

THE FUTURE OF
CLINICAL DEVELOPMENT



PRAHEALTHSCIENCES



US 2018

The Clinical Data Science Conference

June 3rd-6th
Raleigh Convention Center
Raleigh, North Carolina, USA

DH04

Missing Data Imputation and Its Effect on the Accuracy of Classification

Edwin Ponraj Thangarajan, PRA Health Sciences, Chennai, India

Giri Balasubramanian, PRA Health Sciences, Chennai, India

04-Jun-2018



Multivariate data sets frequently have missing observations scattered throughout the data set. Many machine learning algorithms assume that there is no significance in the fact that an observation has an attribute value missing. A common approach in coping with these missing values is to replace the missing value using some plausible value, and the resulting completed data set is analyzed using standard methods. Programs evaluate the effect that some commonly used imputation methods have on the accuracy of classifiers in supervised learning. The effect is assessed in simulations performed on several classical datasets where observations have been made missing at random in different proportions. This paper will describe the below,

- 📄 Missing Data Mechanisms
- 📄 Strategies of Handling Missing Data
- 📄 Method Used to Deal with Missing Values



Acknowledgments

I would like to express my special
thanks of all authors in the references
slide.





Introduction

- Missing Data
- Example of misrepresentation of statistical data
- Missing Data Imputation

Missing Data Mechanisms

- MCAR
- MAR
- NMAR

Strategies of Handling Missing Data

- Complete Case Analysis
- Available Case Analysis
- Weighting Procedures
- Imputation Procedures

Method Used to Deal with Missing Values

- Examine Patterns of Missing Data with SAS Examples
- Mean and Median Imputation – PROC STDIZE
- Hot Deck Imputation – PROC SURVEYIMPUTE
- Multiple Imputation – PROC MI
- Maximum Likelihood imputation – PROC MI



Missing Data

Case	Age	Gender	Home	Education	Occupation
1		Female	No	16	Non-professional
2	22	Male	No		Non-professional
3	39	Male		20	Professional
4		Female	Yes		Professional
5	40		Yes	16	Non-professional
6	22	Female	No	16	
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional



In this data set, data are missing for some respondents as shown by the color **Orange Lighter**. Any case with either missing data in age, home or education is ignored in the analysis. So, only respondents 5-8 would be included in your analysis which would drastically reduce the amount of data to be analyzed and increase the risk of a costly error.

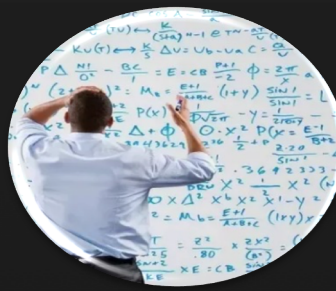


Misrepresentation of Statistical Data

Example



A statistical report: "The number of accidents taking place in the middle of the road is much less than the number of accidents taking place on its side. Hence it is safer to walk in the middle of the road." This conclusion is obviously wrong since we are not given the proportion of the number of accidents to the number of persons walking in the two cases



"The number of students taking up Mathematics Honors in a University has increased 5 times during the last 3 years. Thus, Mathematics is gaining popularity among the students of the university." Again, the conclusion is faulty since we are not given any such details about the other subjects and hence comparative study is not possible.



"75% of the people who drink alcohol die before attaining the age of 80 years. Hence drinking is harmful for longevity of life." This statement, too, is incorrect since nothing is mentioned about number of persons who do not drink alcohol and die before attaining the age of 80 years.

"INCOMPLETE DATA USUALLY LEADS US TO FALLACIOUS CONCLUSIONS."



Case	Age	Gender	Home	Education	Occupation
1	21	Female	No	16	Non-professional
2	22	Male	No	16	Non-professional
3	39	Male	Yes	20	Professional
4	20	Female	Yes	16	Professional
5	40	Female	Yes	16	Non-professional
6	22	Female	No	16	Non-professional
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional



With many classification algorithms, a common approach that is used is to replace the missing values in the data set with some plausible value and the resulting completed data set is analyzed using standard algorithms. The procedure that replaces the missing values using some value is known as imputation. For a layman, this idea conjures an image of a statistician making up the data. This is true only if one were to analyze the data as if the imputed data are real values.



Missing Data Mechanisms





Varied solution and imputation methods were developed for the missing data problem. It is needed to identify the missing data mechanism before such solution and imputation methods. Because the decision which solution and imputation method would be applied to the data set depends upon the missing data mechanisms.

MCAR

- Missing completely at random (MCAR).
- If missing observations is independent of both observed and unobserved values

MAR

- Missing at random (MAR).
- If it is independent of unobserved values and dependent of observed values

NMAR

- Not missing at random (NMAR).
- If it depends on both observed and unobserved values

Knowledge of the mechanism that led to the values being missing is important in choosing an appropriate analysis to use for the data. Hence it is important to consider how the classifier handles the missing data to avoid bias being introduced into the knowledge induced from that classifier.



Strategies of Handling Missing Data





**Complete
Case
Analysis**

Any missing attributes are deleted from the data set

**Available
Case
Analysis**

Uses all observations that have values for a particular attribute, there is no loss of information as all cases are used.

**Weighting
Procedures**

Non-response modify the weights in an attempt to adjust for non-response as if it were part of the sample design

**Imputation
Procedures**

Missing data values are replaced with some value is another commonly used strategy for dealing with missing value. .



