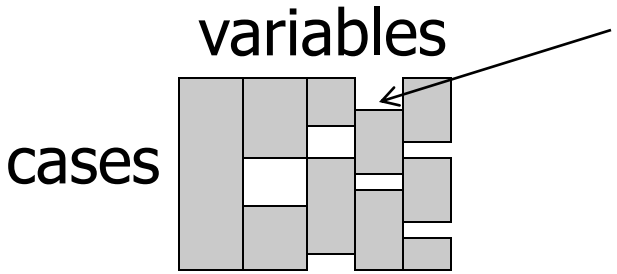
Statistical Analysis with Missing Data

Module 1: introduction, overview

Module 1: Introduction and Overview

- Missing data defined
- The problem: some examples
- Patterns and mechanisms
- Analysis strategies
  - properties of a good method
  - complete-case analysis
  - imputation, and multiple imputation
  - analysis of the incomplete data matrix

Missing data defined



- Always assume missingness hides a meaningful value for analysis
- Examples:
  - Missing data from missed clinical visit(√)
  - In a longitudinal study of blood pressure medications:
    - losses to follow-up (√)
    - deaths (x)

MISSING DATA EVERYWHERE

Cross sectional study

Longitudinal cohort study

Clinical trial

Electronic health records

and more ...

Missing Data in Cross-Sectional or Longitudinal Cohort Study

(*Ex 1.1: Woolson and Clark Data*)

- Muscatine Coronary Risk Factor Study: a longitudinal study of coronary risk factors in schoolchildren.
- Five variables (sex, age, and obesity for three rounds of the survey) are recorded for 4856 units;
- Sex and age are completely recorded, but the three obesity variables are sometimes missing, thereby generating six patterns of missingness.
- Table 1.1 summarizes the pattern of missing data in the data matrix

Table 1.1: Pattern of Missing Data: 0 = observed, 1 = missing

Pattern	Age	Sex	Wt1	Wt2	Wt3	N
A	0	0	0	0	0	1770
B	0	0	0	0	1	631
C	0	0	0	1	0	184
D	0	0	1	0	0	645
E	0	0	0	1	1	756
F	0	0	1	0	1	370
G	0	0	1	1	0	500

- Because age is recorded in five categories and the obesity variables are binary, the data can be displayed as counts in a contingency table.
- Table 1.2 displays the data in this form, with missingness of obesity treated as a third category of the variable, where O = obese, N = not obese, and M = missing. Thus, the pattern MON denotes missing at the first round, obese at the second round, and not obese at the third round, and the other five patterns are defined analogously.

a

Table 1.2: Data for Example 1.1

	Males						Females				
Response	Age Group						Age Group				
Category	5-7	7-9	9-11	11-13	13-15		5-7	7-9	9-11	11-13	13-15
NNN	90	150	152	119	101		75	154	148	129	91
NNO	9	15	11	7	4		8	14	6	8	9
NON	3	8	8	8	2		2	13	10	7	5
NOO	7	8	10	3	7		4	19	8	9	3
ONN	0	8	7	13	8		2	2	12	6	6
ONO	1	9	7	4	0		2	6	0	2	0
OON	1	7	9	11	6		1	6	8	7	6
OOO	8	20	25	16	15		8	21	27	14	15
NNM	16	38	48	42	82		20	25	36	36	83
NOM	5	3	6	4	9		0	3	0	9	15
ONM	0	1	2	4	8		0	1	7	4	6
OOM	0	11	14	13	12		4	It	17	13	23
NMN	9	16	13	14	6		7	16	8	31	5
NMO	3	6	5	2	1		2	3	1	4	0
OMN	0	1	0	1	0		0	0	1	2	0
OMO	0	3	3	4	1		1	4	4	6	1
MNN	129	42	36	18	13		109	47	39	19	11
MNO	18	2	5	3	1		22	4	6	1	1
MON	6	3	4	3	2		7	1	7	2	2
MOO	13	13	3	1	2		24	8	13	2	3
NMM	32	45	59	82	95		23	47	53	58	89
OMM	5	7	17	24	23		5	7	16	37	32
MNM	33	33	31	23	34		27	23	25	21	43
MOM	11	4	9	6	12		5	5	9	1	15
MMN	70	55	40	37	15		65	39	23	23	14
MMO	24	14	9	14	3		19	13	8	10	5

Woolson and Clarke fit multinomial distributions over the 26 categories for each column in Table 1.2.

