

Mission of VA Statisticians' Association (VASA)

- ▶ Promote & disseminate statistical methodological research relevant to VA studies;
- ▶ Facilitate communication & collaboration among VA-affiliated statisticians;
- ▶ Promote good statistical practice;
- ▶ Increase participation & visibility of VA statisticians at national meetings such as the VA HSR&D and ASA Joint Statistical Meetings;
- ▶ Increase participation of statisticians in VA research merit review

Poll

Your affiliation and title:

Your Research areas:

Your level of statistical knowledge: beginner, intermediate, and advanced

Topic of future statistical cyberseminar you like to take:

Need of statistical help in your research: No, Yes.

Problem of Missing Data

Standard statistical methods have been developed to analyze rectangular data sets.

- The rows are observation units

- The columns are variables.

- The entries are values (real numbers).

The concern is what happen when some of these values are not observed.

- Most statistical software creates special codes for the missing values

- Some statistical software exclude subjects with missing values - "complete-case analysis", which will be valid in a very limited cases.

Missing-data Pattern and Mechanism

Which method to choose for the analysis of missing data depends on missing-data patterns, as well as the missing-data mechanism.

Univariate Missing Data Pattern

	A	B	C
1	88	42	63
2	84	82	12
3	26	59	66
4	7	28	73
5	12	75	2
6	79	41	NA
7	81	84	NA
8	64	19	NA

This data set has three variables A, B and C, and the variable C has missing values, denoted by “NA”.

It can be seen that five of the eight subjects in the data set are completely observed, while the values of variable C of the other three subjects are missing.

This is the simplest missing data pattern.

Attrition in Longitudinal Studies

Attrition (dropout) in longitudinal studies is a very frequent occurrence, especially for long follow-ups or difficult study populations such as children.

If in a study, subjects drop out before the end of the study and do not return, then this scenario leads to the so called monotone missing data pattern.

Example on Attrition

A multi-clinic observational study on a prospective cohort of primary care patients with clinical depression.

The study evaluated depressive symptoms, mental and physical health for 966 clinically depressed persons from 6 large U.S. clinics.

These persons responded to several questionnaires that measured their physical and mental health at baseline, 6 weeks, 3 months, and nine months after baseline.

At 9 month after baseline each patient was interviewed by a psychiatrist who determined whether the patient still suffered from clinical depression after nine months.

Example on Attrition, continued

Unit non-response over time (X: observed, M=missing)

Baseline	6 weeks	3 months	nine months	Freq
X	X	X	X	759
X	X	X	M	92
X	X	M	M	27
X	M	M	M	2

File-matching Missing Data Pattern

Sometimes two or more variables are collected on units, but are never jointly observed.

A	B	C
45	34	NA
99	6	NA
16	84	NA
79	37	NA
38	95	NA
32	NA	5
48	NA	34
91	NA	25
NA	19	30
NA	20	31

File-matching Missing Data Pattern, Continued

It is important to know that for this missing data pattern, parameters relating to the association between these jointly unobservable variables (A, B, C) are not estimable from the data.

In the above data, there is no information in the data about the partial associations of B and C given A .

In practice, analyses of data with this pattern require to make strong assumptions about these partial associations.

General Missing Data Pattern)

A	B	C	D
25	26	88	32
NA	42	66	21
27	86	54	NA
28	NA	92	20
29	20	83	NA
30	89	NA	41
NA	NA	NA	35
32	NA	NA	33

Missing Data Mechanism

Understanding the missing data mechanism is important for the analysis of the missing data. Rational of using a method for dealing with a missing data set relies on the missingness mechanism.

To explain the missing data mechanisms, we first define a vector of indicator variables R that identify which variables are observed and which are missing. For example, $R = (0, 1)^\top$ indicates that the first variable is missing and the second variable is observed.

Assumption on Missing Data

Assumption: Missingness indicators hide true values that are meaningful for analysis.

Example on Death

Consider a randomized trial with two treatments, and suppose that a primary outcome of the study is a "quality-of-life health score" Y measured one year after randomization to treatment. For participants who die within a year of randomization to treatment, Y is not available due to death.

Poll

Do you consider patients who died before the end of the study as missing "quality-of-life health score"?

Yes

No

Not Sure

Example on Death, Continued

Do you consider patients who died before the end of the study as missing "quality-of-life health score"?

It does not make sense to treat those outcomes as missing, given that quality of life is meaningless concept for people who not alive.

Example on Non-response in Opinion Polls

We are interested in polling individuals about how they will vote in a future referendum, where the available responses are "yes", "no", or "missing".

Poll

Do you consider patients who with "missing" as missing response?

Yes

No

Not Sure

Example, Continued

Individuals who fail to respond to the question may be refusing to reveal real answers or may have no interest in voting.

Assumption would not apply to individuals who would not vote, and these individuals define a stratum of the population that is not relevant to the outcome of the referendum.