Method Used to Deal with Missing Values



With SAS Examples



- To find the number of missing values for numeric variables PROC MEANS procedure can be used in SAS
- PROC MEANS creates a compact easy-to-read table that summarizes the number of missing values for each numerical variable.
- N and NMISS options in the PROC MEANS statement to count the number of missing values

```
/* count missing values for numeric variables */  
proc means data=Sashelp.Heart nolabels N NMISS;  
    var AgeAtStart Diastolic Systolic Height Weight MRW Smoking  
        Cholesterol;  
run;
```



Examine Patterns of Missing Data

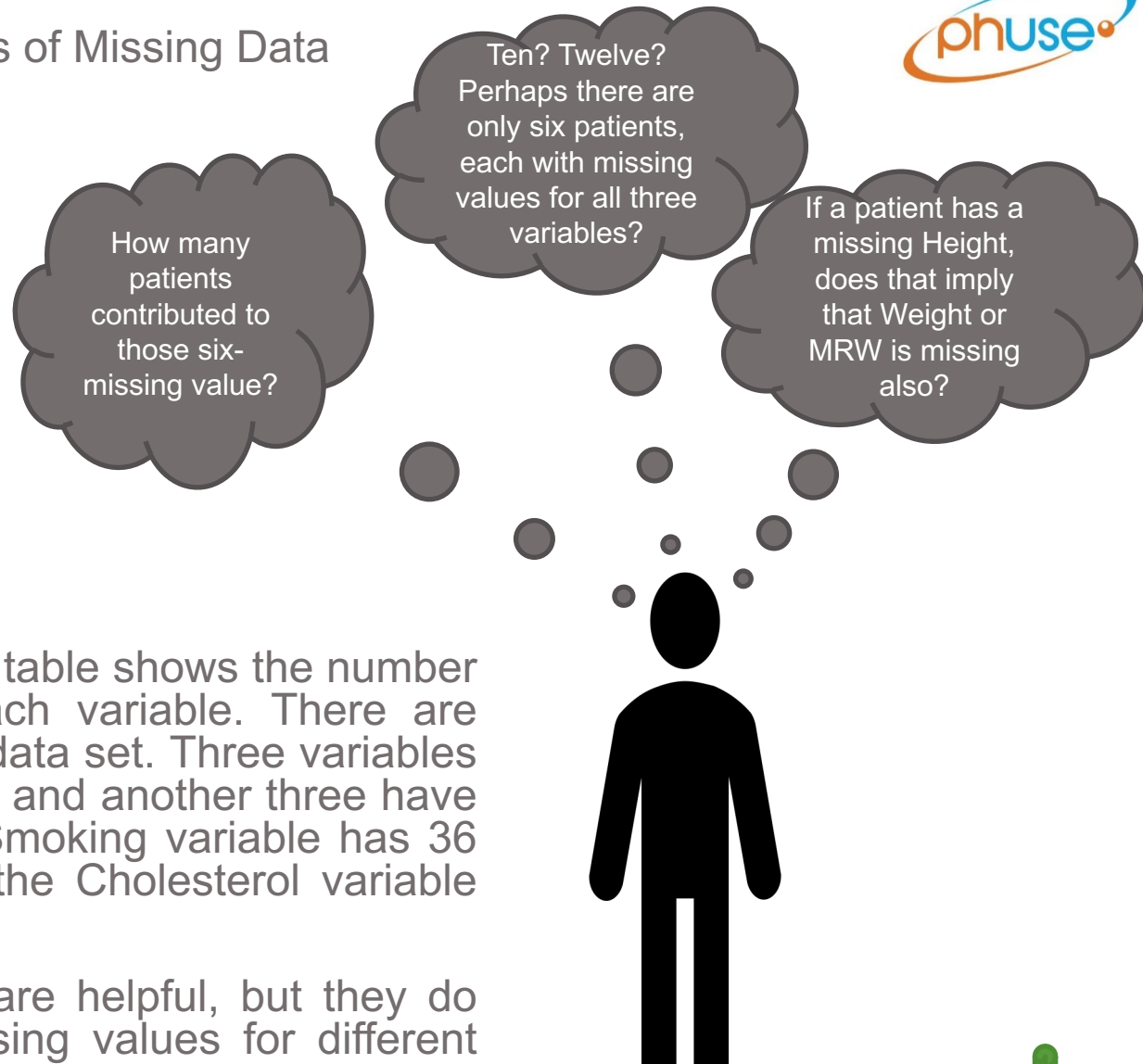
PROC MEANS

The MEANS Procedure

Variable	N	N Miss
AgeAtStart	5209	0
Diastolic	5209	0
Systolic	5209	0
Height	5203	6
Weight	5203	6
MRW	5203	6
Smoking	5173	36
Cholesterol	5057	152

The NMISS column in the table shows the number of missing values for each variable. There are 5209 observations in the data set. Three variables have zero missing values, and another three have six missing values. The Smoking variable has 36 missing values whereas the Cholesterol variable has 152 missing values.

These univariate counts are helpful, but they do not tell you whether missing values for different variables are related.



To answer these questions, you must look at the pattern of missing values



- The MI procedure in SAS is used for multiple imputation of missing values.
- PROC MI has an option to produce a table that summarizes the patterns of missing values among the observations.
- The following call to PROC MI uses the NIMPUTE=0 option to create the "Missing Data Patterns" table for the specified variables

```
ods select MissPattern;  
proc mi data=Sashelp.Heart nimpute=0;  
    var AgeAtStart Height Weight Diastolic Systolic MRW Smoking  
        Cholesterol;  
run;
```




Examine Patterns of Missing Data

PROC MI



5039 observations that have no missing values

The number of missing values for variables, pairs of variables, triplets of variables, and so forth