That is, missingness is regarded as defining strata of the population.

Perhaps better: treat the nonresponse categories as missing value indicators and estimating the joint distribution of the three dichotomous outcome variables from the partially missing data.

Attrition in Longitudinal Studies

- Longitudinal studies often have drop-outs
  - Move out of study catchment area
  - Participation becomes too onerous
- Common analyses have problems:
  - complete case analysis is biased if drop-outs differ
  - Naïve imputation (e.g. last observation carried forward) involves unrealistic assumptions

Missing Data in Sample Surveys

(Exs. 1.3, 1.5)

- Nonresponse in opinion polls:
  - Missing data *if* the nonrespondent votes
- National Health and Nutrition Examination Survey (NHANES) III. Public Use Files subject to:
  - Unit nonresponse
    - noncontact
    - refusal
  - Item nonresponse
    - questionnaire interview complete, health examination missing
    - Individual items – "swiss cheese pattern"

Unit nonrespondents in surveys (Ex. 1.5)

- Unit nonrespondents may differ from respondents, leading to
  - **nonignorable** missing data
  - biased estimates. A simple formula for means:

$$\bar{Y}_R - \bar{Y} = \pi_{NR} \times (\bar{Y}_R - \bar{Y}_{NR})$$

Bias = NR rate * difference in R and NR means

NR = nonrespondent, R = respondent

Unit nonrespondents in surveys (cont.)

- Often very limited info on nonrespondents
- One approach is to link to external data
- Another approach is follow up a subsample of nonrespondents with special efforts:
  - abbreviated interview
  - monetary incentives
  - Data collected can be weighted to represent all nonrespondents, or used to (multiply) impute other nonrespondents

Missing Data in Clinical Trials (Ex 1.9)

- "A long standing issue in clinical trials, and especially in regulatory submissions that contain clinical trials intended to support efficacy and safety and marketing approval
- ICH E9 addresses it briefly but no analysis advice
- FDA's critical path initiative identified this as a topic in the streamlining of clinical trials and PhRMA, in negotiating the PDUFA 4 agreement, wanted FDA to bring consensus to this topic (FDAAA)"
- (From presentation by Robert T. O'Neill Ph.D., Director, Office of Biostatistics Center for Drug Evaluation and Research, FDA ,at the Annual Conference of the International Society for Clinical Biostatistics, Prague, Aug, 2009)

Key Take-Home Messages

- Missing data undermines randomization, the lynchpin of inferences in confirmatory trials
- Limiting missing data should be a major consideration when weighing alternative study designs
  - Analysis methods come with unverifiable assumptions, and limiting these assumptions is crucial
- Careful attention to avoiding missing data in trial conduct can greatly limit the scope of the problem
- Analysis methods need to be driven by plausible scientific assumptions
- Sensitivity analyses to assess robustness to alternative analysis models are needed
  - Lack of robust treatment effect from these analyses reinforces the need to limit missing data in trial design and conduct

Dose-Titration Study of Tacrine for Alzheimer's Disease

- Randomized, double-blind dose-escalation study (Knapp et al. 1994). Outcome - ADAS-COG

Treatment	Time (t)				
	1	2	3	4	5
Placebo	0	0	0	0	0
80mg	40	80	80	80	80
120mg	40	80	120	120	120

The Drop-Out Problem

- Titration to higher dosages to avoid side-effects on liver function
- Patients with side effects removed from double-blind study
- Other drop-outs from lack of compliance, dose-related adverse events
- Substantial differential drop-out rate at t=5:
  - Placebo 44/184 (24%)
  - 80mg 31/61 (51%)
  - 120mg 244/418 (57%)

