: SAQ, CAI

Prevent and then Adjust: Why Adjust?

- Remember: respondent age and education consistently correlate with item-nonresponse:
  - **NOT MCAR**, So standard solution (pairwise/listwise) not adequate
  - Use age & education in adjustment model
- Impute missing data to get a complete data -set
  - All analyses are on ONE data-set
  - Consistent with each other
  - Retain all data

Two Phases In Adjustment

- Phase I: Diagnosis:
  - Think about Missing data (why/how)
  - Inspect Patterns of Missingness
    - Suggest processes
    - Suggest solutions
- Phase II: Cure, Adjust for Missing
  - Use what you know from phase 1
  - Use any available information you have
    - Plan for nonresponse

Patterns of Item Nonresponse

- Various variables missing (Missing = **?**)

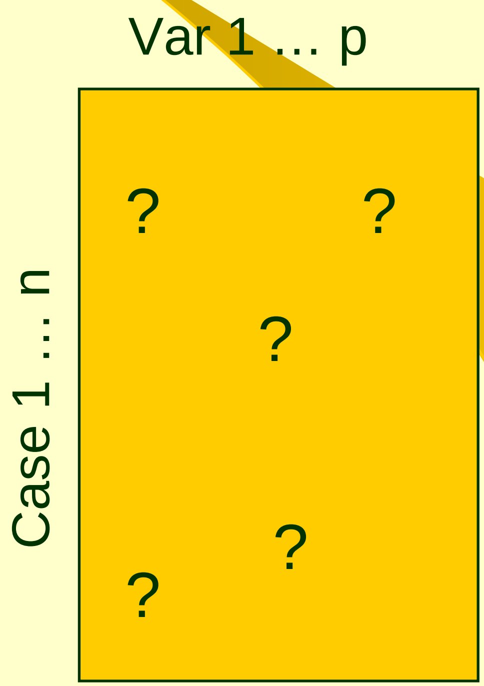
? Unsystematic (MCAR)

or

? MAR

or

? NMAR



Tools for Exploring Missingness

- Descriptive statistics
- Graphical representations
  - View data matrix on screen/Special plots
- Statistical tests
  - Usual tests for MCAR
- Software:
  - SPSS-MVA module
  - Dedicated Programs for Missing Data
    - e.g. SOLAS, NORM
  - Some DIY-tricks using SPSS or any other program

Example Data File Longmis

- Longitudinal data with 5 time points

- Explanatory variable: Sex
- 40 Cases

- Panel attrition
- Incidental missings
- No missings on sex

Longmis Example Data

1	1	49	49	50	58	60
2	1	44	51	(46)	(49)	(48)
3	0	56	53	57	55	52
4	0	57	(52)	58	57	56
5	1	54	55	59	53	(54)
6	0	46	44	44	51	55
7	0	53	53	53	53	57
8	0	44	(52)	(53)	(53)	(54)
9	0	53	54	55	55	56
10	1	53	56	55	52	53
11	0	56	56	56	54	57
12	1	57	55	58	60	59
13	0	54	58	59	61	62
14	0	44	42	(44)	48	47
15	1	56	65	59	63	64
16	1	46	50	50	49	48
17	0	45	50	(50)	55	54
18	1	(66)	61	63	70	71
19	1	50	(48)	(52)	(56)	(53)
20	1	49	(45)	(56)	(48)	(55)
21	1	53	58	60	(59)	58
22	1	48	45	45	50	(51)
23	1	54	52	53	54	55
24	0	50	50	47	(50)	(54)
25	0	50	47	(48)	(49)	(46)
26	0	47	50	54	53	55
27	1	52	53	58	57	58
28	0	45	(45)	(53)	(53)	(55)
29	1	56	57	55	58	61
30	0	47	(46)	(51)	(51)	(50)
31	1	49	49	(49)	(52)	(53)
32	1	46	52	55	52	(54)
33	1	49	51	(55)	(49)	(57)
34	0	59	58	58	61	59
35	0	53	57	50	(55)	55
36	0	40	46	(45)	(47)	(48)
37	1	47	49	47	53	54
38	1	49	52	49	(51)	(52)
39	0	45	50	48	47	(47)
40	1	50	46	45	44	45

Variables

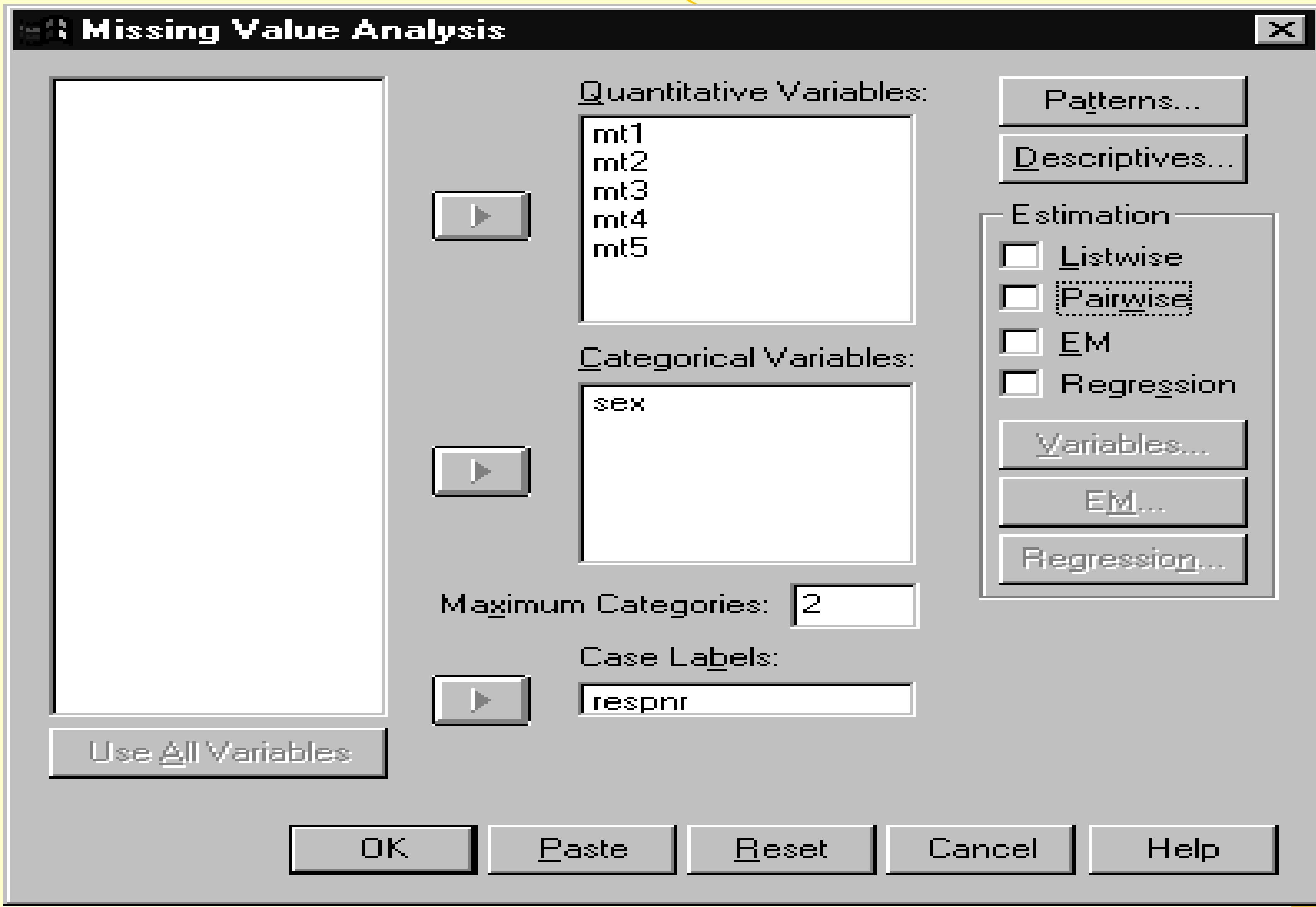
respnr
sex
time1
time2
time3
time4
time5

( ) = missing

SPSS Missing Values Analysis (MVA)

- MVA Patterns
  - Displays missings by pattern
- MVA Descriptives
  - Univariate descriptives
  - MVA Tests (Ho=MCAR)
    - *t*-test for MCAR
    - crosstabs

SPSS MVA



Missing Value Analysis
Quantitative Variables:
mt1
mt2
mt3
mt4
mt5
Categorical Variables:
sex
Maximum Categories:
2
Case Labels:
respnr
Use All Variables
Patterns...
Descriptives...
Estimation
Listwise
Pairwise
EM
Regression
Variables...
EM...
Regression...
OK
Paste
Reset
Cancel
Help

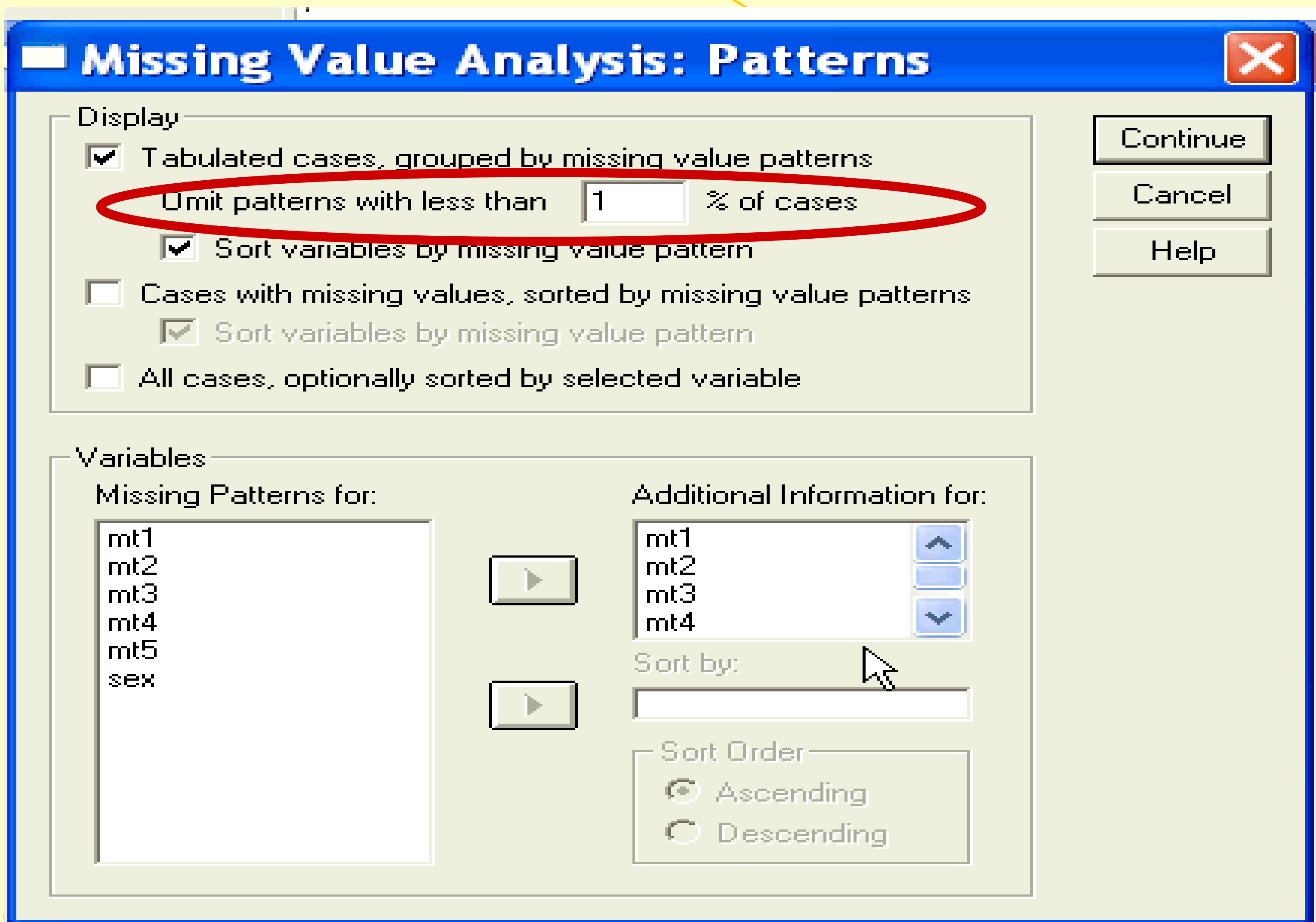
SPSS MVA Display: Patterns

- **Display Tabulated Cases Grouped by Missing Pattern for all cases**
  - **Additional Info**