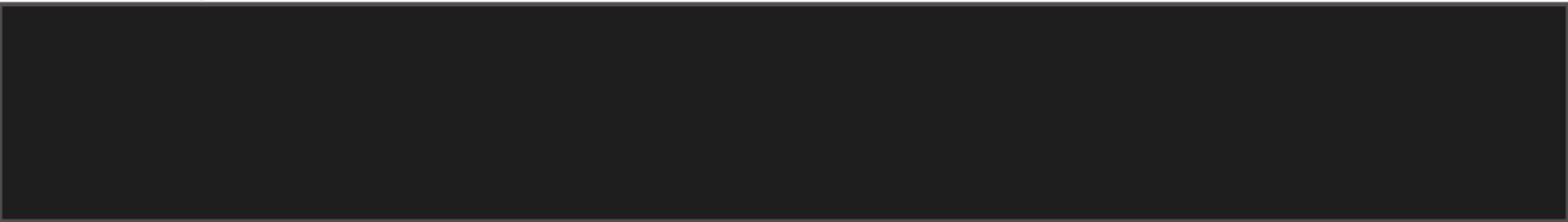
The MI Procedure

Missing Data Patterns

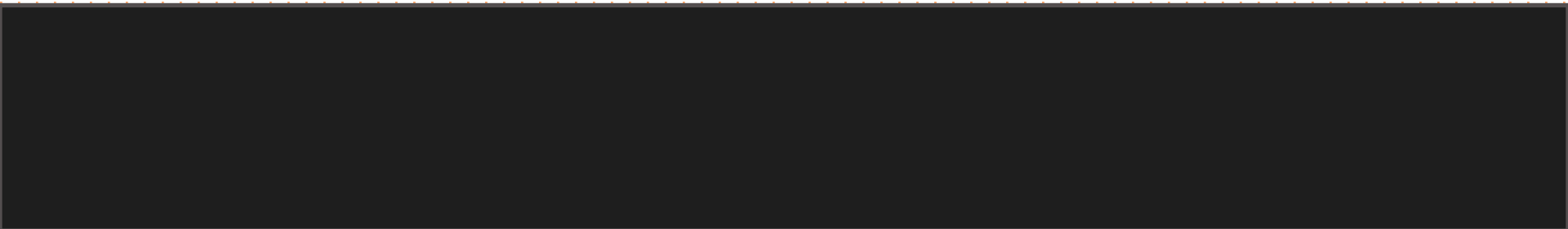
Group	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol	Freq	Percent	Group Means							
											AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
1	X	X	X	X	X	X	X	X	5039	96.74	44.073030	64.826206	153.060131	85.411193	136.987498	119.864457	9.370907	227.410796
2	X	X	X	X	X	X	X	.	124	2.38	43.653226	64.137097	154.193548	83.459677	134.040323	124.387097	8.685484	.
3	X	X	X	X	X	X	.	X	8	0.15	42.875000	65.312500	158.125000	83.500000	131.625000	121.750000	.	210.125000
4	X	X	X	X	X	X	.	.	28	0.54	45.214286	65.482143	151.392857	87.714286	140.571429	115.928571	.	.
5	X	X	.	X	X	.	X	X	4	0.08	40.000000	63.687500	.	73.500000	126.500000	.	27.500000	256.500000
6	X	.	X	X	X	X	X	X	4	0.08	55.000000	.	154.000000	81.250000	129.000000	124.500000	10.250000	254.250000
7	X	.	.	X	X	.	X	X	2	0.04	34.000000	.	.	77.000000	125.000000	.	2.500000	201.500000

The table reports the number of observations that have a common pattern of missing values

- Most patients (96.7%) have complete data.
- The most likely measurement to be missing is Cholesterol, followed by information about whether the patient smokes.
- You can see exactly how many patients are missing Height measurements, Weight measurements, or both.
- It is obvious that the MRW variable has a missing value if and only if the Weight variable is missing.



Assumptions and Patterns of *missingness* are used to determine *which methods* can be used to deal with missing data.





Method Used to Deal with Missing Values



Mean and Median
Imputation



- Commonly Used
- Ignores any relationship between variables
- Underestimate the variance covariance matrices for that data
- Mean and median imputation can only be used on continuous attributes
- For categorical data, the mode is often imputed whilst using either mean or median imputation.



- The easiest way to perform mean imputation in SAS is to use PROC STDIZE.
- PROC STDIZE supports the REONLY and the METHOD=MEAN options, which tells it to replace missing values with the mean for the variables on the VAR statement.

PROC STDIZE

<options>;

BY variables; You can specify a BY statement with PROC STDIZE to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

FREQ variable; If one variable in the input data set represents the frequency of occurrence for other values in the observation, specify the variable name in a FREQ statement. PROC STDIZE treats the data set as if each observation appeared n times, where n is the value of the FREQ variable for the observation. Nonintegral values of the FREQ variable are truncated to the largest integer less than the FREQ value. If the FREQ variable has a value that is less than 1 or is missing, the observation is not used in the analysis.

LOCATION variables; The LOCATION statement specifies a list of numeric variables that contain location measures in the input data set specified by the METHOD=IN option

SCALE variables; The SCALE statement specifies the list of numeric variables that contain scale measures in the input data set specified by the METHOD=IN option.

VAR variables; The VAR statement lists numeric variables to be standardized. If you omit the VAR statement, all numeric variables not listed in the BY, FREQ, and WEIGHT statements are used.

WEIGHT variable; The WEIGHT statement specifies a numeric variable in the input data set with values that are used to weight each observation. Only one variable can be specified.



Mean and Median Imputation

PROC STDIZE



```
/* Create "original data" by randomly inserting missing values for some heights */  
data Have;  
    set sashelp.class;  
    call streaminit(12345);  
    Replaced = rand("Bernoulli", 0.4); /* indicator variable is 1 about 40% of time */  
    if Replaced then Height = .;  
run;  
  
/* Mean imputation: Use PROC STDIZE to replace missing values with mean */  
proc stdize data=Have out=Imputed  
    oprefix=Orig_ /* prefix for original variables */  
    reponly /* only replace; do not standardize */  
    method=MEAN; /* or MEDIAN, MINIMUM, MIDRANGE, etc. */  
    var Height; /* you can list multiple variables to impute */  
run;  
  
proc print data=Imputed;  
    format Orig_Height Height BESTD8.1;  
    var Name Orig_Height Height Weight Replaced;  
run;
```

The METHOD= option in PROC STDIZE supports several statistics. You can use METHOD=MEDIAN to replace missing values by the median, METHOD=MINIMUM to replace by the minimum value, and so forth.