GVHD Clinical Trial

- Longitudinal study of two treatments of GVHD in transplant patients
- Outcome: summary score of extent/severity of skin disease in 10 areas -- % change from baseline after 12 weeks (one of a set of repeated measures)
- Primary analysis: Wilcoxon Rank Sum Test
- Drops-outs for various reasons:
  - Death; withdrawal for adverse events; administrative reasons; transfer to treatment arm by protocol with adverse GVD outcome
- Initial protocol specified LOCF imputation, rejected by FDA

Missing Data in Electronic Health Records

EHRs from OSU Wexner Medical Center

- Data consist of EHRs from 24655 type 2 diabetes(T2D) patients diagnosed between January 1, 2013, and December 31, 2017.
- Many variables were extracted from the Ohio State University Wexner Medical Center Information Warehouse and they contain important biomarkers related to T2D including:
  systolic blood pressure (SBP), total cholesterol (TC), HbA1c, high-density lipoprotein (HDL), body mass index (BMI) and T2D medications.
- There were a total of over 100 medications (mono- or combined therapies) recorded in the data.

EHRs are particularly useful for ITRs

- Big benefit of using EHRs include cost effectiveness and reflection of real time and real-world evidence that are often missing in standard trials or cohort studies.
- Sound analysis of EHRs provides a great opportunity to learn patient's disease progression, so as to make accurate disease prognosis and treatment optimization.
- In a big data framework, EHRs contain health records from thousands of patients (volume), have rich information from heterogenous patients regarding a list of treatments and a range of outcomes (variety), and actually reflect real practice in one or more particular health systems (veracity).

Data time-windows for each analysis

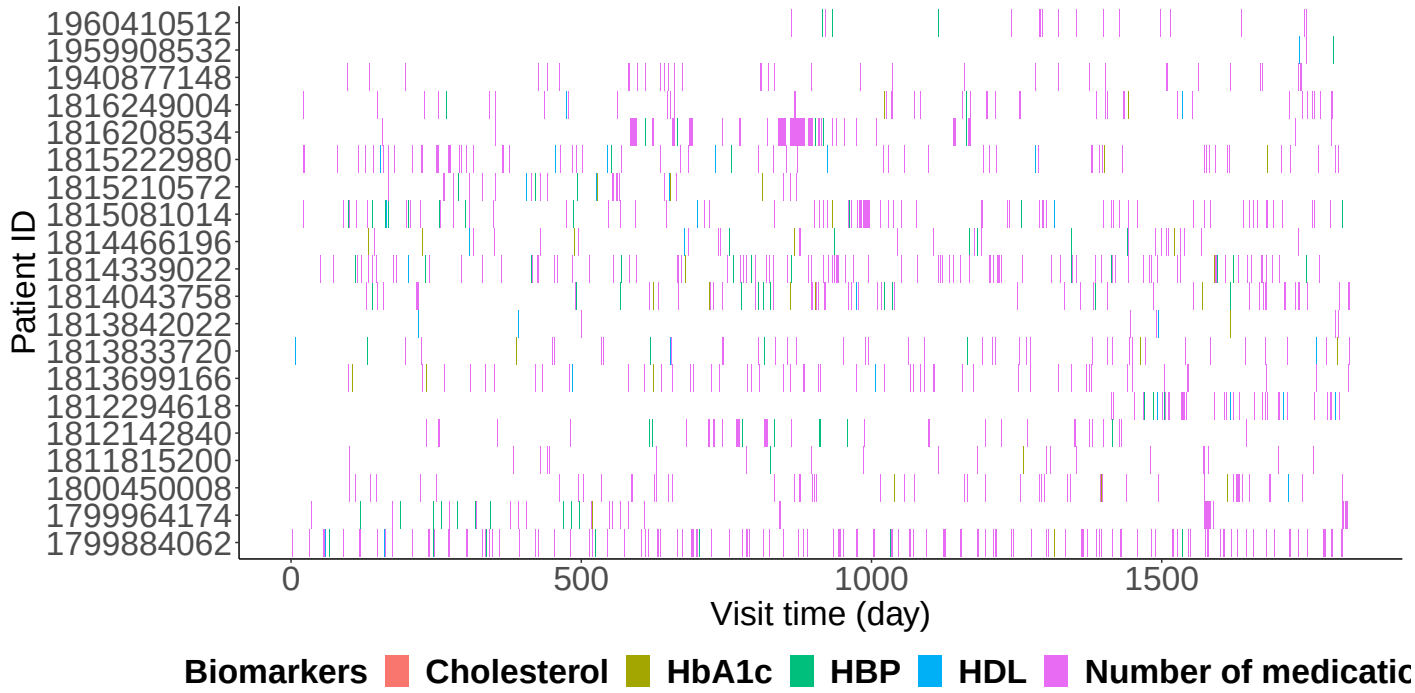
- Year 2013-2016: data from this wide window to learn latent heterogeneity (patient's disease chronicity)
- Year 2015-2016: data to extract short term feature variables
- Last encounter in year 2016: time zero to obtain treatment information
- Year 2016-2017: data in this window to derive outcome (HbA1c in 6 months)