- Display Individual Cases with Missings Sorted by Missing Value Pattern

- Display Cases Sorted by Variable
  - Example variable sex

SPSS MVA Patterns
Three Choices



Display *Table* by Pattern for All

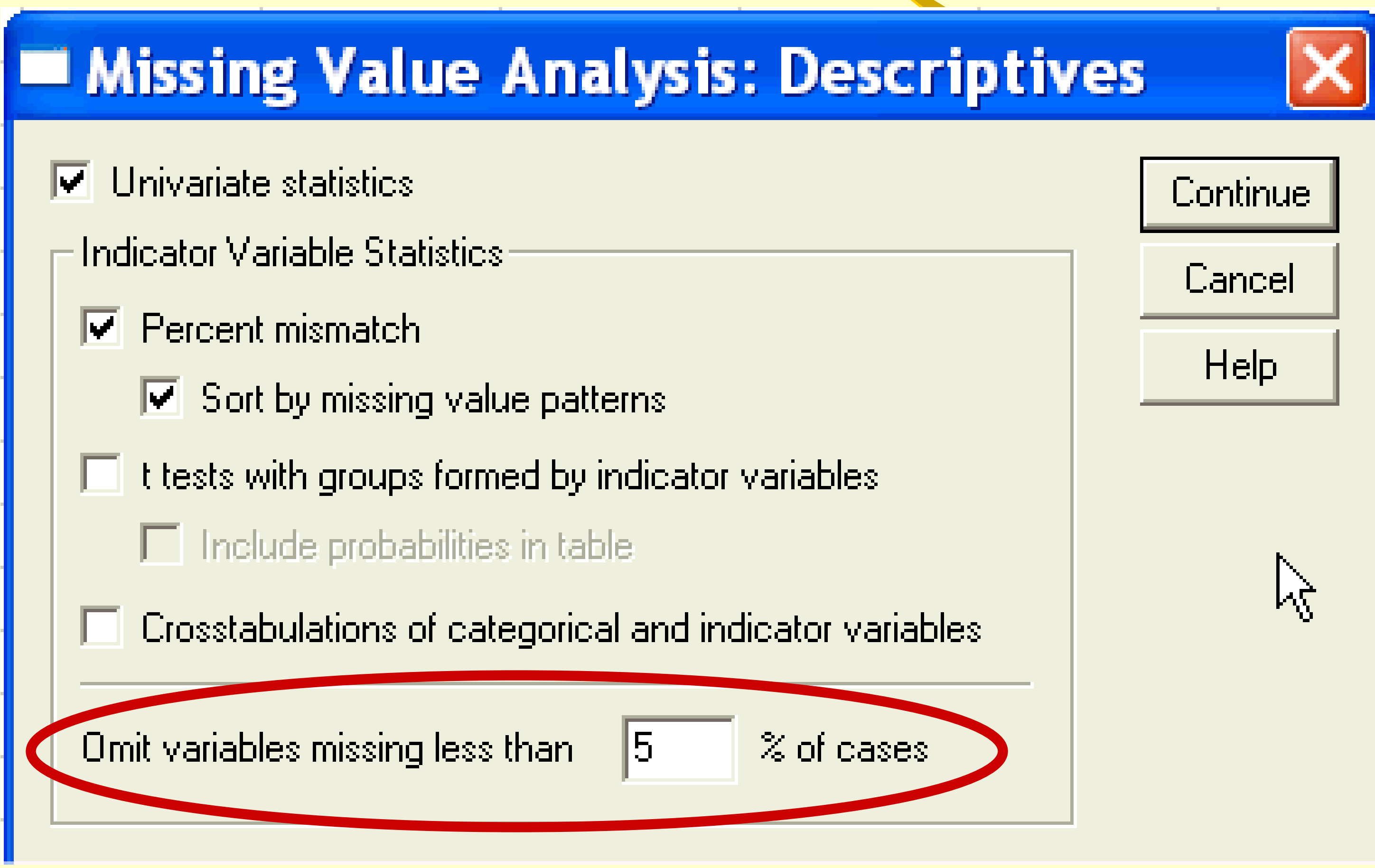
Tabulated Patterns

Number of Cases	Missing Patterns[a] SEX	MT1	MT2	MT3	MT4	MT5	Complete if ...[b]	MT1[c]	MT2[c]	MT3[c]	MT4[c]	MT5[c]
17							17	52.24	53.24	53.47	55.06	56.35
5						X	22	49.80	51.00	52.80	51.40	.
2					X	X	26	49.50	51.00	48.00	.	.
2					X		19	53.00	57.50	55.00	.	56.50
1		X					18	.	61.00	63.00	70.00	71.00
1			X				18	57.00	.	58.00	57.00	56.00
2				X			19	44.50	46.00	.	51.50	50.50
5				X	X	X	33	46.40	48.80	.	.	.
5			X	X	X	X	39	47.00	.	.	.	.

SPSS MVA : Descriptives

- Univariate Statistics
- Pairwise Mismatch
- **Patterns: t-test with indicator variables (missingness indicator)**
- **Patterns: Crosstabulations**
  - **categorical var & indicator var**

SPSS MVA Descriptives
Four Choices



MVA Descriptives 1

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
MT1	39	50.13	4.55	1	2.5	0	0
MT2	34	52.18	4.99	6	15.0	0	0
MT3	28	53.57	5.23	12	30.0	0	0
MT4	26	54.73	5.50	14	35.0	0	1
MT5	23	56.48	5.54	17	42.5	1	1
SEX	40			0	.0		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 +

MVA Descriptives 2

Percent Mismatch of Indicator Variables. [a,b]

	MT2	MT3	MT4	MT5
MT2	15.00			
MT3	20.00	30.00		
MT4	25.00	15.00	35.00	
MT5	32.50	22.50	17.50	42.50

The diagonal elements are the percentages missing, and the off-diagonal elements are the mismatch percentages of indicator variables.

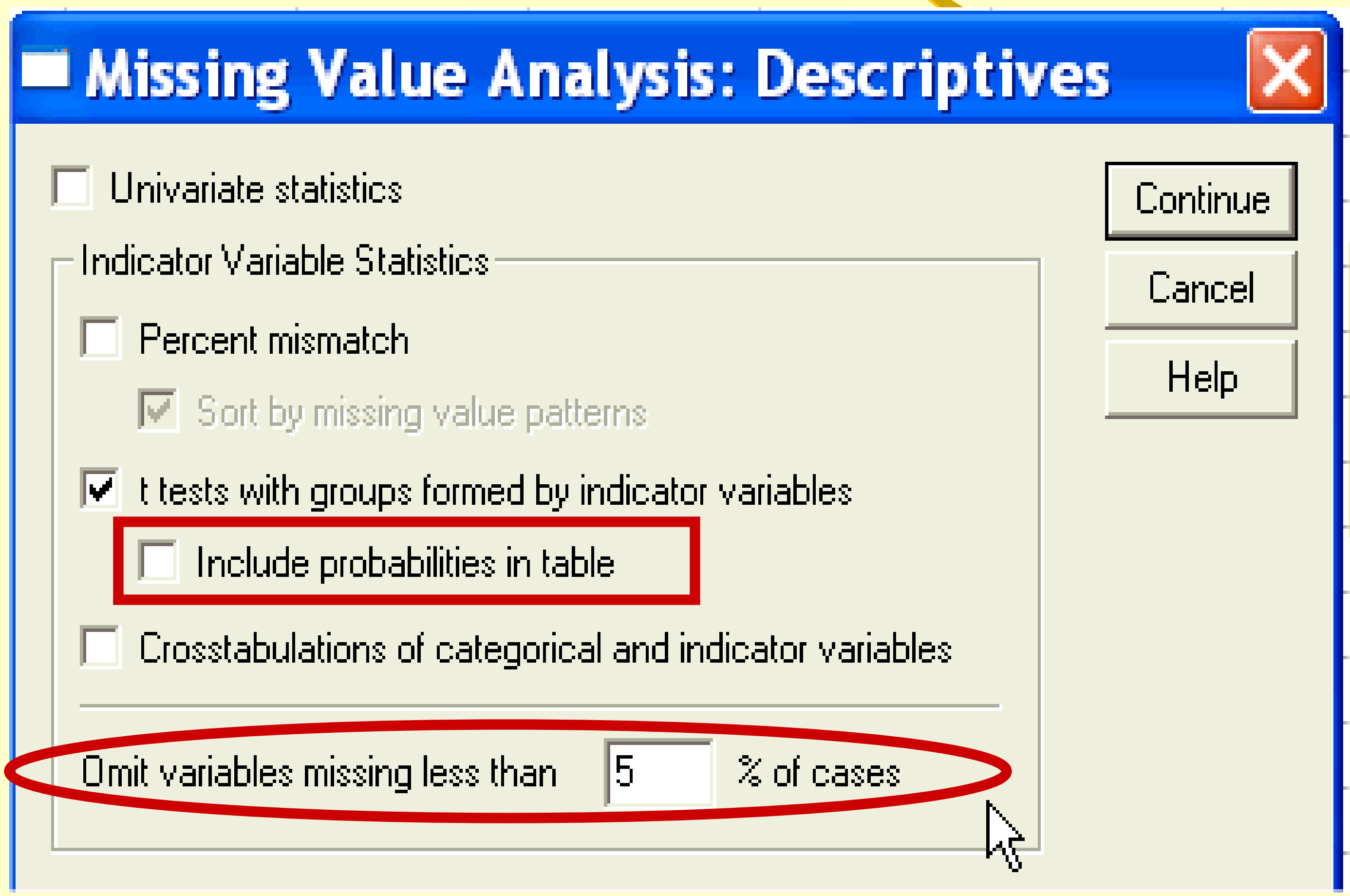
a. Variables are sorted on missing patterns.

b. Indicator variables with less than 5% missing values are not displayed.

MVA tests

- *t*-test for MCAR (Ho: MCAR)
- What does MVA Descriptives do?
  - For each variable with missing values, indicator variables coded as *present* vs. *missing*
  - Performs t-test to compare these groups on other variables
  - Default no p-values
  - Default omit vars less than 5% missing

SPSS MVA Descriptives
t-test for MCAR



MVA t-test for MCAR

Separate Variance t Tests

		MT1	MT2	MT3	MT4	MT5
MT1	t	.	.	.	.	.
	df	.	.	.	.	.
	P(2-tail)	.	.	.	.	.
	# Present	39	33	27	25	22
	# Missing	0	1	1	1	1
	Mean(Present)	50.13	51.91	53.22	54.12	55.82
	Mean(Missing)	.	.	.	.	.
MT2	t	.8	.	.	.	.
	df	6.8	.	.	.	.
	P(2-tail)	.432	.	.	.	.
	# Present	33	34	27	25	22
	# Missing	6	0	1	1	1
	Mean(Present)	50.39	52.18	53.41	54.64	56.50
	Mean(Missing)	48.67	.	.	.	.
MT3	t	4.6	3.4	.	1.0	1.8
	df	27.1	13.7	.	1.2	1.2
	P(2-tail)	.000	.004	.	.492	.289
	# Present	27	27	28	24	21
	# Missing	12	7	0	2	2
	Mean(Present)	51.81	53.26	53.57	55.00	57.05
	Mean(Missing)	46.33	48.00	.	51.50	50.50

MVA tests 2

- *Cross-tabulation* Ho: MCAR
- What does MVA Descriptives do?
  - For each variable with missing values, indicator variables coded as *present* vs. *missing*
  - Gives a crosstabulation of categorical variables with indicator variables (missingness indicators)
  - No formal chi-square test, no p-values
  - Default omit vars with less than 5% missing

Crosstabulations Sex and Missingness Indicators

			Total	male	female
MT1	Present	Count	39	19	20
		Percent	97.5	100.0	95.2
	Missing	% 99	2.5	.0	4.8
MT2	Present	Count	34	15	19
		Percent	85.0	78.9	90.5
	Missing	% 99	15.0	21.1	9.5
MT3	Present	Count	28	12	16
		Percent	70.0	63.2	76.2
	Missing	% 99	30.0	36.8	23.8
MT4	Present	Count	26	12	14
		Percent	65.0	63.2	66.7
	Missing	% 99	35.0	36.8	33.3

SOLAS

- Missing data analysis and imputation
- Used in bio-medical and pharmaceutical research
- (non)parametric
- Stand-alone program
- But reads SPSS
- And SAS, BMDP, et cetera..

