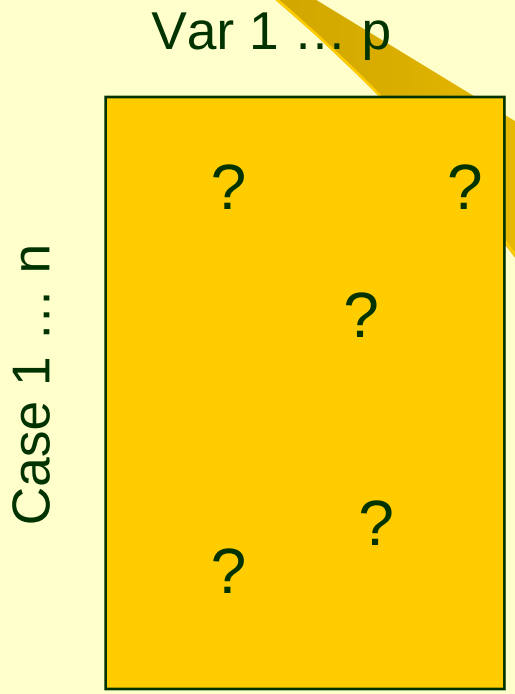
Missing Data: Patterns, Mechanisms & Prevention

Edith de Leeuw

Thema middag Nonresponse en Missing Data, Universiteit Groningen, 30 Maart 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

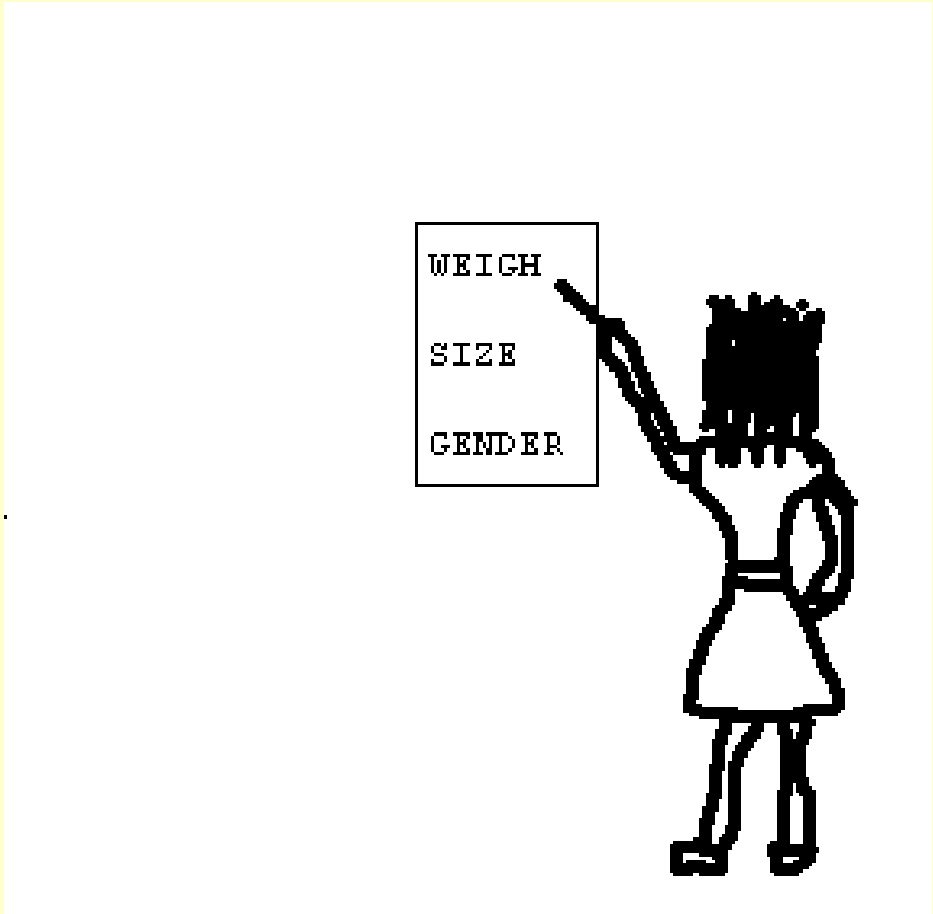
- Potential for *biased* results
  - Univariate analysis and (general) low item-nonresponse: bias is generally small
  - Multivariate analysis, even with low item nonresponse for each question, cumulates to a substantial proportion of records that are missing
- So: do something
- Simply ignoring (standard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

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

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 understandable)
  - Pretest this: Expert reviews, cognitive interviewing,etc
- Lay-out should be clear and guide from question to question
- Use graphical language consistently
  - SAQ, such as web/internet questionnaire

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 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, 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 = **?**)