2013-2016	2015-2016	Baseline Treatment Date	2016-2017
36-month-data 8456 patients	12-month-data 5133 patients	5133 patients	12-month-data 5133 patients
Fit the latent process model to estimate patient subgroups	Calculate the recent 12-month mean biomarkers to estimate ITRs	Last clinical visit in 2016 at which patients received T2D treatments	Fit regression models to create the outcome variable

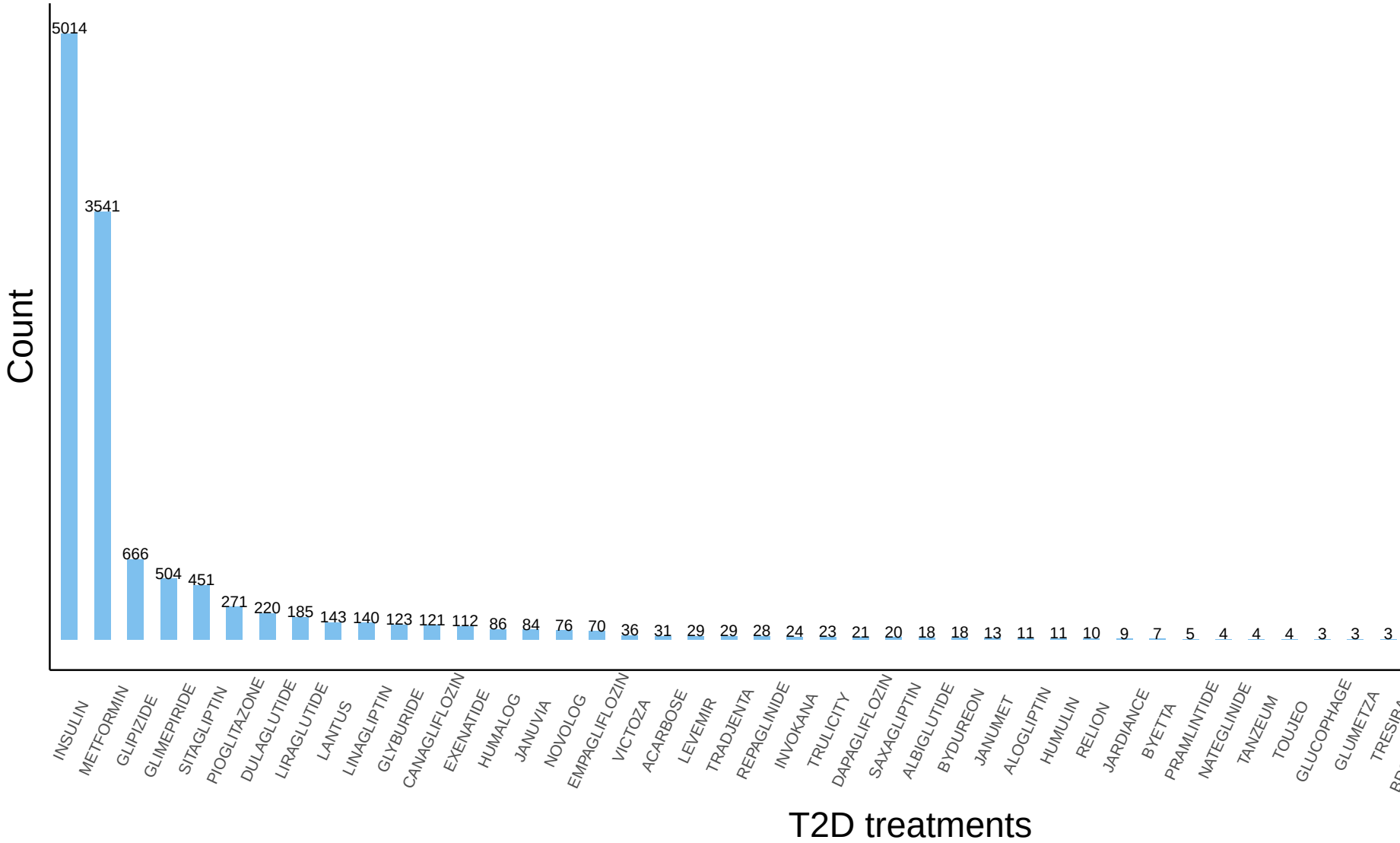
Challenges presented in EHRs

- Longitudinal biomarker measurements in EHRs are multivariate and have mixed modes or types: continuous (lab measurements), binary (diagnosis codes), counts (medications).
- Biomarker data are collected at each clinical encounter so they are irregular and sparse. Additionally, not all biomarkers are collected at the same time.
- When the measurements are taken is heterogeneous across patients and time patterns are potentially informative.
- Treatments are many and there are significant heterogeneity between patients receiving different treatments.

A snapshot of data measurements



Medication frequency pattern



Other problems formulated as missing data

- Finite population inference: nonsampled units are "missing"