Datasheet : LONGMIS

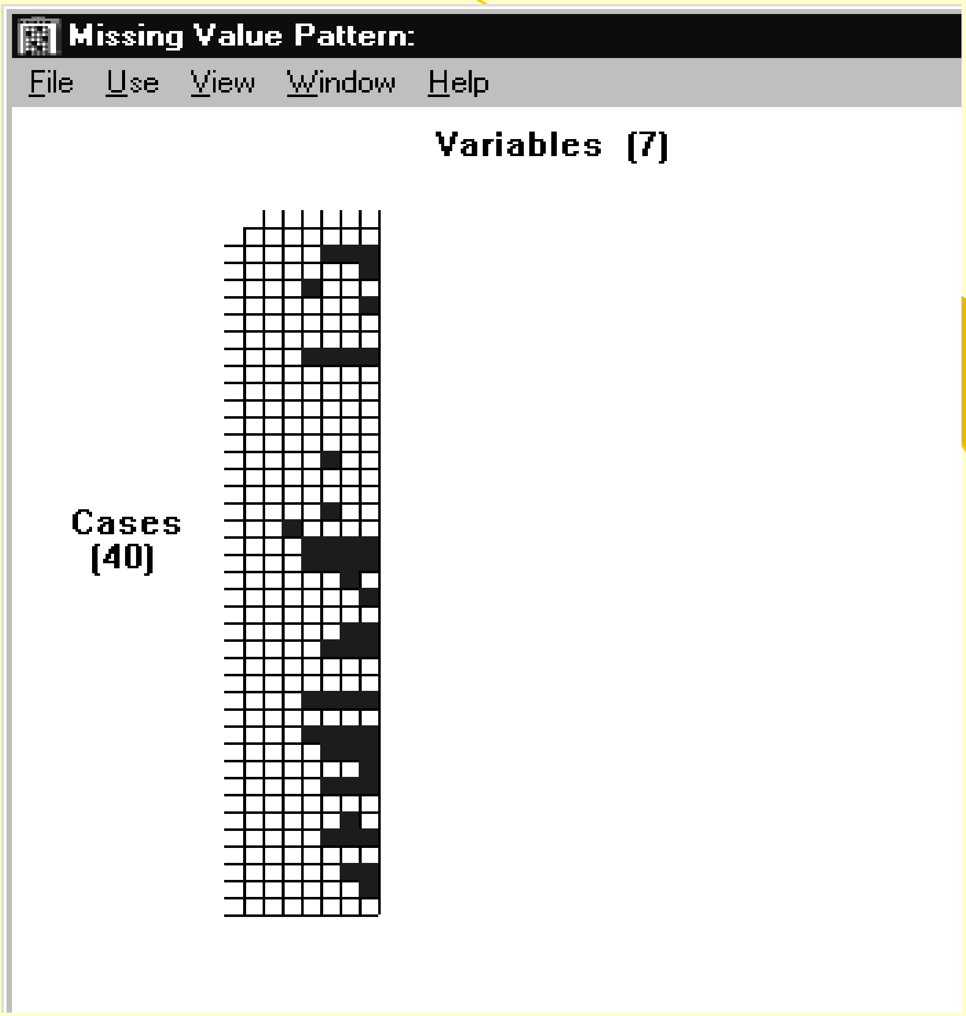
File Edit Variables Use Analyze Plot Format View Window Help

7 vars 40 cases	RESPNR	SEX	MT1	MT2	MT3	MT4	MT5
1	1.000000	1	49	49	50	58	60
2	2	1	44	51			
3	3	0	56	53	57	55	
4	4	0	57		58	57	56
5	5	1	54	55	59	53	
6	6	0	46	44	44	51	55
7	7	0	53	53	53	53	57
8	8	0	44				
9	9	0	53	54	55	55	56
10	10	1	53	56	55	52	53
11	11	0	56	56	56	54	57
12	12	1	57	55	58	60	59
13	13	0	54	58	59	61	62
14	14	0	44	42		48	47
15	15	1	56	65	59	63	64
16	16	1	46	50	50	49	48
17	17	0	45	50		55	54
18	18	1		61	63	70	71
19	19	1	50				

LONGMIS: First Look with SOLAS

SOLAS
Missing Data
Pattern

Variables
respnr
sex
mt1
mt2
mt3
mt4
mt5



Simple Procedures DIY

- **Recode** all variables into **new** variables with values: 1 = missing, 0 = observed
  - These variables are missingness indicators
- Use your *favorite standard* program and do simple tests like SPSS MVA does
  - Descriptives on the recoded variables
  - Cross-tabulation missingness indicator with (substantive) categorical variables
  - T-tests with (substantive) interval variables

DIY-MVA and *MORE*...

- Use **new** variables (missingness indicators)
- Use favorite standard program
- Examples
  - SPSS Explore
  - Graphs
    - Boxplot with missingness indicator on category axis
  - Correlations between missingness indicators
  - PCA
  - Correlations substantive vars with indicators
    - Pairwise deletion! **Why?**
  - …….

Example MSCOHORT.SAV

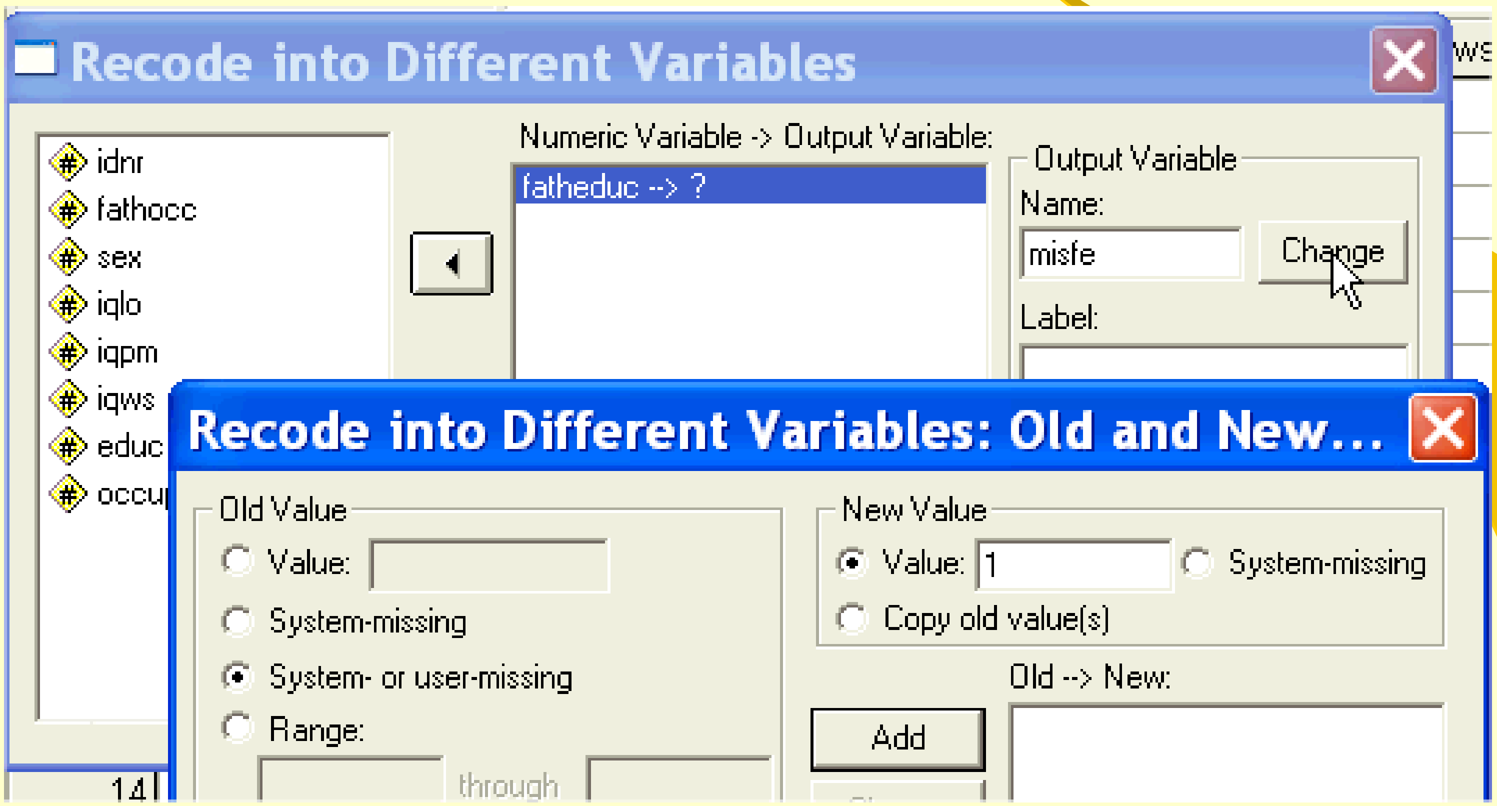
- Data set from educational research
  - Order of variables: idnr, father education (fatheduc), father occupation (fathocc), sex, iqlo, iqpm, iqws, education (educ), occupation (occup)
  - Note 1: iglo,iqpm,iqws are three IQ-tests
  - Note 2: 2 variables measure 'father of pupil' rest of variables measure pupil!
  - Note 3: Missing data are indicated by missing value 999

Step 1: Make Indicator Variables

value 1 if missing, 0 if not!

- RECODE fatheduc (MISSING=1) (ELSE=0) INTO misfe
- RECODE fathocc (MISSING=1) (ELSE=0) INTO misfo
- RECODE sex (MISSING=1) (ELSE=0) INTO missex
- RECODE iqlo (MISSING=1) (ELSE=0) INTO misiqlo
- RECODE iqpm (MISSING=1) (ELSE=0) INTO misiqpm
- RECODE ....

SPSS Recode
Into Different Variable



Step 2: SPSS Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
MISFE	5690	.00	1.00	.4374	.4961
MISFO	5690	.00	1.00	.1065	.3085
MISSEX	5690	.00	1.00	6.854E-03	8.251E-02
MISIQLO	5690	.00	1.00	8.471E-02	.2785
MISIQPM	5690	.00	1.00	.1230	.3285
MISIQWS	5690	.00	1.00	.1253	.3311
MISEDUC	5690	.00	1.00	.5557	.4969
MISOCC	5690	.00	1.00	.5891	.4920
Valid N (listwise)	5690				

Step 3 Test MCAR
How about gender?: Crosstabs

MISOCC * SEX Crosstabulation

			SEX 0	SEX 1	Total
MISOCC	.00	Count	1586	751	2337
		% within MISOCC	67.9%	32.1%	100.0%
		% within SEX	54.0%	27.7%	41.4%
		Adjusted Residual	$\boxed{20.1}$	$\boxed{-20.1}$	
	1.00	Count	1352	1962	3314
		% within MISOCC	40.8%	59.2%	100.0%
		% within SEX	46.0%	72.3%	58.6%
		Adjusted Residual	$\boxed{-20.1}$	$\boxed{20.1}$	
Total		Count	2938	2713	5651
		% within MISOCC	52.0%	48.0%	100.0%
		% within SEX	100.0%	100.0%	100.0%
		Adjusted Residual			

chi^2 : 4003
df = 1
p=.00
Phi= 0.27

1=f
0= m

Misoc=missing on occupation Is this MCAR?

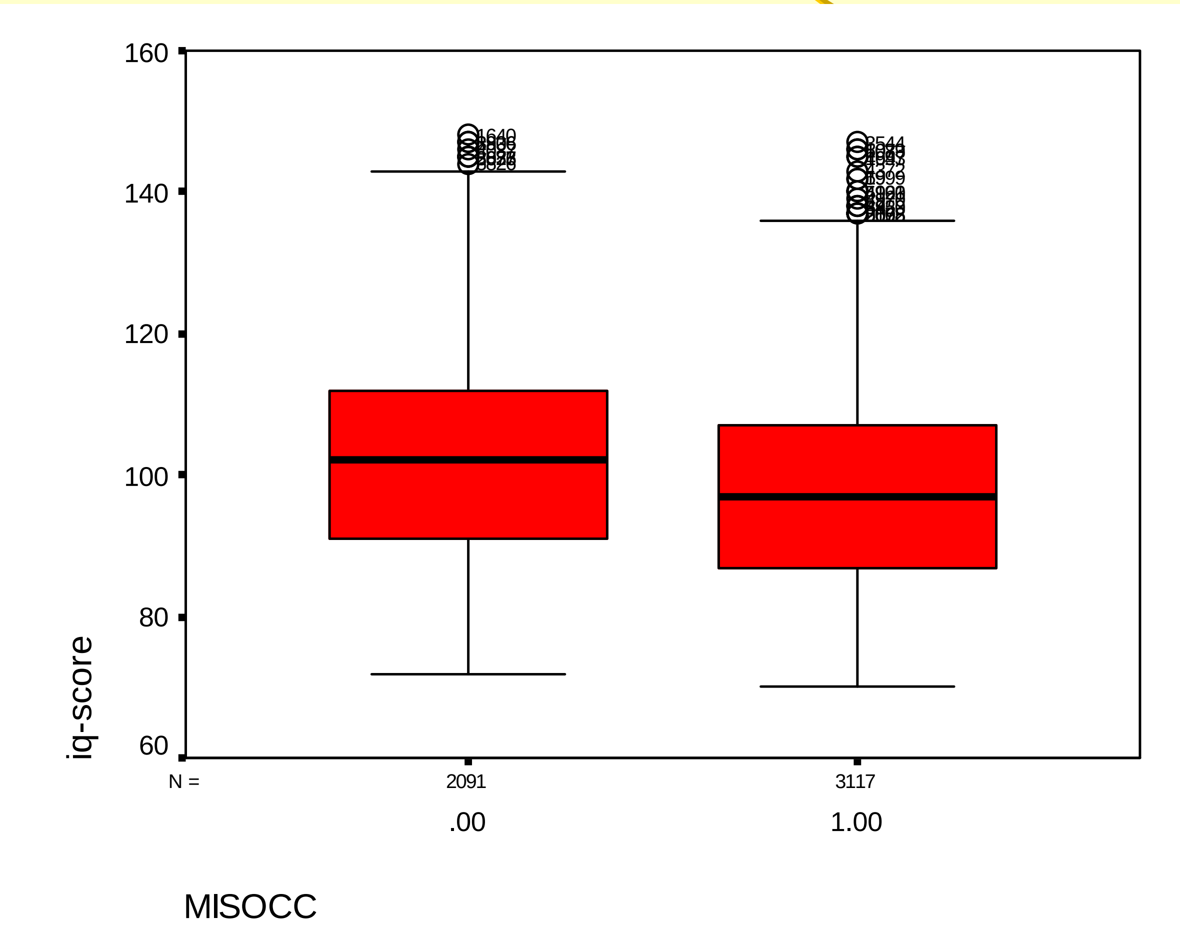
Step 4 Test MCAR continued
How about IQ?: T-test