Assumption would apply to individuals who do not respond to the initial poll but would vote in the referendum. For these individuals it would make sense to treat their responses on the referendum as missing.

Missing Data Mechanism

Since some of the variables Y in the data set may be missing, we partition Y into two parts, Y_{obs} and Y_{mis} , the observed part and the missing part.

Below we introduce the three missing data mechanisms, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Missing Completely at Random (MCAR)

The data are called MCAR if missingness does not depend on the value of data Y , missing or observed. Statistically, that is

$$f(R|Y) = f(R) \text{ for all } Y. \quad (1)$$

The consequence of the above relationship means that missingness indicator is conditionally independent of the study variables, so data on the study variables, regardless of missing or observed, have nothing to do with the mechanism that the data are missing.

Example on MCAR

Consider an experiment involving random sampling. Under MCAR people with missing values are really just a random sample of the study population.

As a result, people with observed data still make up a valid random sample from the study population for the ultimate inference, though with some loss in efficiency because of the smaller sample size.

Therefore when data are MCAR, you can ignore the units with missing values and proceed with the complete data analysis, and still make valid inference.

Missing at Random (MAR)

MAR is a less restrictive assumption on missing data mechanism than MCAR. Data are called MAR when missingness depends only on the observed components Y_{obs} of Y , but not on the missing components Y_{mis} . Statistically, that is

$$f(R|Y) = f(R|Y_{obs}) \text{ for all } Y_{mis}. \quad (2)$$

Despite its name, MAR does not mean that the missing data are a simple random sample of all data values, which is the case for MCAR.

MAR lessens MCAR in that the missing values can behave like a random sample of all values within subpopulations defined by the observed data rather than the entire data values.

Example on MAR

In many VA studies, race is a the key variable but missing substantially.

If the reason for missing race is only related to age of a veteran, and if age is observed for all study subjects, the MAR holds.

Missing not at random (MNAR)

The mechanism is called missing not at random (MNAR) if the distribution of R depends on the missing values in Y .

Strictly speaking, MNAR comes in two forms: (i) missingness depends on the missing value itself or (ii) missingness depends on an unobserved variable.

The latter is common in latent variable models or hidden Markov models.

Example on MNAR

If the reason for missing race is related to race itself or some unmeasured covariates, the MAR does not hold.

Illustration Example

We demonstrate the three mechanisms and their consequences through the following simple example.

Consider a simple linear relationship between two variables x and y , where the predictor variable x is fully observed, but some of y are missing according to different missing mechanisms.

The variables x and y follow the linear regression $y = x + \varepsilon$, where x and ε independently follows standard normal distributions.

We generate an independent sample of (x, y) with size 300. The sample is referred to as the full data, because all values of x and y are known. Then three samples with y -missing data are generated from the full data, according to the missing mechanisms MCAR, MAR and MNAR, respectively.

Result

	Full data	MCAR	MAR	MNAR
(Intercept)	0.08(0.06)	0.01(0.08)	-0.07(0.13)	0.44(0.09)
x	1.02(0.06)	1.01(0.08)	1.15(0.12)	0.70(0.09)
R ²	0.53	0.54	0.36	0.23
Adj. R ²	0.53	0.53	0.35	0.23
Num. obs.	300	148	163	201

Methods for Dealing with Missing Data- Complete data Analysis

As the name suggests, the complete-case method (also known as list-wise deletion) makes use of cases with complete data on all variables of interest in the analysis.

Cases with incomplete data on one or more variables are discarded. This approach is easiest to implement since standard full data analysis can be applied without modification, and is the default in many statistical software (i.e., cases with any missing values are automatically ignored).

The disadvantages include loss of precision due to a smaller sample size, and bias in complete-data summaries when the cases with missing data differ systematically from the completely observed cases (i.e., when data are not missing completely at random).

Weighted Complete Data Method

The earlier complete-case strategy can potentially lead to biased summaries because the sample of observed cases may not be representative of the full sample.

Another strategy reweights the sample to make it more representative. For example, if response rate is twice as likely among men as women, data from each man in the sample could receive a weight of 2 in order to make the data more representative.

This strategy is commonly used in sample surveys, especially in the case of unit nonresponse, where all the survey items are missing for cases in the sample that did not participate.

This method usually requires the MAR assumption.

Maximum Likelihood (ML) Method under MAR

The fundamental idea behind the ML methods is conveyed by its name: find the values of the parameters that are most probable, or most likely, for the data that have actually been observed.

Denote D_{obs} as the observed data, which have the probability density $p(D_{obs}|\theta)$. Here θ is a set of the parameters of interest. Then the *likelihood function* $L(\theta|D_{obs})$ is equal to $p(D_{obs}|\theta)$, i.e. $L(\theta|x) = p(D_{obs}|\theta)$, and is thought of as a function of θ where the data D_{obs} are fixed.

Maximum Likelihood Method under MAR, continued

The next step is to deduce a numerical value for θ using our knowledge of $L(\theta|x)$. The ML principle says we should choose an estimate of θ that maximizes the likelihood function.