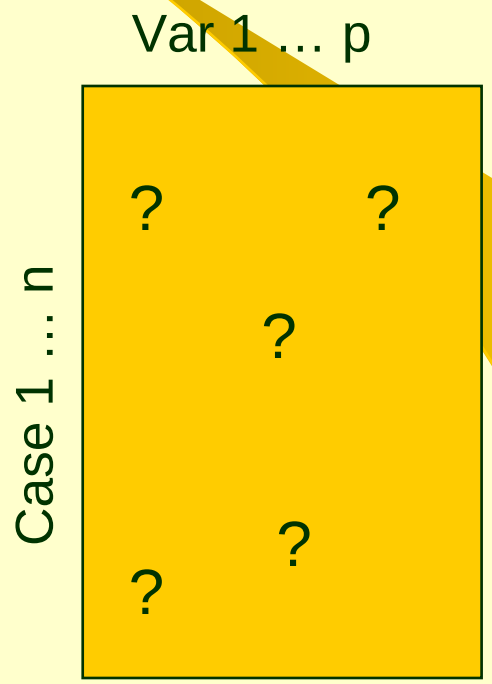
? Unsystematic (MCAR)

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

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 (missingness indicators)
  - 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

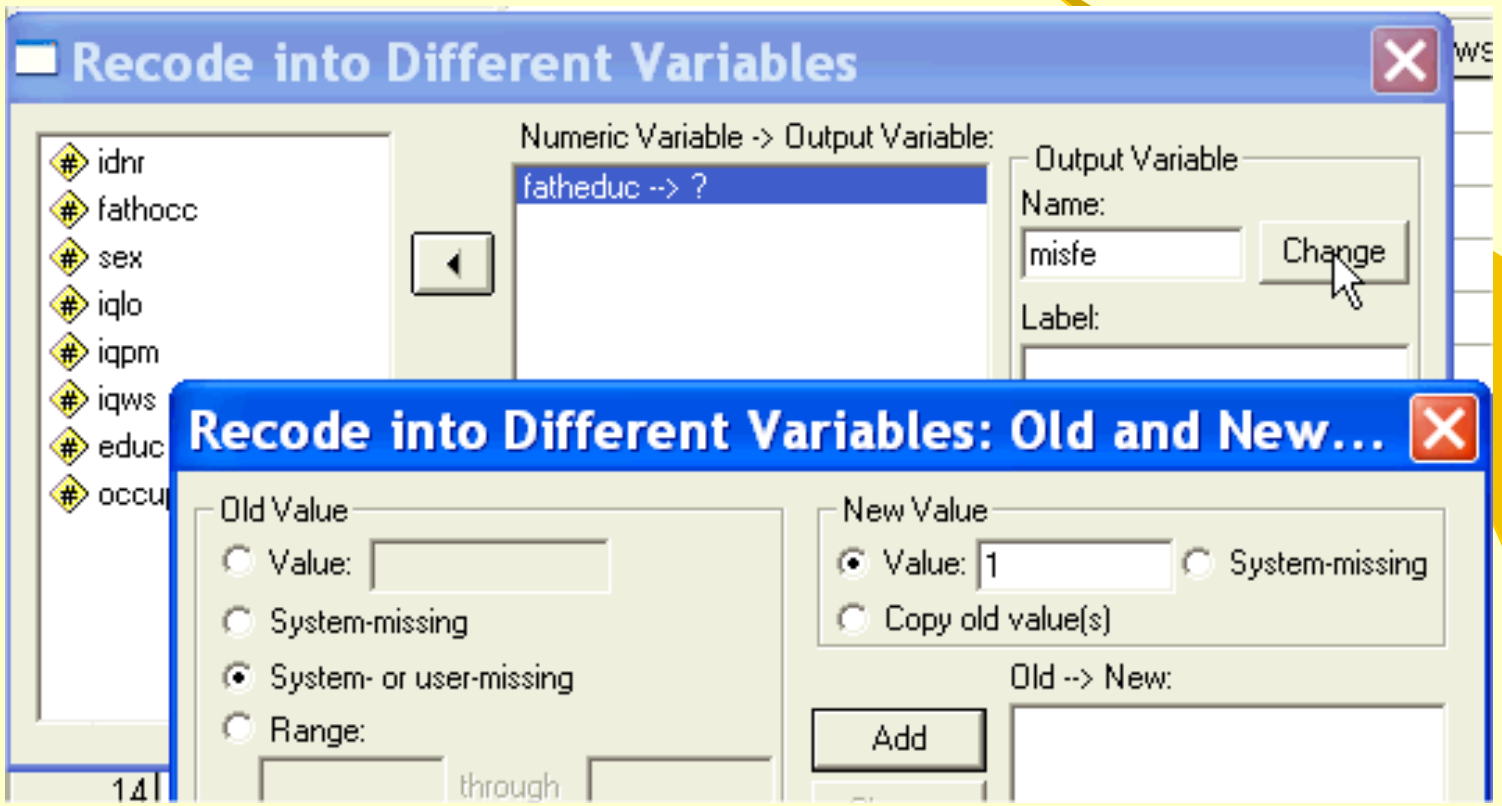
- Data set from educational research
  - Order of variables: idnr, father education (fatheduc), father occupation (fathocc), sex, iqlo, iqpm, iqws, education (educ), occupation (occup)
  - Note 1: iglo,iqpm,iqws are three IQ-tests
  - Note 2: 2 variables measure ‘father of pupil’ rest of variables measure 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	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	20.1	-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	-20.1	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?