Group Statistics **(P=.00)**

	MISOCC	N	Mean	Std. Deviation	Std. Error Mean
IQLO	.00	2091	102.21	14.29	.31
	1.00	3117	97.98	13.87	.25

Misoc=missing on occupation Is this MCAR?
1=missing 0= data available

Boxplot of IQ-score grouped by Missingness indicator Occupation



Patterns in Missingness 1: Correlations between Missingness Indicators (ignore significance)

		MISFE	MISFO	MISSEX	MISIQLO	MISIQPM	MISIQWS	MISEDUC	MISOCC
MISFE	Pearson Correlation	1.000	$\boxed{.220^{**}}$	.038**	.036**	-.017	-.022	.164**	.189**
	Sig. (2-tailed)	.	.000	.004	.007	.187	.092	.000	.000
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISFO	Pearson Correlation	.220**	1.000	.061**	.110**	.072**	.071**	.027*	.017
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000	.044	.190
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISSEX	Pearson Correlation	.038**	.061**	1.000	.036**	.021	.014	.070**	.065**
	Sig. (2-tailed)	.004	.000	.	.007	.117	.305	.000	.000
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISIQLO	Pearson Correlation	.036**	.110**	.036**	1.000	$\boxed{.614^{**}}$	$\boxed{.609^{**}}$	-.054**	-.063**
	Sig. (2-tailed)	.007	.000	.007	.	.000	.000	.000	.000
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISIQPM	Pearson Correlation	-.017	.072**	.021	.614**	1.000	$\boxed{.970^{**}}$	-.078**	-.093**
	Sig. (2-tailed)	.187	.000	.117	.000	.	.000	.000	.000
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISIQWS	Pearson Correlation	-.022	.071**	.014	.609**	.970**	1.000	-.086**	-.100**
	Sig. (2-tailed)	.092	.000	.305	.000	.000	.	.000	.000
	N	5690	5690	5690	5690	5690	5690	5690	5690
MISEDUC	Pearson Correlation	.164**	.027*	.070**	-.054**	-.078**	-.086**	1.000	$\boxed{.898^{**}}$

Patterns in Missingness 2: PCA Missingness Indicators (Varimax)

Rotated Component Matrix [a]

	Component		
	1	2	3
MISFE	-.039	.151	.727
MISFO	.070	-.115	.794
MISSEX	.029	.074	.274
MISIQLO	.790	-.031	.111
MISIQPM	.961	-.037	.001
MISIQWS	.959	-.046	-.005
MISEDUC	-.041	.963	.101
MISOCC	-.057	.963	.110

Patterns 3: Correlations Missingness Indicators and Substantive Vars

(Pairwise Deletion!)

		MISFE	MISFO	MISSEX	MISIQLO	MISIQPM	MISIQWS	MISEDUC	MISOCC
FATHEDUC	Pearson Correlation	.	.072	.036	.025	.097	.094	.027	.023
FATHOCC	Pearson Correlation	.063	.	-.002	-.021	-.037	-.035	.018	.008
SEX	Pearson Correlation	.193	-.010	.	-.074	-.157	-.160	.184	.267
IQLO	Pearson Correlation	-.447	-.044	-.023	.	.102	.101	-.124	-.146
IQPM	Pearson Correlation	-.271	-.012	-.034	.047	.	-.016	-.074	-.091
IQWS	Pearson Correlation	-.415	.004	-.004	.040	-.019	.	-.057	-.091
EDUC	Pearson Correlation	-.245	-.031	.047	.034	.087	.081	.	-.154
OCCUP	Pearson Correlation	-.192	-.029	.032	.037	.095	.091	-.029	.

In Sum: Missing But How?

- Missing Completely at Random (MCAR)
  - Missingness is not related to the variables under study
    - strongest assumption, simple and quick solutions
    - SPSS listwise deletion or **complete case** analysis, but there are better ways (impute)
- Missing at Random (MAR)
  - Missingness is related to the observed data but not to the missing data
    - weaker assumption, more complicated solutions
    - SPSS special module, other dedicated programs

In Sum:
Missing But How? continued

- Non-ignorable or Not Missing at Random (**NMAR**)
  - Missingness is related to the variables under study
    - Weakest assumption
    - Complicated solutions
    - Special models necessary
    - Need information on process of missingness
      - **Propensity model**

So, What is the Case?

? MCAR

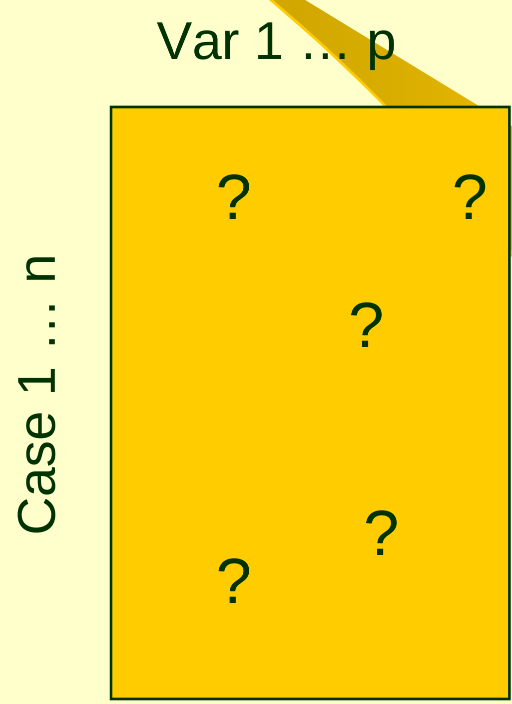
or

? MAR

or

? NMAR

- **Decision based on**
  - **A priori knowledge**
  - **Theory**
  - **Study of missing data pattern**



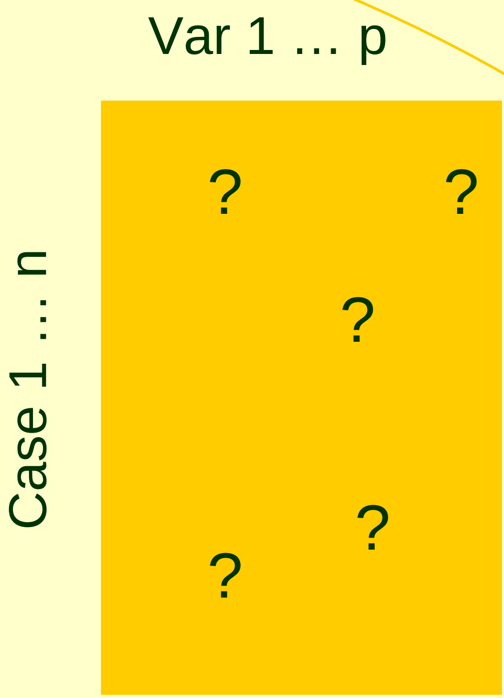
? = missing

Suggested Readings

- De Leeuw, E.D., Hox, J., and Huisman, M. (2003). Prevention and treatment of item nonresponse. Journal of Official Statistics, 19, 2, 153-176.
- Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

Part II: Extra Slides Prevention

See also De Leeuw et al (2003) www.jos.nu



Edith de Leeuw

Universiteit Utrecht

University of Essex

Sources Item-Nonresponse

- Researcher (by design)
- Interviewer
- Respondent
- Questionnaire
- Method of Data Collection

- Interaction between sources, e.g, respondent and questionnaire

What Can Be Done

- Missing by Design
  – Special analyses (e.g., multi-level analysis)
- Partial Non-Response (e.g., break-of)
  – Prevent
  – Adjust:
    - Delete cases and treat as unit-nonresponse (weighting)
    - Keep cases and impute missing answers
- Item Non-Response
  – Prevent
  – Adjust (impute!)

Mechanisms I: Interviewer

- Interviewer fails to:
  - Ask question
  - Record answer
  - Record answer correctly
    - In post-interview editing this will often be coded as missing
  - Fails to probe (ask again)

- Causes of failure:
  - Mistakes (e.g., wrong routing)
  - Purpose, cheating (e.g., fast interview, not wanting to go to much trouble)

Prevention I: Interviewer

- Mistakes:
  - Train interviewers in correct procedures
  - Give instruction about the questionnaire
  - Avoid mistakes by:
    - Ergonomic lay-out questionnaire or interviewer schedule (e.g., far less chance of skipping, routing errors, etc)
    - Use of computer-assisted interviewing (e.g., no routing errors, range checks)
- Cheating:
  - Stricter supervision
  - CAI

Mechanisms II: Respondent

- Respondent
  - Skips question by mistake
  - Refuses to answer
  - Not able to provide (correct) answer
- Causes:
  - Badly designed self-administered questionnaire (mistake)
  - Sensitive question (refusal)
  - A problem in the total question-answer process (not able to provide, e.g. memory in retrospective questions)

Prevention II: Respondent

- Write good questions and test them:
  - Comprehension question & answer categories
  - Inclusion of all relevant answer categories
- Avoid mistakes (cf. Interviewer mistakes)
  - Provide help (good instructions, etc)
  - Ergonomic lay-out questionnaire
  - CSAQ
- Pretest!
- Special formats
  - Sensitive questions
  - Retrospective questions

Mechanisms and Prevention III: The Questionnaire

- Good questionnaire helps to avoid mistakes of interviewer and/or respondent
- Question should be understood, categories should fit and be exhaustive (keep questions simple & understandable)
  - Pretest this
- Lay-out should be clear and guide from question to question
- Use graphical language consistently
  - SAQ, such as web/internet questionnaire

Suggested Readings

- De Leeuw, E.D., Hox, J., and Huisman, M. (2003). Prevention and treatment of item nonresponse. Journal of Official Statistics, 19, 2, 153-176.
- Downloadable without costs at
- www.jos.nu

Handling Incomplete Data in Longitudinal Surveys

Joop Hox

Edith de Leeuw

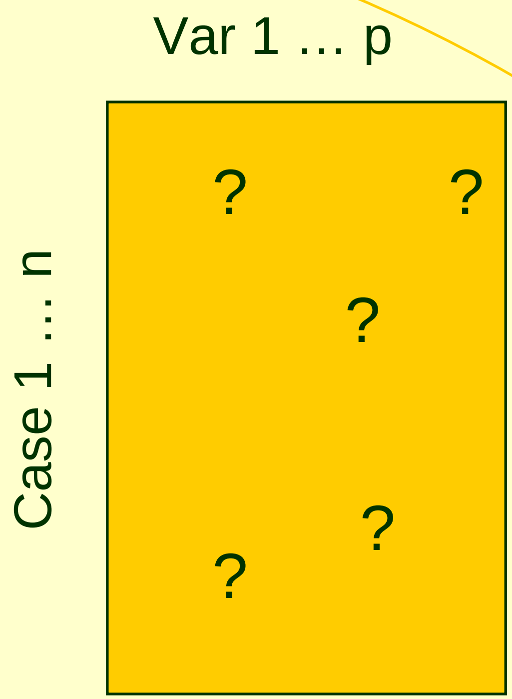
University of Essex

Universiteit Utrecht

Methodology of Longitudinal Surveys (MOLS) Short Course July 2006

Part III: Treatment
Analysing Missing Data

Simple solutions



Joop Hox

Universiteit Utrecht

MOLS
University of Essex

July 2006

Contents

- Ad hoc solutions and their (dis)advantages
- Principled solution: direct modeling of incomplete data
- Principled solution: multiple imputation

Important Distinctions

- Missing Completely At Random (MCAR)
  - missing data not related to anything
- Missing At Random (MAR)
  - missing data unrelated to unobserved value
  - but may be related to other observed variables
- Not Missing At Random (NMAR)
  - missingness related to unobserved (missing) value
- MCAR & MAR: Ignorable
  - under appropriate model
- NMAR: Nonignorable/Informative

Ad Hoc Solutions

- Analyze only observed part
  - Complete Cases
    - (Complete Cases with Weighting)
  - Available Cases

- Single imputation
  - Many methods

Complete Cases

- Delete incomplete cases
  - weigh complete cases to compensate selection
- SPSS: listwise deletion

Complete Cases: (Dis)Advantages

+ Simple
+ Standard Analysis Methods
- Inefficient
- Assumes MCAR

- Use: *If less than ±5% Is missing*

Available Cases

- Compute various statistics on cases available for each specific calculation
- Example:
  - compute means and standard deviations for all variables, using all available cases for each variable
  - compute correlations for all pairs of variables, using all available cases for each pair of variables (SPSS pairwise deletion)

Available Cases: (Dis)Advantages

+ *Appears* more efficient than Complete Cases

- May result in correlations outside [-1,+1]

- May result in ill-conditioned covariance or correlation matrix

- such as r_{12}= 1, r_{13}= 1, r_{23}= -1

- Assumes MCAR

- Sample size undefined

Use: *Never*

Complete and Available Case Analysis (SPSS)

1	10	15	8
2	3	2	8
3	6	4	11
4	4	10	2
5	17	11	26
6	10	99	16
7	10	99	5
8	11	99	12
9	14	99	14
10	10	99	13
11	4	10	7
12	14	21	23
13	15	17	13
14	5	3	7
15	22	19	22

Used for r_{12} r_{13} r_{23}

Thrown away in listwise, used for r_{13} in pairwise

Used for r_{12} r_{13} r_{23}

Listwise Deletion

		X1	X2	X3
Pearson Correlatio	X1	1.000	.765	.852
	X2	.765	1.000	.558
	X3	.852	.558	1.000
Sig. (2-tailed)	X1	.	.010	.002
	X2	.010	.	.093
	X3	.002	.093	.

a. Listwise N=10

Pairwise deletion

		X1	X2	X3
Pearson Correlation	X1	1.000	.765	.801
	X2	.765	1.000	.558
	X3	.801	.558	1.000
Sig. (2-tailed)	X1	.	.010	.000
	X2	.010	.	.093
	X3	.000	.093	.
N	X1	15	10	15
	X2	10	10	10
	X3	15	10	15

Example of Impossible Correlation Matrix (SPSS)

Data Matrix

1	1	1	99
2	2	2	99
3	3	3	99
4	4	4	99
5	5	5	99
6	1	99	1
7	2	99	2
8	3	99	3
9	4	99	4
10	5	99	5
11	99	1	5
12	99	2	4
13	99	3	3
14	99	4	2
15	99	5	1

Listwise Deletion

		X1	X2	X3
Pearson Correlation	X1	.[a]	.[a]	.[a]
	X2	.[a]	.[a]	.[a]
	X3	.[a]	.[a]	.[a]
Sig. (2-tailed)	X1	.	.	.
	X2	.	.	.
	X3	.	.	.