Mean and Median Imputation PROC STDIZE



Obs	Name	Orig_Height	Height	Weight	Replaced
1	Alfred	69	69	112.5	0
2	Alice	56.500	56.500	84.0	0
3	Barbara	65.300	65.300	98.0	0
4	Carol	62.800	62.800	102.5	0
5	Henry	63.500	63.500	102.5	0
6	James	.	61.500	83.0	1
7	Jane	59.800	59.800	84.5	0
8	Janet	.	61.500	112.5	1
9	Jeffrey	62.500	62.500	84.0	0
10	John	59	59	99.5	0
11	Joyce	51.300	51.300	50.5	0
12	Judy	64.300	64.300	90.0	0
13	Louise	.	61.500	77.0	1
14	Mary	66.500	66.500	112.0	0
15	Philip	.	61.500	150.0	1
16	Robert	.	61.500	128.0	1
17	Ronald	.	61.500	133.0	1
18	Thomas	57.500	57.500	85.0	0
19	William	.	61.500	112.0	1

- The output shows that the missing data (such as observations 6 and 8) are replaced by 61.5, which is the mean value of the observed heights.
- For a subsequent visualization, a binary variable (Replaced) that indicates whether an observation was originally missing.



- Mean imputation, which is easy to implement, enables analysts to use every observation.
- However, mean imputation has three serious disadvantages that can lead to problems in your statistical analysis.
- Mean imputation is a univariate method that ignores the relationships between variables and makes no effort to represent the inherent variability in the data.
- In particular, when you replace missing data by a mean, you commit three statistical sins:
 - Mean imputation reduces the variance of the imputed variables.
 - Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
 - Mean imputation does not preserve relationships between variables such as correlations.



Method Used to Deal with Missing Values



Hot Deck
Imputation



- Commonly used, especially in survey samples
- It can cope with both continuous and categorical attributes
- Replacing the missing values using values from one or more similar instances that are in the same classification group



PROC SURVEYIMPUTE

<options>;

The
SURVEYIMPUTE
SAS procedure
provides several
imputation
methods for
replacing missing
values in an item
by observed
values from the
same item

These imputation
methods can be
used to impute
missing values for
survey data.

BY variables; You can specify a BY statement with PROC SURVEYIMPUTE to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

CELLS variables; The CELLS statement names the variables that identify the imputation cells. The imputation cells divide the data into groups of similar units. The combination of levels of CELLS variables defines the imputation cells. If you do not use this statement, then all observation units are assumed to be in one imputation cell.

CLASS variable < (options) > . . . < variable < (options) > > / options >; The CLASS statement names the classification variables for the analysis.

CLUSTER variables; The CLUSTER statement names variables that identify the first-stage clusters in a clustered sample design. First-stage clusters are also known as primary sampling units (PSUs). The combinations of categories of CLUSTER variables define the clusters in the sample. If you also use the STRATA statement, clusters are nested within strata.

ID variable; The ID statement names a variable in the DATA= input data set to identify observation units.

IMPJOINT <variables>; The IMPJOINT statement specifies the names of variables that are to be imputed jointly for the fully efficient fractional imputation (FEFI) method.

OUTPUT <OUT=SAS-data-set ><OUTJKCOEFS=SAS-data-set >< keyword=name...keyword=name >; The OUTPUT statement creates a SAS data set that contains the imputed data. You must use the OUTPUT statement to store the imputed data in a SAS data set.

REPWEIGHTS variables; The REPWEIGHTS statement names variables that provide replicate weights.

STRATA variables; The STRATA statement names one or more variables that form the strata in a stratified sample design. The combinations of levels of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently.

VAR variables; The VAR statement names the analysis variables to be imputed.

WEIGHT variable; The WEIGHT statement names the variable that contains the sampling weights.



Hot Deck Imputation

PROC SURVEYIMPUTE [Bayesian Bootstrap method as an example]



- Using a simulated data set from a fictitious survey of drug abusers
- A stratified clustered sample of drug abuse treatment centers is taken from a list of available treatment centers.
- The list is first stratified based on geographic locations.
- From each strata, two or three treatment centers are sampled as the primary sampling units (PSU).
- Data are collected from individual patients within the selected treatment centers
- The survey collects information about the substances that the patients used (such as drugs, alcohol, and marijuana) along with insurance information and treatment information.
- The data set contains 736 observation units in 35 PSUs and 10 strata.
- The sum of the weights is 19,600. Therefore, the survey data represent a population of 19,600 patients from the study area. Some participants did not respond to all questions. The data set contains missing values in many variables.

```
data DrugAbuse;
  input Strata PSU ObsWeight ImputationCell Age Sex Race Insurance
        Drug Alcohol Treatment;
  datalines;
1 1 5 1 74 1 1 3 2 2 1
1 1 5 1 20 0 3 1 2 2 1
1 1 5 3 42 1 2 1 1 2 1
1 1 5 3 65 1 3 2 1 2 1
1 1 5 2 53 1 1 1 1 1 1
1 1 5 3 49 1 1 1 1 2 1
1 1 5 3 51 0 2 1 2 1 1
1 1 5 2 77 0 3 1 1 1 1
1 1 5 2 26 1 1 1 1 2 1
1 1 5 3 28 0 3 1 1 1 1
1 1 5 1 71 1 1 1 2 2 1
1 1 5 2 72 1 1 3 2 2 1
1 1 5 3 24 1 1 1 1 1 1
1 1 5 2 65 1 1 2 1 1 2
1 1 5 3 47 1 1 1 1 1 1
1 1 5 2 37 1 1 2 1 2 1
1 1 5 2 46 1 1 3 1 1 1
1 1 5 2 52 1 1 1 1 2 2
1 1 5 3 60 0 3 1 1 2 1
1 1 5 1 31 0 1 1 1 2 1
1 2 5 1 23 0 3 3 1 1 1
1 2 5 2 78 1 1 . . . .
1 2 5 2 29 1 1 1 1 1 1
1 2 5 2 21 . . . . . .
... more lines ...
10 4 55.5556 1 40 0 3 2 1 1 2
10 4 55.5556 1 32 1 3 1 2 2 1
10 4 55.5556 3 68 0 1 2 2 1 2
10 4 55.5556 3 35 1 1 2 1 2 2
;
run;
```



Hot Deck Imputation

PROC SURVEYIMPUTE [Bayesian Bootstrap method as an example]



```
proc surveyimpute data=DrugAbuse method=hotdeck(selection=abb)
    ndonors=5 seed=773269;
    var Sex Race Insurance Drug Alcohol Treatment;
    cells ImputationCell;
    output out=DrugAbuseABB;
run;
```

Identifies the imputation cell variable ImputationCell

Requests the approximate Bayesian bootstrap method,

Requests five donor units for every recipient unit

- To impute the missing items, you first need to decide whether to impute within imputation cells.
- Imputation cells divide the data into groups of similar units such that the recipient units share similar characteristics with the donor units in the same group.
- For example, it is reasonable to believe that different age groups, races, and income categories might have different responses to the drug abuse survey.
- You can use these characteristics to create imputation cells.
- Characteristics for imputation cells might come from the same survey or might come from other sources such as census data or previous surveys.
- In this example, assume that the imputation cells are available as a variable called ImputationCell in the data set.