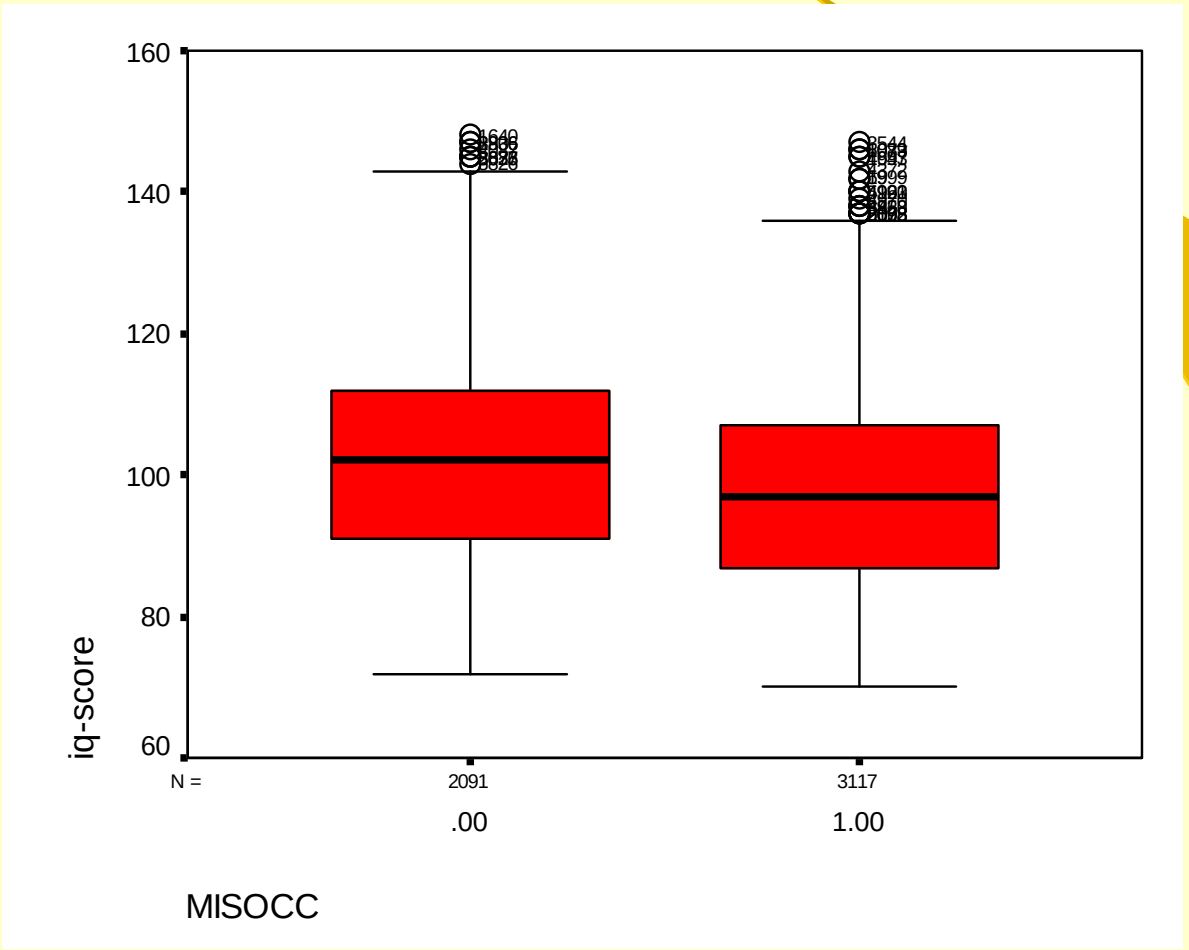
Step 4 Test MCAR continued How about IQ?: T-test

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	.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	.614**	.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	.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	.898**

Patterns 2: 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	.

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.