Pairwise Deletion

		X1	X2	X3
Pearson Correlation	X1	1.000	1.000	1.000
	X2	1.000	1.000	-1.000
	X3	1.000	-1.000	1.000
Sig. (2-tailed)	X1	.	.000	.000
	X2	.000	.	.000
	X3	.000	.000	.
N	X1	10	5	5
	X2	5	10	5
	X3	5	5	10

Imputation Methods

- Fill holes in data with plausible values
- Many methods, depending on ‘plausible’
- Impute with model based values
  - mean
  - regression
  - cold deck
- Impute with real values
  - hot deck
  - regression hot deck

Mean Imputation

- Replace missing value by the variable's mean computed for all available cases
  - unconditional mean imputation

+ Simple

- Assumes MCAR

- Underestimates variance

- Underestimates sampling error

Regression Imputation

- Replace missing value by value predicted from regression on observed variables
  - regression coefficients usually estimated on complete cases
  - conditional mean imputation

+ Assumes MAR if regression is linear

- Underestimates variance
  - but less than mean imputation
- Underestimates sampling error
  - but less than mean imputation

Cold Deck Imputation

- Replace missing value by a value, that is completely independent of the data set
  - for example: replace with population mean, expected value under random response

+ Simple

- Assumes MCAR

- Underestimates variance

- Underestimates sampling error

Hot Deck Imputation

- Replace missing value by a value, taken from similar but observed cases in data
  - there are a variety of 'hot deck' procedures
- 'Similar' defined by grouping variables
  - 'adjustment cells'
- 'Similar' defined by distance measure
  - 'nearest neighbor hot deck'

+ Often MAR

+ Better variance estimate than cold deck/mean

\- Imprecise control of sampling error

Regression Hot Deck Imputation

- Also called Predictive Mean Matching
- Use observed predictor variables to predict variable with missing values
  - regression equation based on complete cases
  - predictions for complete and incomplete cases
- Match each incomplete case to the complete case with closest predicted value
- Replace missing value by observed value of matched complete case

(Little, 1986; Landerman, Land & Pieper, 1997; Laaksonen, 1998)

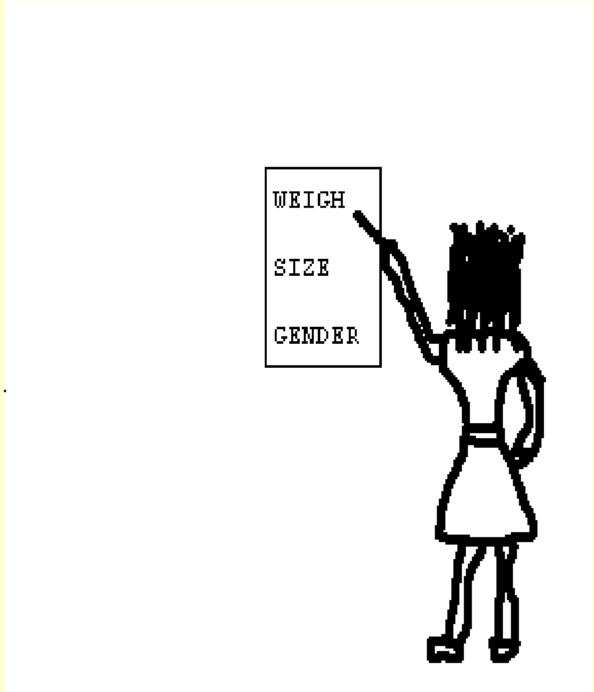
Imputation: (Dis)Advantages

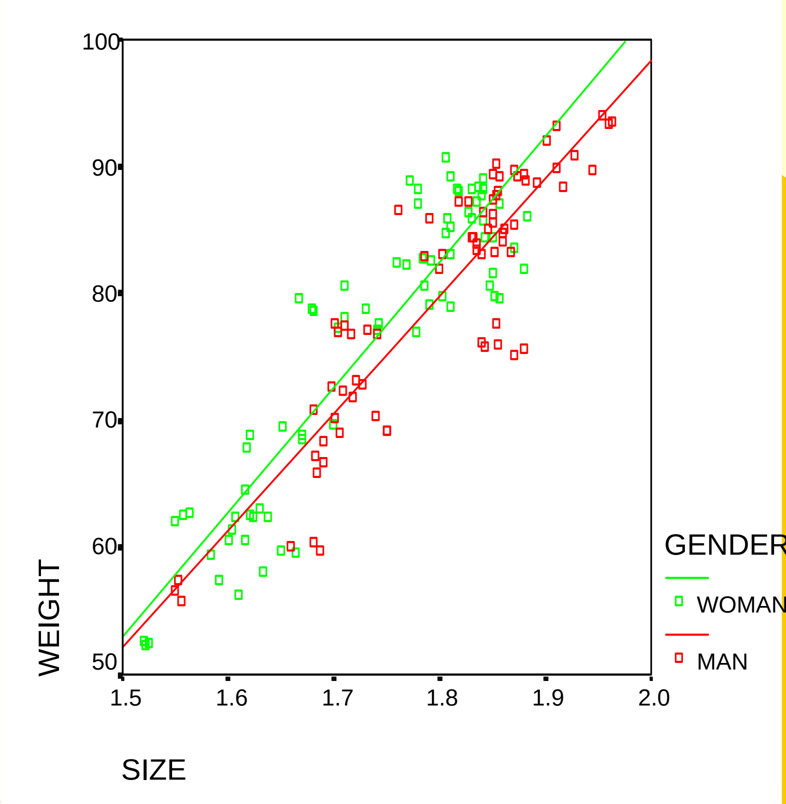
+ Fairly Simple
+ Imputation creates complete data set → standard analysis methods apply
- Often underestimate variance → underestimate sampling error
- Correct sample size undefined → underestimate sampling error
- Univariate method → distorts relationships

Silly example (again)

Hoytink, 2004

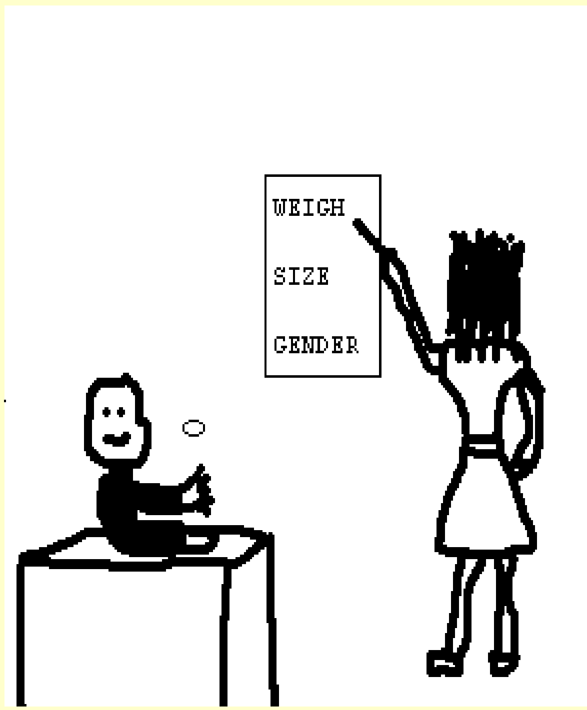
- Normal situation: complete data

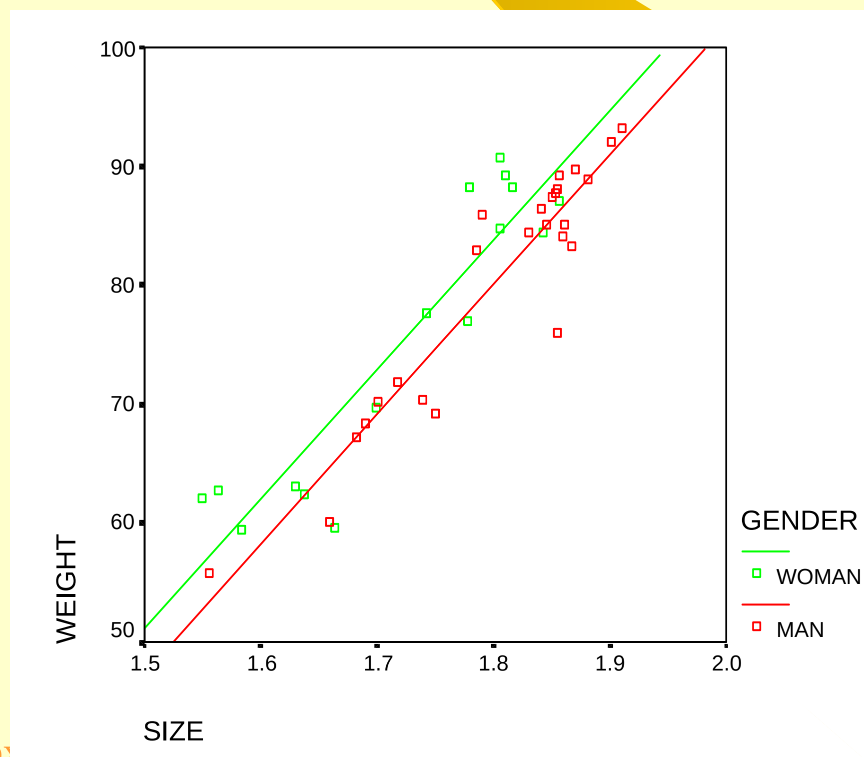




Silly example, complete cases

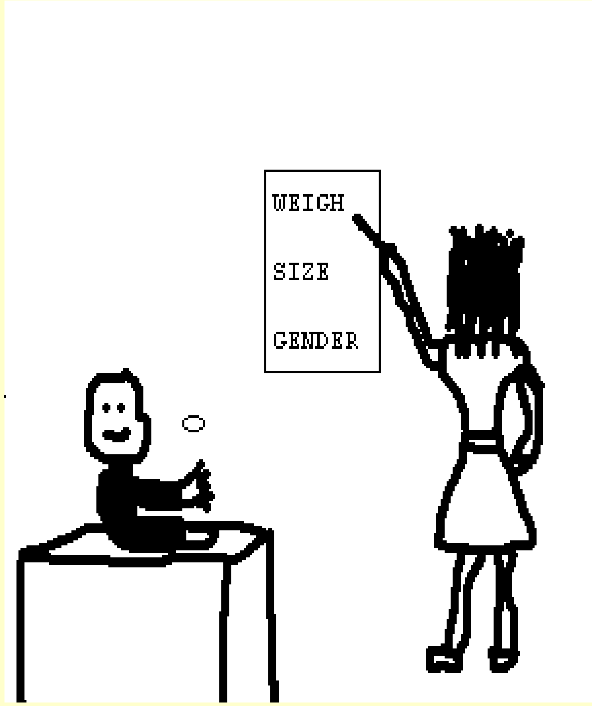
- Data Missing Completely At Random
  - The dog ate the interview forms!