- Measurement errors (Ex 1.15): true and measured variables, where true are missing or only observed for a calibration sample
- Disclosure limitation (Ex 1.16): replace some values by imputations to reduce disclosure risk
- Causal Inference under potential outcome framework
- Detection limit or Censored data

Data Augmentation: Complete-data problems formulated as missing data

**Some complete data can be "artificially" augmented with underlying unobserved data, allowing incomplete-data methods of analysis, e.g. EM algorithm

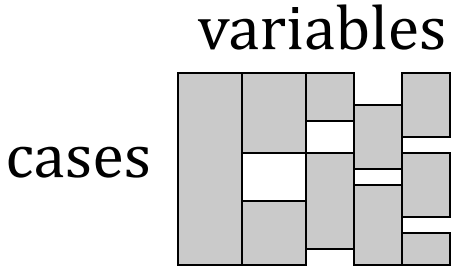
- Factor analysis (Ex 1.8) -- multivariate regression with unobserved regressors
- Mixed-effects models -- random effects are unobserved "missing" data
- Transformation models; interval censored data

Pattern Analysis versus Mechanism Analysis

Pattern Analsyis of Missing Data

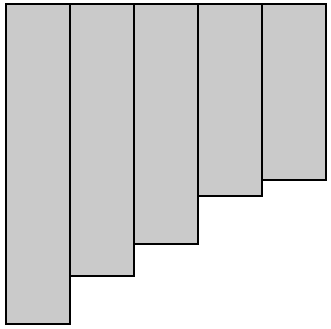
(analysis of observed data or descriptive analysis)

- Some methods work for a general pattern

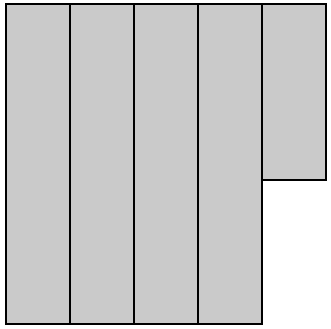


- Other methods apply only to special patterns

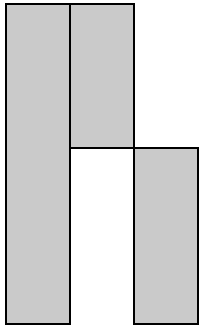
monotone



univariate



file matching



Pattern versus mechanism

- Pattern: Which values are missing?
- Mechanism: Why? Reasons related to the study variables?