In other words, we choose a value of the parameter that best explains the observed data, and this value is called the *ML estimate* of θ .

Calculation of the ML estimates

Finding the maximum can be easy or hard, depending on the form of $p(D_{obs}|\theta)$. For the normal (gaussian) distribution, θ consists of the mean μ and variance σ^2 , and we can find the derivative of the $\log(L\theta|D_{obs})$, set it equal to zero, and solve directly for μ and σ^2 .

But for most problems, these analytic expressions are hard to come by, and we have to use more elaborate techniques. One might use Newton-Raphson or quasi-Newton algorithms to directly maximize the likelihood of the observed data. One of the most popular techniques for maximizing the likelihood function is the expectation-maximization (EM) algorithm.

Imputation-based Methods

implicit modeling methods

- ▶ Hot deck imputation: replacing missing values by values from "similar" responding units in the sample.
- ▶ Cold deck imputation: replace a missing value of an item by a constant value from an external source.

Model-based procedures: define a model for the observed data and base imputation on a draw from the posterior distribution of missing data conditional on observed data under that model.

Single Imputation

Single imputation substitutes some reasonable guess (imputation) for each missing value and then perform the analysis as if there were no missing data.

Mean imputation: A popular approach replaces each missing value with the unconditional mean of the observed values for a particular variable. This single-imputation approach can lead to underestimating standard errors. Additionally, this method distorts relationships between variables since it pushes the correlations between variables towards zero.

Single Imputation, continued

Conditional mean imputation: For data sets with few missing variables, conditional mean imputation is an improvement on the unconditional mean approach since it replaces each missing value with the conditional mean of the variable based on other fully observed variables in the data set.

Regression imputation: replace missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present. Mean imputation is a special case.

Stochastic regression imputation: replace missing values by a value predicted by regression imputation plus a residual, drawn to reflect uncertainty in the predicted value.

Nearest Neighbor Hot Deck on imputing Missing Outcome

Define a metric to measure distance between units, based on the values of observed covariates

Choose imputed values that come from responding units close to the units with missing value.

Nearest Neighbor Hot Deck, Cont

Let $X = (X_1, \dots, X_K)$ be K covariates that are observable for all units, and let $x_i = (x_{i1}, \dots, x_{iK})^T$ be the values of K covariates for a unit i for which y_i is missing.

Let $d(i, j)$ be the metric between two covariates x_i and x_j of unit i and unit j .

Nearest Neighbor Hot Deck, Cont

For the unit i , we choose an imputed value for y_i from those units j that are such that (1) $y_j, x_{j1}, \dots, x_{jK}$ are observed, and (2) $d(i, j)$ is less than some value d_0 .

Comments on Imputation

Imputation should generally be

Conditional on observed variable, to reduce bias due to response, improve precision, and preserve association between missing and observed variables.

Multivariate, to preserve associations between missing variables.

Draws from the predictive distribution rather than means.

Account for imputation uncertainty \rightarrow multiple imputation.

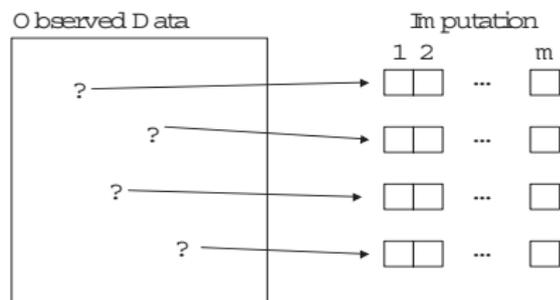
Multiple Imputation, cont

Imputation: Create D imputations of the missing data, $Y_{mis}^{(1)}, \dots, Y_{mis}^{(d)}$, under a suitable model.

Analysis: Analyze each of the m completed datasets in the same way.

Combination: Combine the m sets of estimates and SE's using Rubin's (1987) rules.

Picture Illustration



Software for Multiple Imputation - SAS

PROC MI implements multiple imputation by the multivariate normal model.

PROC MIANALYZE takes the results of the complete-data analysis per dataset.

IVEware is a SAS callable software application that implements multiple imputation using chained equations, called sequential regressions.

Software for Multiple Imputation - STATA

The ice package: a user-contributed Stata package that provides multiple imputation by chained equations.

Software for Multiple Imputation - SPSS

Command, MULTIPLE IMPUTATION: a part of the Missing Values module, supports multiple imputations by chained equations.

Imputation and analysis can be done in a largely automatic fashion and is well integrated with the software for complete-data analysis.

Software for Multiple Imputation - R

Amelia: multiple imputations based on the multivariate normal model.

BaBooN: multiple imputations by chained equations.

cat: multiple imputation of categorical data according to the log-linear model.

kmi: a Kaplan-Meier multiple imputation, specifically designed to impute missing censoring times.

mi: a chained equations approach based on Bayesian regression methods.

mice: MI by the chained equations.