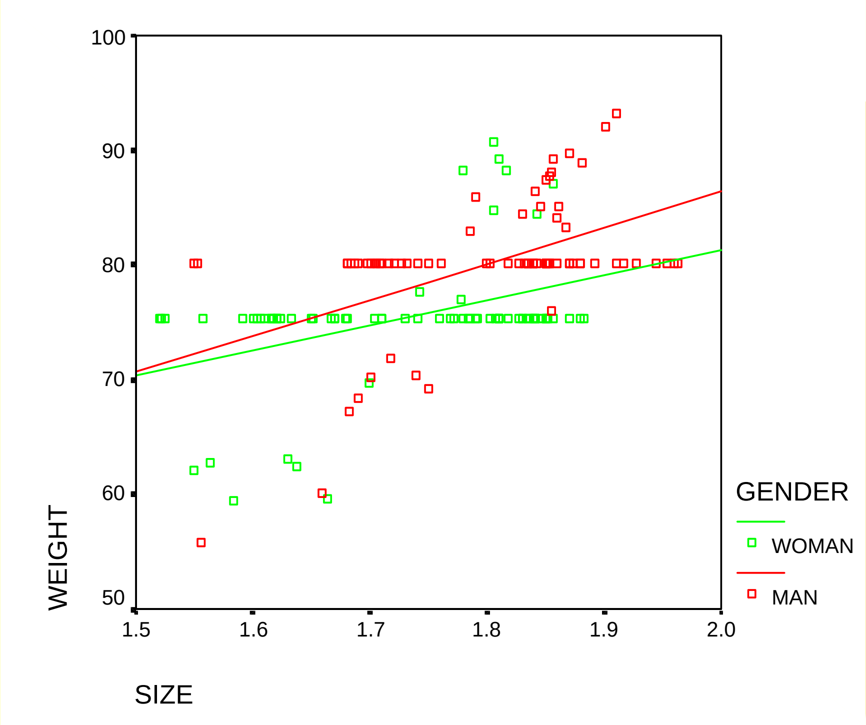




Silly example, mean substitution

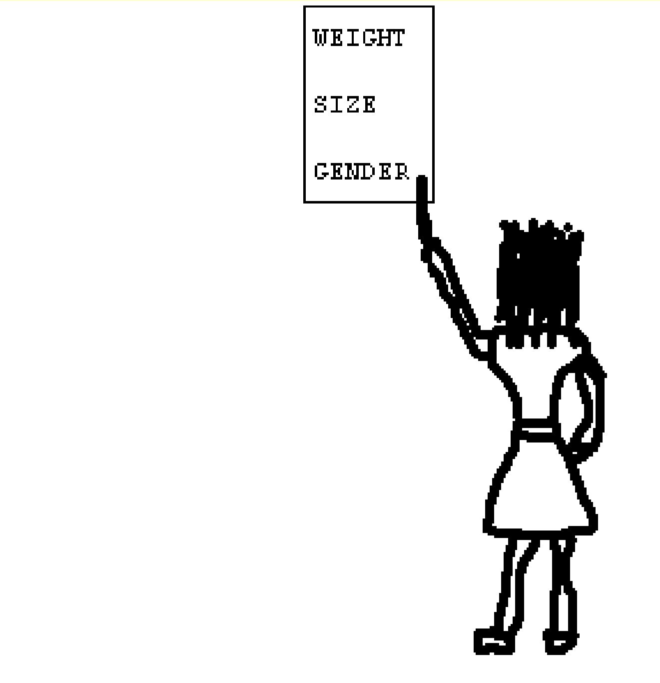
- Data Missing Completely At Random

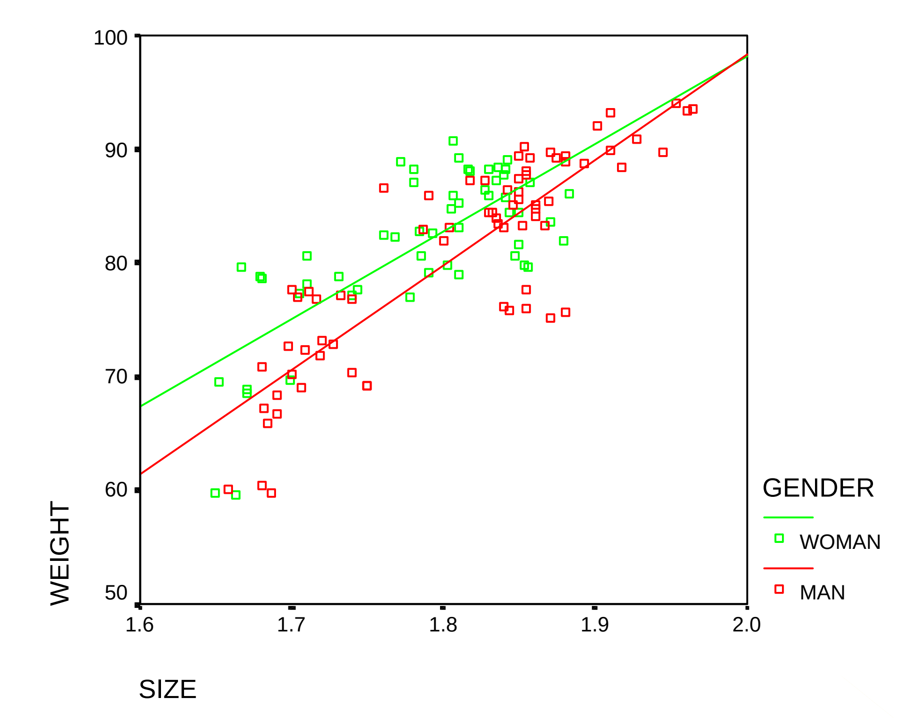




Silly example, MAR

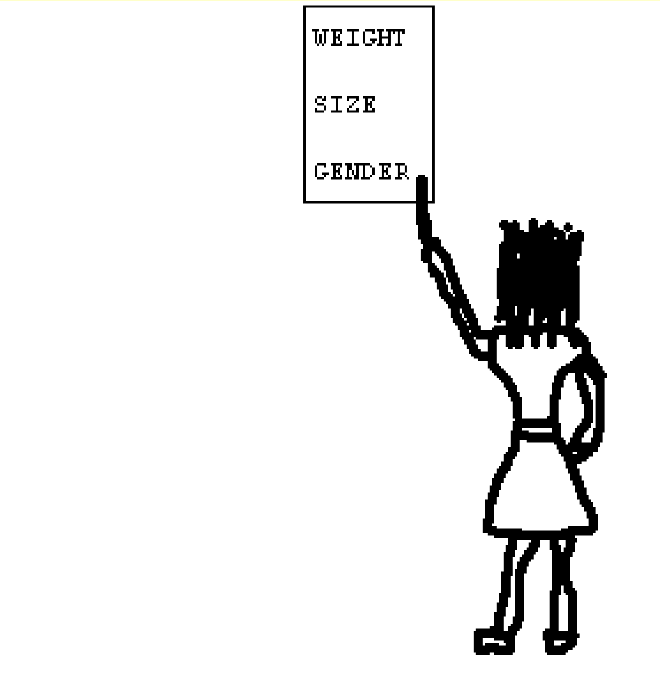
- Data Missing At Random:
  - Persons height < 1.65 meter cannot reach line 'weight'
- Default option 'do nothing' (complete cases)
  - Clearly biased!

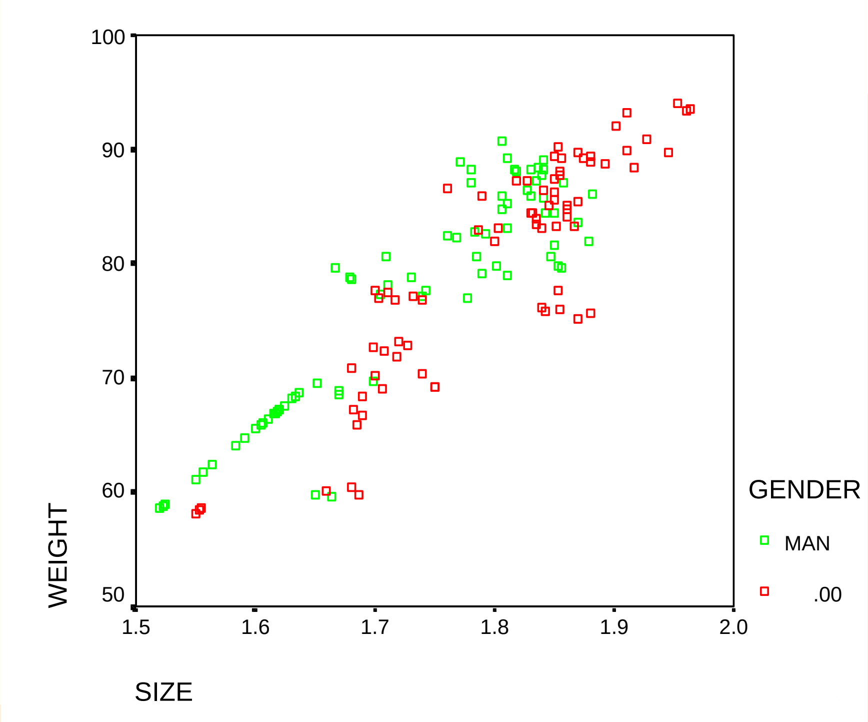




Silly example, MAR

- Data Missing At Random:
  - Persons height < 1.65 meter cannot reach line ‘weight’
- Regression imputation using gender & size
  - Reasonable, but not perfect!





Imputation with Errors Added

- Most imputation methods underestimate the variance
- Remedy: Add random error to imputed value
  - from statistical distribution
    - parametric, model based value
  - residual from similar case
    - nonparametric, real value

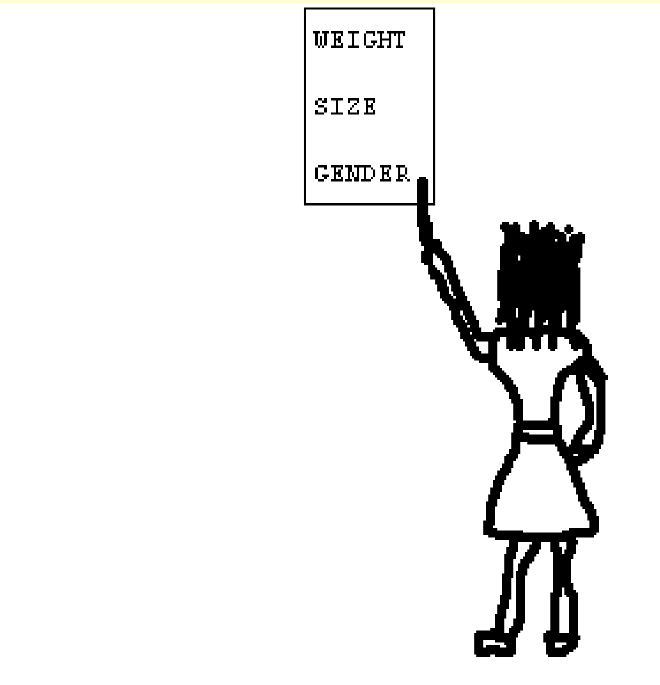
\+ Restores correct variance

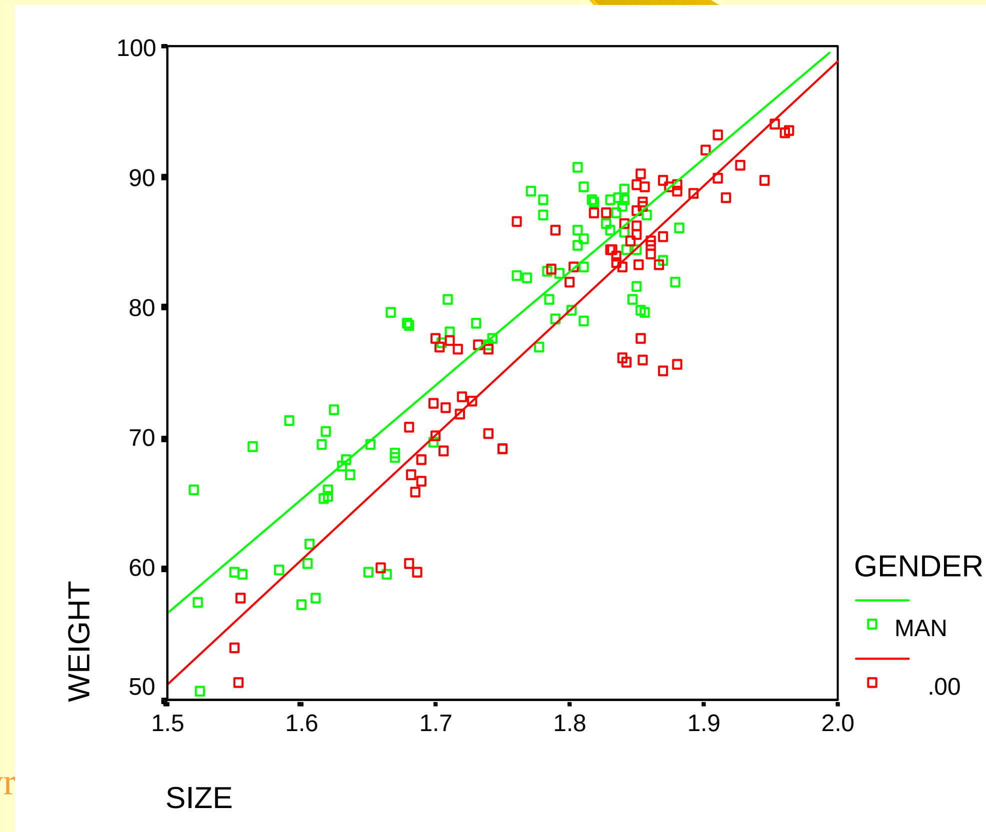
\- Correct sample size undefined

\- Not exactly replicable

Silly example, MAR

- Data Missing At Random:
  - Persons height < 1.65 meter cannot reach line ‘weight’
- Regression imputation using gender & size + error
  - Looks good!