Hot Deck Imputation

PROC SURVEYIMPUTE [Bayesian Bootstrap method as an example]



SURVEYIMPUTE Procedure Output

The SURVEYIMPUTE Procedure

Imputation Information	
Data Set	WORK.DRUGABUSE
Imputation Method	HOTDECK
Selection Method	ABB
Random Number Seed	773269

Number of Observations Read	736
Number of Observations Used	736

Missing Data Patterns															
Group	Sex	Race	Insurance	Drug	Alcohol	Treatment	Freq	Sum of Weights	Unweighted Percent	Weighted Percent	Group Means				
											Sex	Race	Insurance	Drug	Treatment
1	X	X	X	X	X	X	681	681	92.53	92.53	0.566814	1.491924	1.718062	1.292217	1.716593
2	X	X	X	X	X	.	34	34	4.62	4.62	0.500000	1.441176	1.588235	1.294118	1.588235
3	X	X	.	.	.	.	13	13	1.77	1.77	0.692308	1.230789	.	.	.
4	.	.	.	.	.	.	8	8	1.09	1.09	.	.	.	.	.

Imputation Summary		
Observation Status	Number of Observations	Sum of Weights
Nonmissing	681	681
Missing	55	55
Missing, Imputed	55	55
Missing, Not Imputed	0	0



Hot Deck Imputation

PROC SURVEYIMPUTE [Bayesian Bootstrap method as an example]



Unit Identification: The recipient index, and all the variables from the input data set DrugAbuse

UnitID	Recipient	Strata	PSU	ObsWeight	ImputationCell	Age	Sex	Race	Insurance	Drug	Alcohol	Treatment
20	0	1	1	5	1	31	0	1	1	1	2	1
21	0	1	2	5	1	23	0	3	3	1	1	1
22	1	1	2	5	2	78	1	1	1	1	1	1
22	2	1	2	5	2	78	1	1	1	1	1	1
22	3	1	2	5	2	78	1	1	2	1	2	1
22	4	1	2	5	2	78	1	1	3	1	2	1
22	5	1	2	5	2	78	1	1	1	1	2	2
23	0	1	2	5	2	29	1	1	1	1	1	1
24	1	1	2	5	2	21	1	2	2	1	2	2
24	2	1	2	5	2	21	1	3	1	1	1	1
24	3	1	2	5	2	21	1	2	2	1	2	1
24	4	1	2	5	2	21	1	1	2	1	2	1
24	5	1	2	5	2	21	1	2	2	2	1	1
25	0	1	2	5	3	85	0	1	3	1	2	1

Units that are complete respondents have one row. unit 21 is a complete respondent, so it has only one row in the output data set and its Recipient value is 0

Units that are incomplete respondents have five rows in the output data set. Unit 22 is an incomplete respondent, so it has five rows in the output data set and its Recipient values range from 1 to 5.



Method Used to Deal with Missing Values



Multiple
Imputation



- Multiple Imputation (MI) is not simply a technique for imputing missing data
- Obtaining estimates and correct inferences for statistics ranging from simple descriptive statistics to the parameters of complex multivariate models
- Solves the limitations of single imputation by introducing an additional form of error based on variation in the parameter estimates across the imputation, which is called “between imputation error”.
- It replaces each missing item with two or more acceptable values, representing a distribution of possibilities