  Y = data matrix, if no data were missing

  M = missing-data indicator matrix

  (i,j) th element indicates whether (i,j) th element of Y is missing (1) or observed (0)

  – Pattern concerns distribution of M

  – Mechanism concerns distribution of M given Y

More on mechanisms

- Missingness mechanism is:
  - missing completely at random (MCAR) if missingness independent of Y:

$$p(M \mid Y) = p(M) \text{ for all } Y$$

  - missing at random (MAR) if missingness only depends on observed components Y_{obs} of Y:

$$p(M \mid Y) = p(M \mid Y_{obs}) \text{ for all } Y$$

  - missing not at random (MNAR) if missingness depends on missing components of Y , after conditioning on observed data

MAR for univariate nonresponse

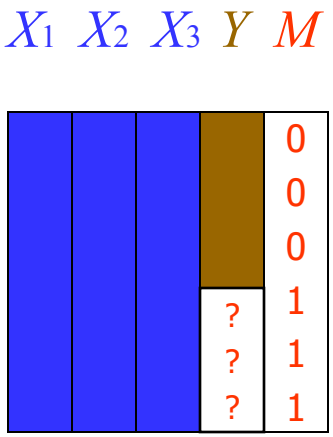
X_j = complete covariates

Y = incomplete variable

M = 1, Y missing

0, Y observed

$R = I - M$



MAR: missingness independent of Y given $X_1 \ldots X_k$

That is, M can depend on X's …

but not on Y given X's

MAR for monotone missing data

MAR if dropout depends on values recorded prior to drop-out

MNAR if dropout depends on values that are missing (that is, after drop-out)

Censoring by end of study: plausibly MCAR

Drop-out from side effect: MAR if side effect is measured and included in analysis