Comparison of Ad Hoc Solutions on longitudinal *longmis* data: means

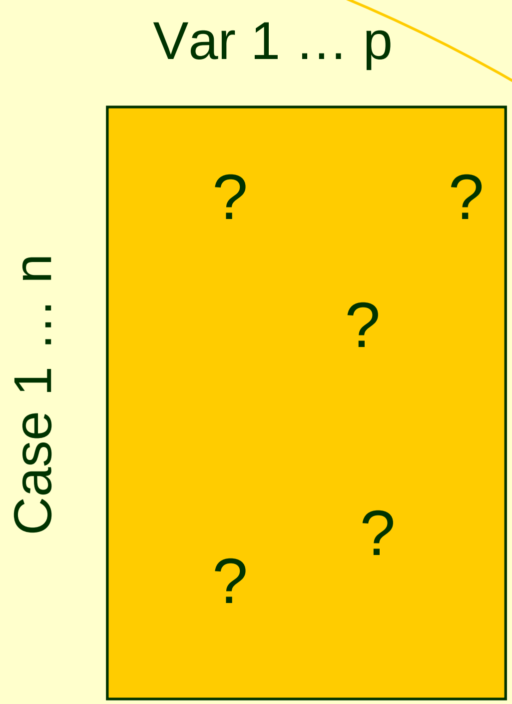
Means	T1	T2	T3	T4	T5
Complete data	50.5	51.5	52.5	53.6	54.6
Complete cases	52.2	53.2	53.4	55.0	56.4
Mean imputation	50.1	52.2	53.6	54.7	56.5
Regression imputation	50.2	52.0	52.7	54.3	55.7
Regression + error	50.3	51.8	52.6	54.0	55.6
Hot deck	50.1	51.4	52.2	53.7	54.4

- Data are MAR: dropout more probable after low outcome

Comparison of Ad Hoc Solutions on longitudinal *longmis* data: correlations

Correlation between T1 and	T2	T3	T4	T5
Complete data	.74	.74	.76	.71
Complete cases	.80	.77	.64	.57
Mean imputation	.65	.48	.43	.33
Regression imputation	.77	.69	.63	.50
Regression + error	.75	.71	.67	.50
Hot deck	.63	.65	.50	.53

Part II: Treatment Missing Data



Principled solution:

modeling of incomplete data

Likelihood Based Procedures

- Maximum Likelihood (ML): General procedure to estimate model parameters
- Special ML procedures for partially observed data
  - EM algorithm
  - Factored likelihood

Maximum Likelihood Estimation (ML)

- Data Y are assumed generated by a model with probability function (probability density) $f(Y/\vartheta)$
- ϑ are model parameters
- The Likelihood Function $L(\vartheta/Y)$ is a function of the parameters ϑ, which specifies the *Likelihood* of the data Y
- The Maximum Likelihood estimate of ϑ is the value that maximizes the likelihood $L(\vartheta /Y)$
  - For convenience often log-likelihood $l(\vartheta /Y)=ln(L(\vartheta /Y))$

Mathematics of ML with missing data

- Data *Y* and missingness pattern *R* have a joint probability function $f\ (Y,R\ /\ \vartheta,\psi)$
- Parameters ϑ for Y, ψ for R
- The Likelihood function for the joint model is $L(\vartheta,\psi\ /\ Y_{obs},R)$
  - So we need to estimate parameters for the data model and for the response model

BUT

- If missingness is MAR (or MCAR)

  then ϑ and ψ are *independent*
- We can use $L(\vartheta/Y_{obs})$ instead of $L(\vartheta,\psi\ /Y_{obs},R)$

ML estimation: MAR and NMAR

- If MAR (ϑ and ψ independent) we use $L(\vartheta/Y_{obs})$ instead of $L(\vartheta,\psi\ /Y_{obs},R)$
- We still need an algorithm to maximize $L(\vartheta/Y_{obs})$ with incomplete data
  - standard algorithms may not work on data with holes
- However, if NMAR (ϑ and ψ dependent) we *must* use $L(\vartheta,\psi\ /Y_{obs},R)$
  - and need a model for R

    (about which we seldom have information...)
- If MAR is tenable the model is *much* simpler

ML under MAR: EM Algorithm

- Two steps: **E**xpectation and **M**aximization step
- Expectation: given model parameters θ, compute expected value for all missing data in Y
- Maximization: given complete data Y, estimate θ by ML using standard procedure
- Thus the **EM** algorithm:
  - fill holes in data with plausible start values
  - estimate θ on completed data using standard ML
  - estimate missing data using model and current θ
  - repeat until convergence

(Dempster, Laird & Rubin, 1977)

(Dis)advantages EM algorithm

+ Under MAR unbiased estimates
+ Simple to program

- Special programs needed for different models
- Standard errors not included
  – Obtained by other means after EM convergence

Example of EM (SPSS)

Missing Data			Filled-in Data		
110	15	8	10	15	8
2 3	2	8	3	2	8
3 6	4	11	6	4	11
4 4	10	2	4	10	2
517	11	26	17	11	26
610	99	16	10	10.3	16
710	99	5	10	13.4	5
811	99	12	11	12.5	12
914	99	14	14	15.1	14
1010	99	13	10	11.1	13
11 4	10	7	4	10	7
1214	21	23	14	21	23
1315	17	13	15	17	13
14 5	3	7	5	3	7
1522	19	22	22	19	22

EM Correlations [a]

	X1	X2	X3
X1	1.000		
X2	.701	1.000	
X3	.801	.446	1.000

a. Little's MCAR test: Chisquare = .690, df = 2, Prob = .708

- Actually, EM does not fill in values, only sufficient statistics.
- Test shows MCAR assumption tenable
- Note no significances given (what *N*?)

EM as General Missing Data Method

- Use EM to estimate a very general model
  - SPSS: ‘data are multivariate normal’
- Use sufficient statistics from this model elsewhere
  - use correlations for factor analysis
- Impute missing data and use them elsewhere
  - use completed data to calculate sum score on scale

\+ Simple

\- No standard errors (what N?)

\- Standard significance tests biased (N too large)

- If single imputation is used, EM at least uses all available information assuming MAR

Maximum Likelihood on Incomplete Data

- ML estimation procedure can be adapted to work with incomplete data
  - *raw data likelihood*
- But needs appropriate software
- Structural Equation Modeling (SEM)
  - Assuming multivariate normality (Amos, Lisrel, Eqs)
  - For more data types in Mplus
- Multilevel analysis can also deal with incomplete longitudinal data using ML estimation

Comparison of Likelihood Based Solutions

Means	T1	T2	T3	T4	T5
Complete data	50.5	51.5	52.5	53.6	54.6
Complete cases	52.2	53.2	53.4	55.0	56.4
Hot deck (best ad hoc)	50.1	51.4	52.2	53.7	54.4
EM + ML (identical)	50.4	51.9	52.2	53.7	54.6

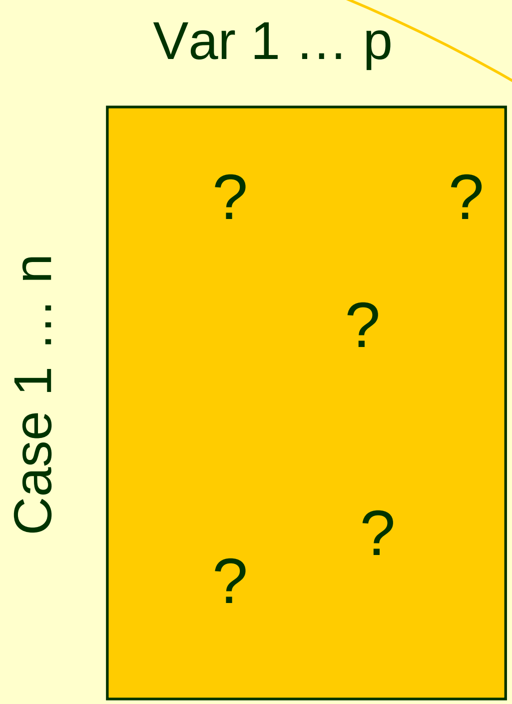
Comparison of Likelihood Based Solutions

Correlation between T1 and	T2	T3	T4	T5
Complete data	.74	.74	.76	.71
Complete cases (best ad hoc)	.80	.77	.64	.57
EM	.77	.75	.68	.66
ML	.77	.75	.69	.67