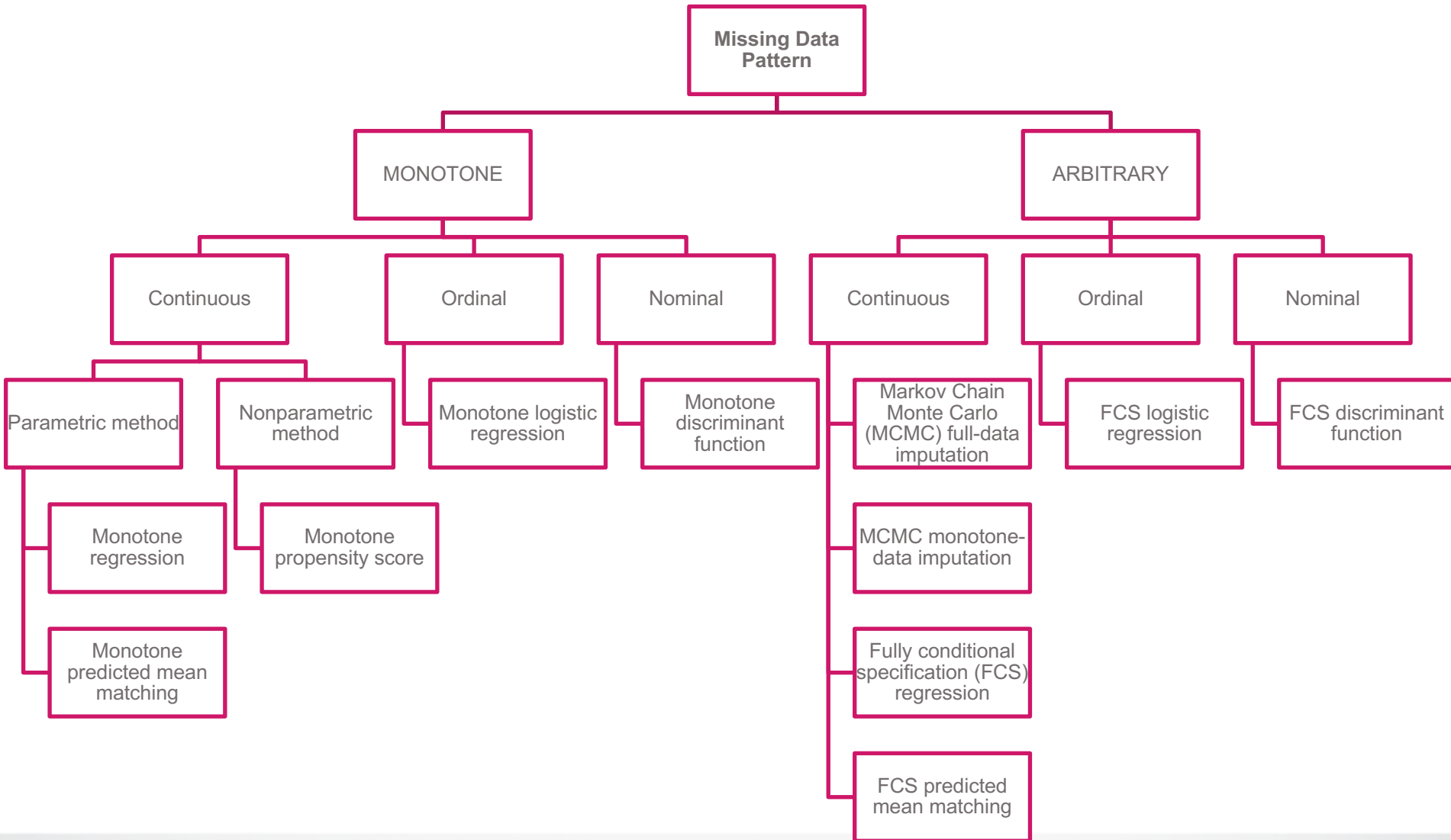


- PROC MI in SAS is a simulation-based procedure.
- Its purpose is not to re-create the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference.
- There are several imputation methods in PROC MI
- The method of choice depends on the pattern of missingness in the data and the type of the imputed variable



Multiple Imputation

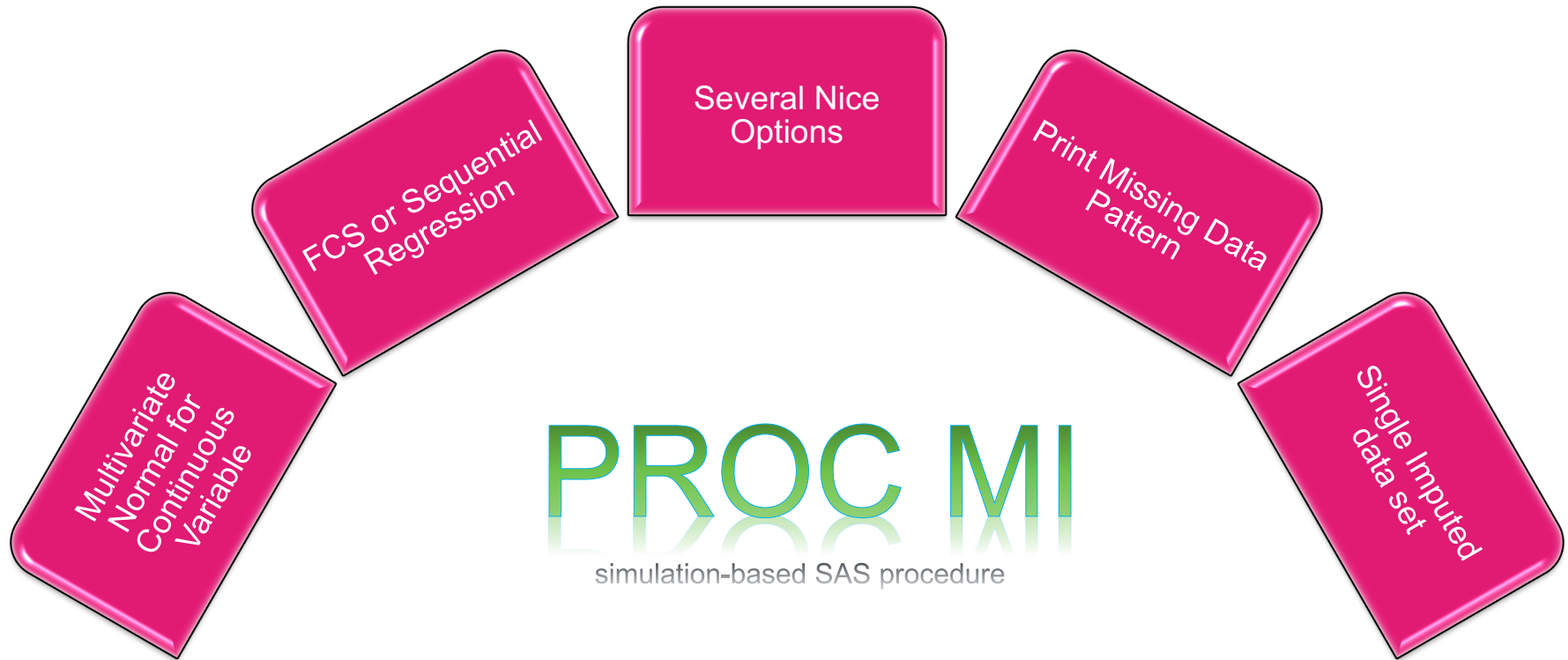
PROC MI - Method of Choices





Multiple Imputation

PROC MI - Features





Multiple Imputation

PROC MI - Statements



PROC MI
<options>;

BY variables; You can specify a BY statement with PROC MI to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used

CLASS variables; The CLASS statement specifies the classification variables in the VAR statement. Classification variables can be either character or numeric. The CLASS statement must be used in conjunction with either an FCS or MONOTONE statement.

EM <options>; The expectation-maximization (EM) algorithm is a technique for maximum likelihood estimation in parametric models for incomplete data

FCS <options>; The FCS statement specifies a multivariate imputation by fully conditional specification methods. If you specify an FCS statement, you must also specify a VAR statement.

FREQ variable; To run a procedure on an input data set that contains observations that occur multiple times, you can use a variable in the data set to represent how frequently observations occur and specify a FREQ statement with the name of that variable as its argument (variable) when you run the procedure..