A non-monotone example

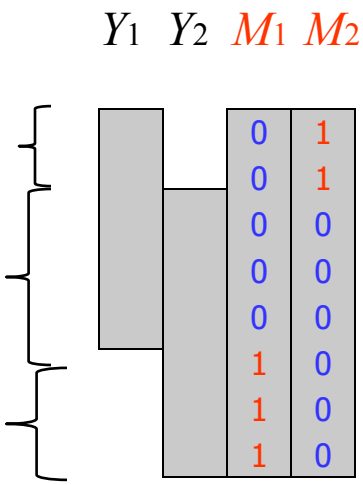
Mechanism is MAR if

$\Pr(Y_2 \text{ missing}) = g(Y_1)$

$\Pr(Y_1 \text{ missing}) = f(Y_2)$

$\Pr(\text{complete}) =$

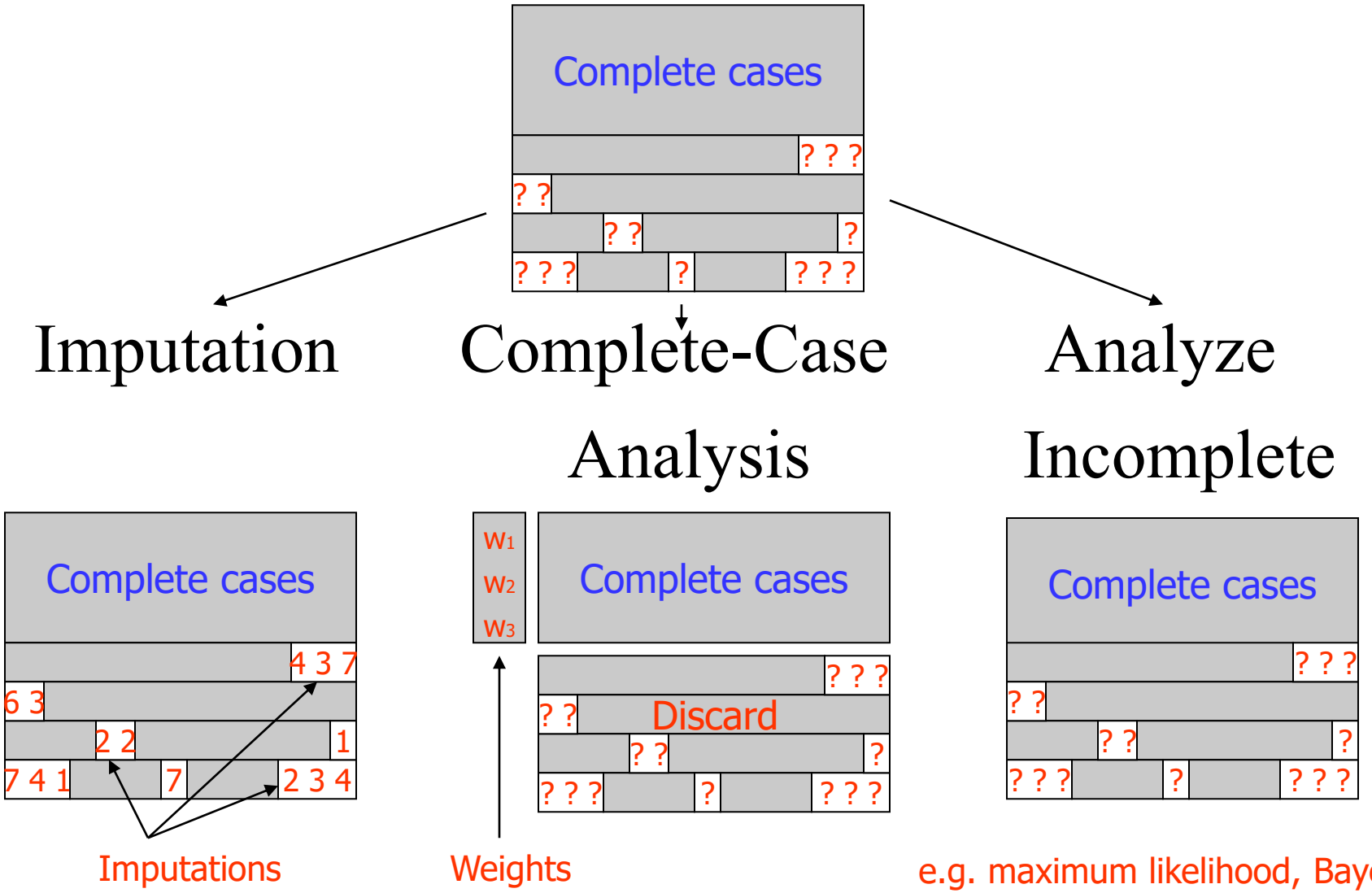
$1 - f(Y_2) - g(Y_1)$



Properties of a good missing-data method

- Makes use of partial information on incomplete cases, for reduced bias, increased efficiency
- Frequency valid ("calibrated") inferences under plausible model for missing data (e.g. confidence intervals have nominal coverage)
- Propagates missing-data uncertainty, both within and between imputation models
- Favor likelihood based approaches
  - Maximum Likelihood (ML) for large samples
  - Multiple Imputation/Bayes for small samples

General Strategies



History of Missing Data Analysis

Missing data methods in statistics -- history

1. Pre-1970s
   - Ad-hoc imputation/complete-case analysis
   - Maximum Likelihood (ML) for simple problems – factored likelihood (Anderson 1957)
   - ML for complex problems too hard
2. 1970's – mid 1980's: Maximum likelihood era
   - Modeling missing data mechanism, definition of MAR (Rubin 1976)
   - Better computing, EM (Dempster Laird and Rubin 1977) and extensions facilitate ML for complex problems
   - ML for various models beyond multivariate normal (LR 1987)

Missing data methods -- history

3. Mid 1980's – 1990's: Bayes and Multiple Imputation (MI)
   - Bayes via "data augmentation" for the multivariate normal problem (Tanner and Wong 1984)
   - Rubin proposes MI, justified via Bayes (1978, 1987)
   - MCMC facilitates Bayes and MI as an alternative to ML, with better small sample properties (see e.g. LR 2002)
4. 1990's – present: weakening assumptions, improving robustness
   - James Robins and colleagues propose Augmented Inverse-Probability Weighting (AIPW) methods for missing data (Robins & Rotnitsky, 1995)
   - Robust Bayesian models, more attention to model checks
   - more complex models – e.g. latent class models for categorical data, BART
   - Algorithmic methods – weakening or burying assumptions?