ML (in SEM) also gives standard errors: all correlations are significant

Part II: Treatment Missing Data

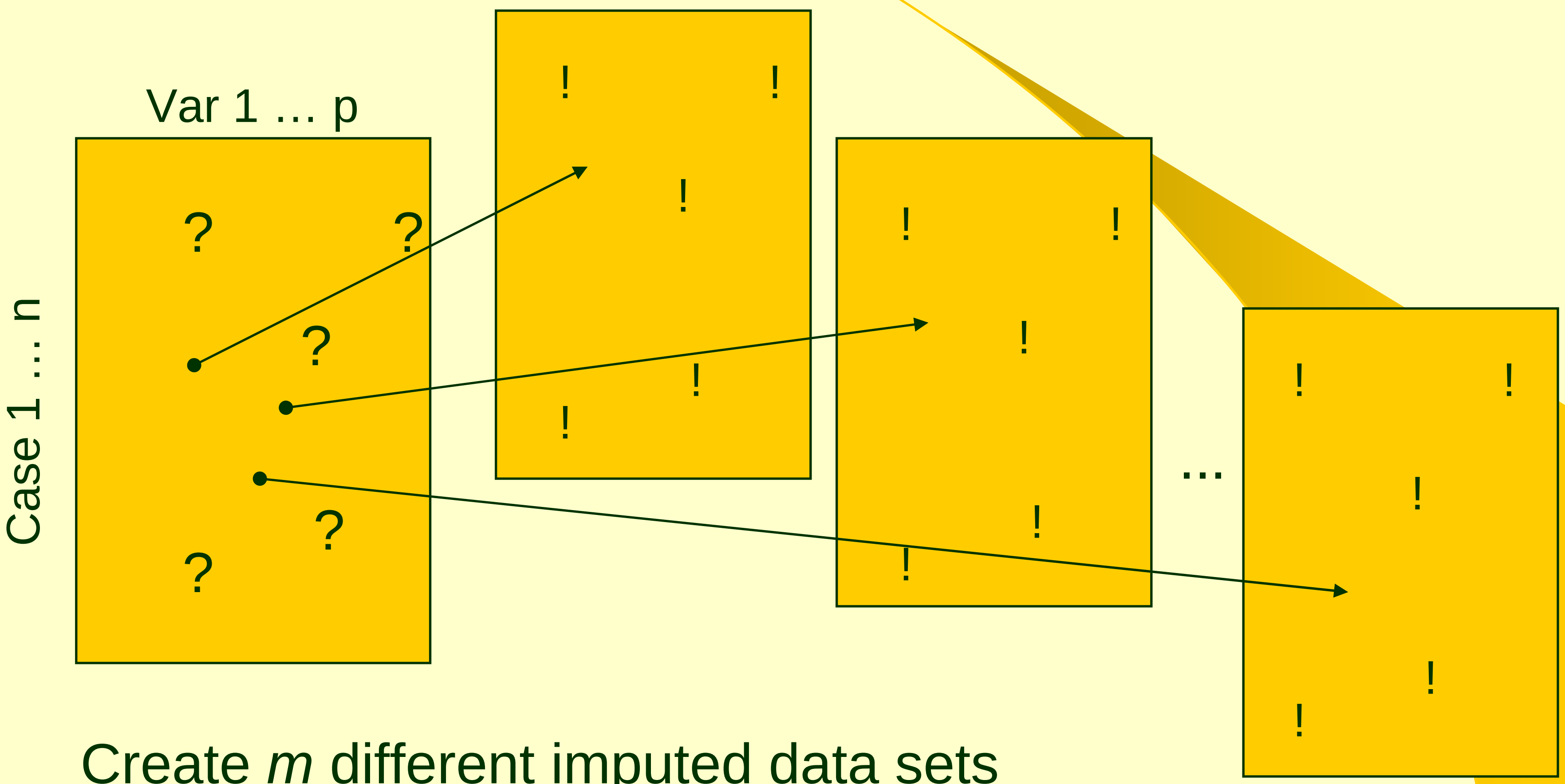
Multiple imputation



Single versus Multiple Imputation

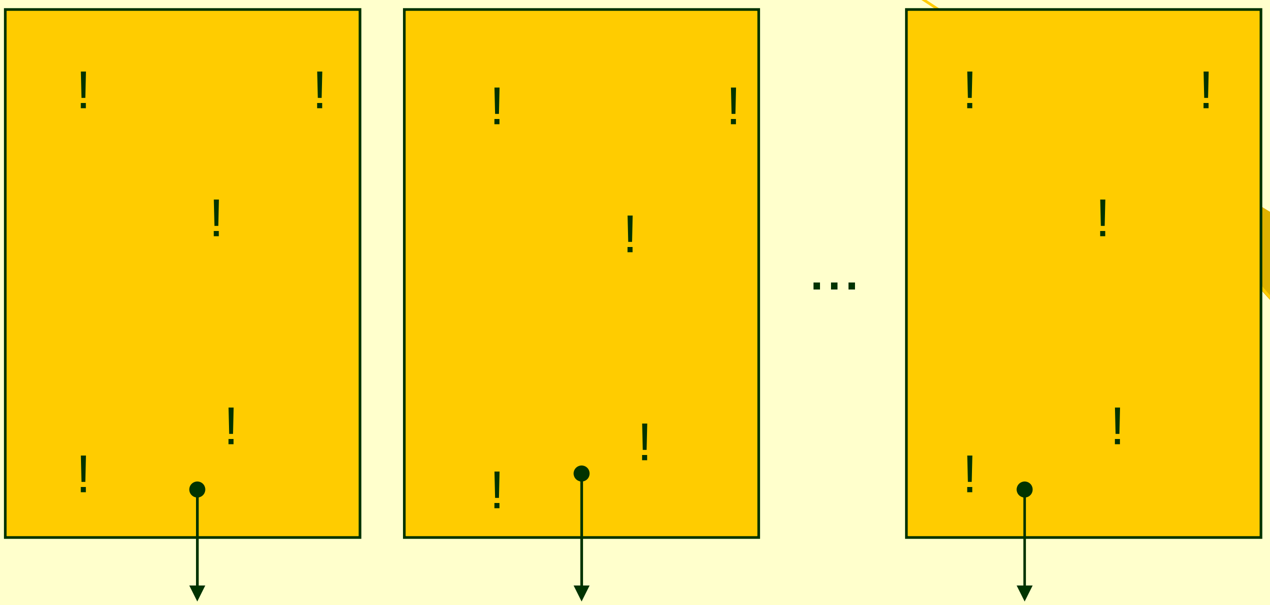
- Imputation = fill the holes in the data
  - usually with best possible estimate
  - followed by standard analysis
  - overestimates sample size, underestimates error
- Multiple Imputation (MI) = do this *m* times
  - with randomly chosen estimate from distribution of possible estimates
  - followed by *m* standard analyses
  - the *m* outcomes are then combined
  - the variation of *m* imputations restores the error

Multiple Imputation: Imputation



Create *m* different imputed data sets

Multiple Imputation: Analysis



Do standard complete data analysis *m* times

Combine results

Multiple Imputation: Key Idea

- Multiple Imputation *does not create extra data*
- It represents partially observed data so that it can be analyzed with standard complete-data techniques

Steps in Multiple Imputation

1. Create imputations
2. Analyze completed data sets
3. Combine the results

Create Imputations

- Parametric method
  - specify a model for complete data
  - for each missing data point:
    - estimate predictive distribution of the missing data
    - impute with a random value from this distribution
- Nonparametric method
  - group similar cases into adjustment cells
  - for each missing data point
    - collect non-missing cases from adjustment cell
    - impute with value from randomly selected non-missing case

Create Imputations: How Many?

- An estimator based on $m < \infty$ imputations has efficiency

$$\left(1 + \frac{\gamma}{m}\right)^{-1}$$

with γ = proportion missing *information*

  - note that $\gamma \neq$ proportion *missing data*

How Many? 3-5 Is Enough!

	γ				
m	.1	.3	.5	.7	.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Analyze *m* Completed Data Sets

- Standard complete data analysis techniques
- Obtain *m* sets of point estimates Q_i and variances (SE^2) U_i
- Combine *m* results into single outcome

Combine the Results

- Simply compute mean of *m* estimates

$$\overline{Q} = \frac{1}{m}\sum \hat{Q}_i$$

Combining Standard Errors

- *U* = Within imputation variance = mean of *m* variances

$$\overline{U} = \frac{1}{m}\sum U_i$$

- *B* = Between imputation variance = variance of point estimates

$$B = \frac{1}{m-1}\sum(\hat{Q}_i - \overline{Q})^2$$

- *T* = Total error variance

$$T = \overline{U} + (1 + m^{-1})B$$

MI Confidence Interval and Tests

- MI confidence interval
$$\bar{Q} \pm t_{df}\sqrt{T}$$
- MI significance test
$$t_{df} = \bar{Q}/\sqrt{T}$$
- Degrees of freedom *df*
$$df = (m-1)\left(1+\frac{m\bar{U}}{(m+1)B}\right)^2$$

Missing Information

- Estimate of the proportion of missing information

$$\gamma = \frac{r + 2(df + 3)}{r + 1}$$

$$r = (T - \bar{U}) / \bar{U}$$

with

Creating Imputations

- Generating MI data sets is difficult and requires special software

- Two approaches
  - ➔ Parametric
  - ➔ Nonparametric

Creating MI's, Parametric Approach

- MI data sets are simulated draws from a predictive distribution of the missing data
- Requires a model for the complete data
- With uncertainty about both missing values and parameters of predictive distribution

- Complex computations use Markov Chain Monte Carlo (MCMC) methods
  - data augmentation: Gibbs sampler, Metropolis-Hastings

Example: Univariate Normal Data

- Assume $y_1, y_2, \ldots, y_n \sim N(\mu, \sigma^2)$

$y_1, y_2, \ldots, y_a$ observed

$y_{a+1}, y_{a+2}, \ldots, y_n$ missing (MCAR or MAR)

how do we impute the missing Y's?

Univariate Normal Data (continuation)

- Assume $y_1, y_2, \ldots, y_n \sim N(\mu, \sigma^2)$
  $y_1, y_2, \ldots, y_a$ observed, $y_{a+1}, y_{a+2}, \ldots, y_n$ missing

$$\bar{Y}_{obs} = \frac{1}{a} \sum Y_i \qquad S^2_{obs} = \frac{1}{a-1} \sum (y_i - \bar{y}_{obs})^2$$

- Draw $y_{a+1}, \ldots, y_n$ from $N(\bar{Y}_{obs}, S^2_{obs})$?

- *Almost!*
  – But this ignores uncertainty about μ and σ^2

Univariate Normal Data (continuation)

- Assume $y_1, y_2, \ldots, y_n \sim N(\mu, \sigma^2)$

  $y_1, y_2, \ldots, y_a$ observed, $y_{a+1}, y_{a+2}, \ldots, y_n$ missing

Right way: $\sigma^2 \sim (a-1)S^2_{obs} / \chi^2_{a-1}$

- Take $\mu \sim N(\bar{y}_{obs}, \sigma^2 / a)$
- Take
- Take $y_{a+1}, \ldots, y_n$ from $N(\mu, \sigma^2)$ *!*
  - Repeat *m* times

Creating MI's, Nonparametric Approach

- Use logistic regression on complete variables to predict nonresponse on incomplete variable
- Divide the sample into imputation classes based on predicted nonresponse probability (propensity score)
- Randomly impute observed value from imputation class
- *Almost!*
  - But this ignores uncertainty about logistic regression parameters

Creating MI's, *Correct* Nonparametric Approach

- *Right way*: Bootstrap logistic regression
- Use bootstrapped regression equation to predict nonresponse on incomplete variable
- Divide the sample into imputation classes based on predicted nonresponse probability (propensity score)
- Randomly impute observed value from imputation class
  - This restores the variability we have because we must estimate the propensity scores

Multiple Imputation: Models and Software

- SPSS Regression + Error is not correct!
  - SAS MI procedures are correct
- NORM multivariate normal (Splus, Windows)
- CAT categorical (Splus)
- MIX continuous and categorical (Splus)
- PAN panel data (Splus)
  - available at http://www.stat.psu.edu/~jls/
- Amelia multivariate normal & longitudinal (Windows)
  - http://gking.harvard.edu/stats.shtml
- Mice multivariate normal (Windows)
  - http://www.multiple-imputation.com/
- SOLAS nonparametric bootstrap solution
  - commercial, http://www.statsol.ie

Multiple Imputation Free Windows Software

- NORM multivariate normal
- Amelia multivariate normal & longitudinal
- Mice multivariate normal

- Normality assumption applies only to incomplete variables
- Normalizing transformations followed by backtransformations
- Categorization of ordinal, nominal information
  - Automatic in Norm, Amelia
- In general, MI appears robust against mild violations of scale assumptions

Multiple Imputation versus Likelihood Based Procedures

- ML procedures
  - + efficient
  - - model specific
  - - complicated
- MI procedures
  - + general, uses standard complete data techniques

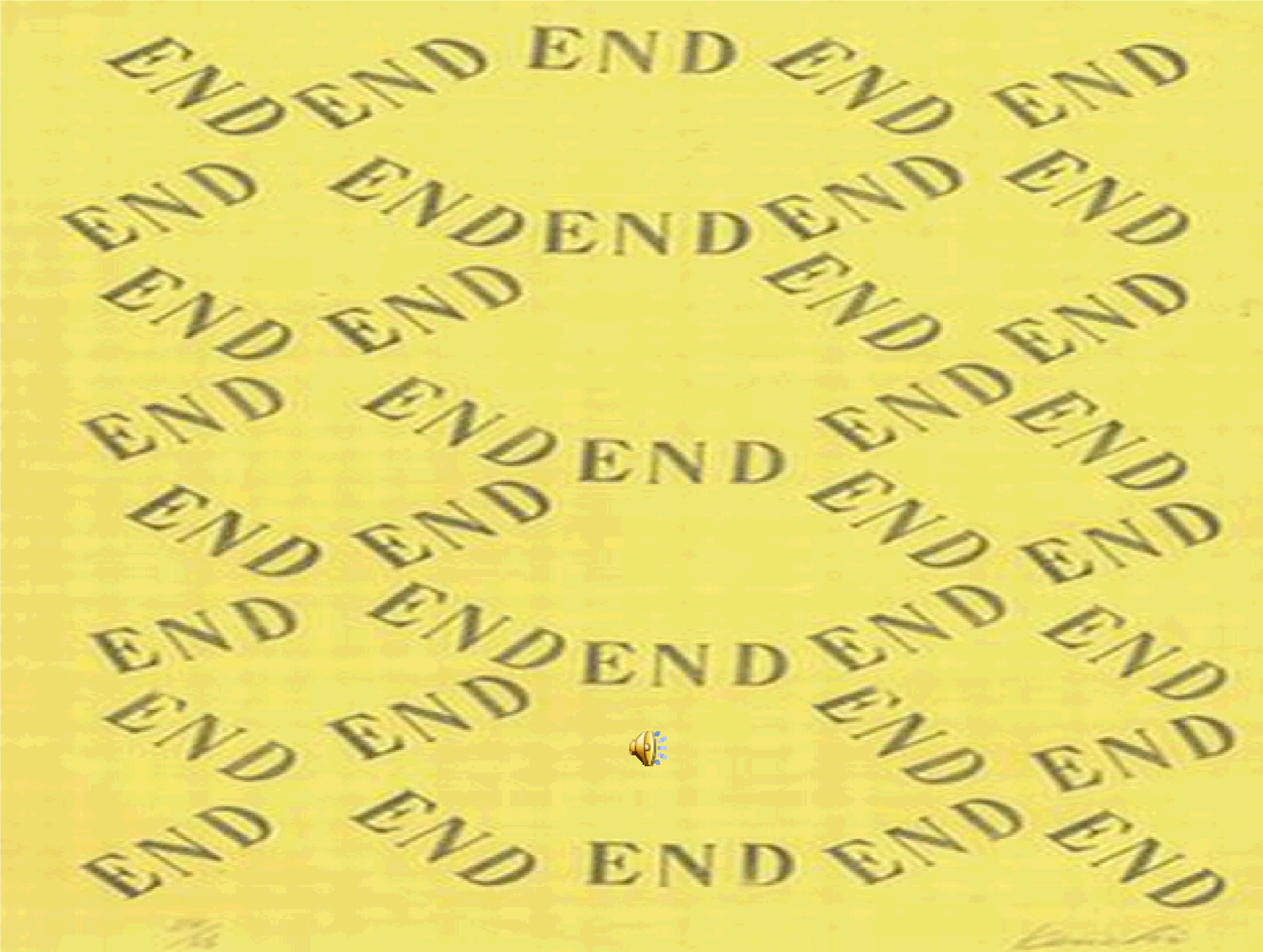
    (which need not be Likelihood-based)
  - - complicated

Suggested Reading
Introductory

- De Leeuw, E.D., Hox, J., and Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 19, 2, 153-176.
- A. Acock (1997). Working with missing values. *Family Science Review*, 10, 76-102.
- Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.

Suggested Reading
Statistical

- R.J.A. Little & D.B. Rubin (1987). *Statistical analysis with missing data*. New York: Wiley.
- J.L. Schafer (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.


END END END END END
END END END END END
END END END END
END END END END END
END END END END
END END END END END
END END END END
END END END END END