MCMC <options>; The MCMC statement specifies the details of the MCMC method for imputation

MNAR options; The MNAR statement imputes missing values by using the pattern-mixture model approach, assuming the missing data are missing not at random (MNAR).

MONOTONE <method <(imputed <= effects>> </ options>)>> <...method <(imputed <= effects>> </ options>)>> ; The MONOTONE statement specifies imputation methods for data sets with monotone missingness. You must also specify a VAR statement, and the data set must have a monotone missing pattern with variables ordered in the VAR list..

TRANSFORM transform (variables</ options>)<...transform (variables</ options>)> ; The TRANSFORM statement lists the transformations and their associated variables to be transformed. The options are transformation options that provide additional information for the transformation

VAR variables; The VAR statement lists the variables to be analyzed..



The pattern of missingness is arbitrary and all variables with missing values are continuous.

Full-data imputation, which
uses a single chain

create 5 imputations

```
proc mi data=Sashelp.Heart nimpute=5 out=ImputedHeart seed=55555;  
  mcmc;  
  var AgeAtStart Height Weight Diastolic Systolic MRW Smoking  
      Cholesterol;  
run;
```



Multiple Imputation

PROC MI – Example 1 [Print out of Variance Information and Parameter Estimates]



The pattern of missingness is arbitrary and all variables with missing values are continuous.

Variance Information							
Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Height	0.000000292	0.002464	0.002464	5205.1	0.000142	0.000142	0.999972
Weight							
MRW							
Smoking							
Cholesterol							

The output variance information presents the between-imputation variance, within-imputation variance, and total variance for combining complete-data inferences for each variable, along with the degrees of freedom for the total variance, relative increase in variance, fraction missing information and relative efficiency.

Parameter Estimates									
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0 Pr > t
Height	64.811979	0.049637	64.7147	64.9093	5205.1	64.811077	64.812496	0	1305.71 <.0001
Weight	153.062451	0.401226	152.2759	153.8490	5154.3	153.044720	153.083799	0	381.49 <.0001
MRW	119.945077	0.277348	119.4014	120.4888	5120.9	119.929548	119.961558	0	432.47 <.0001
Smoking	9.369757	0.							
Cholesterol	227.483481	0.							

The parameter estimates output summarizes basic descriptive statistics for the imputed values by variable. The imputed data sets are stored in the ImputedHeart data set, with the index variable _Imputation indicating the imputation numbers. The data set can now be analyzed using standard statistical procedures with _Imputation_ as a BY variable



Multiple Imputation

PROC MI – Example 1 [Print out EM (Posterior Mode) Estimates]



The pattern of missingness is arbitrary and all variables with missing values are continuous.

EM (Posterior Mode) Estimates									
TYPE	_NAME_	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
MEAN		44.068727	64.812354	153.072405	85.358610	136.909580	119.949917	9.369819	227.405950
COV	AgeAtStart	73.388922	-4.040112	23.314275	30.577926	77.080562	35.000499	-17.265641	105.208970
COV	Height	-4.040112	12.809062	53.504809	-0.661331	-6.060066	-9.733376	12.372054	-12.814415
COV	Weight	23.314275	53.504809	834.684737	122.801814	180.715779	442.555772	30.746791	95.456715
COV	Diastolic	30.577926	-0.661331	122.801814	167.978558	244.697759	99.649738	-10.130760	106.662408
COV	Systolic	77.080562	-6.060066	180.715779	244.697759	562.488389	171.676364	-26.537106	211.853591
COV	MRW	35.000499	-9.733376	442.555772	99.649738	171.676364	398.571693	-30.102483	123.111525
COV	Smoking	-17.265641	12.372054	30.746791	-10.130760	-26.537106	-30.102483	144.495380	-6.161358
COV	Cholesterol	105.208970	-12.814415	95.456715	106.662408	211.853591	123.111525	-6.161358	2015.365252

The posterior mode, the highest observed-data posterior density, with a noninformative prior, is computed from the expectation-maximization (EM) algorithm and is used as the starting value for the chain.



Multiple Imputation

PROC MI – Example 1 [Output SAS dataset: ImputedHeart – Imputation=1]



...	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
127	Alive		.	Female	57	66.5	137	78	138	104	0	.	261
128	Dead	Other	.	Female	34			84	130		0	48	242
129	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
131	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242

	Imputation	Status	DeathCause	AqeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	1	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	1	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	1	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
127	1	Alive		.	Female	57	66.5	137	78	138	104	0	.	261
128	1	Dead	Other	.	Female	34	58.004845896	95.41981434	84	130	104.41486668	0	48	242
129	1	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	1	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
131	1	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242



Multiple Imputation

PROC MI – Example 1 [Output SAS dataset: ImputedHeart – Imputation=5]



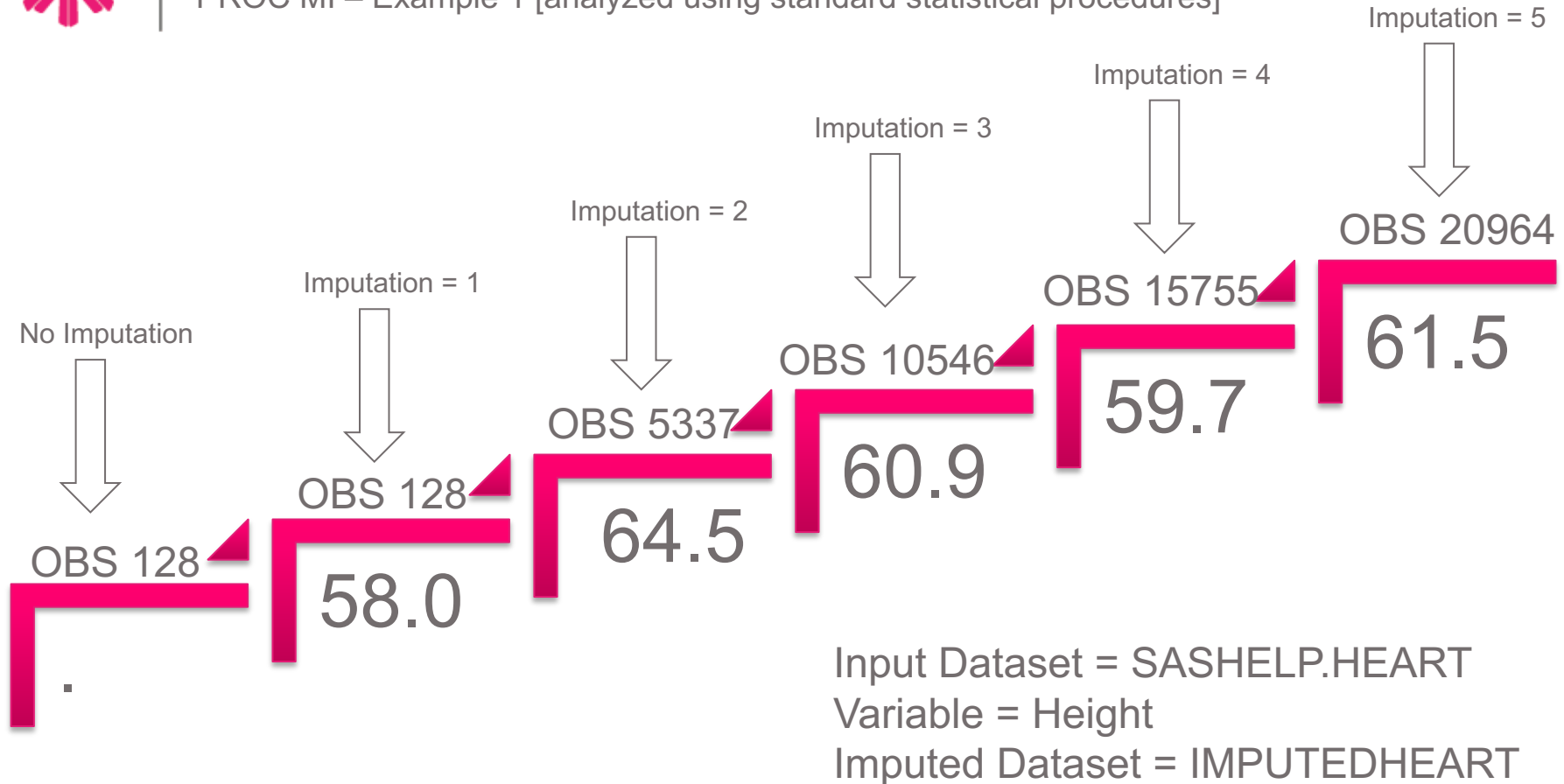
	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
127	Alive		.	Female	57	66.5	137	78	138	104	0	.	261
128	Dead	Other	.	Female	34			84	130		0	48	242
129	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
131	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242

	Imputation	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
20960	5	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
20961	5	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
20962	5	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
20963	5	Alive		.	Female		61.5	137	78	138	104	0	.	261
20964	5	Dead	Other	.	Female	34	61.542391821	151.98430173	84	130	140.63715669	0	48	242
20965	5	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
20966	5	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
20967	5	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242



Multiple Imputation

PROC MI – Example 1 [analyzed using standard statistical procedures]



The data set can now be analyzed using standard statistical procedures with `_Imputation_` as a BY variable



Multiple Imputation

PROC MI – Example 2 [MAR Assumption]



Multiple imputation often assumes that missing values are missing at random (MAR). The below example the MI procedure to impute missing values under this assumption

Obs	Trt	y0	y1
1	0	11.4826	.
2	0	10.009	10.8667
3	0	11.3643	10.666
4	0	11.3098	10.8297
5	0	11.3094	.
6	1	10.3815	10.5587
7	1	11.2001	13.7616
8	1	9.7002	10.346
9	1	10.0801	.
10	1	11.2667	11.0634

Sample Trial Data (First 10 Observations)

- Suppose that a pharmaceutical company is conducting a clinical trial to test the efficacy of a new drug.
- The trial is conducted with two groups that have an equal number of patients: a treatment group that receives the new drug and a placebo control group.
- The variable TRT is an indicator variable, with a value of 1 for patients in the treatment group and a value of 0 for patients in the control group.
- The variable Y0 is the baseline efficacy score, and the variable Y1 is the efficacy score at a follow-up visit.
- Suppose in a trial that the variables TRT and Y0 are fully observed and the variable **Y1 contains missing values** in both the treatment and control groups.

```
proc mi data=Mono2 seed=14823 out=outmi;  
  class Trt;  
  monotone reg;  
  var Trt y0 y1;  
  
run;
```

The MAR assumption should be examined after the analysis of the imputed dataset.



Multiple Imputation

PROC MI – Example 2



Adjusting Imputed Classification Levels in Sensitivity Analysis: This example generate imputed dataset for a specified sequence of shift parameters, which adjust the imputed values for observations in the treatment group (TRT=1)

Model Information	
Data Set	WORK.MONO2
Method	Monotone
Number of Imputations	5
Seed for random number generator	14823

Monotone Model Specification	
Method	Imputed Variables
Regression	y0 y1

Missing Data Patterns							
Group	Trt	y0	y1	Freq	Percent	Group Means	
						y0	y1
1	X	X	X	7	70.00	10.747371	11.156014
2	X	X	.	3	30.00	10.957367	.

MNAR Adjustments to Imputed Values		
Imputed Variable	Observations	Shift
y1	Trt = 1	-1.5300

Variance Information							
Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
y1	0.016696	0.197003	0.217038	6.7103	0.101699	0.096154	0.981132

```
proc mi data=Mono2 seed=14823
out=outmi;
class Trt;
monotone reg;
mnar adjust( y1/shift=-1.53
adjustobs=(Trt='1')) ;
var Trt y0 y1;
run;
```

Parameter Estimates									
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0 Pr > t
y1	10.964555	0.465873	9.853226	12.07588	6.7103	10.779723	11.122095	0	23.54 <.0001

In the MI procedure the MNAR statement used to impute missing values under various MNAR scenarios and examine the results. If the results under MNAR differ from the results under MAR, then you should question the conclusion under the MAR assumption.



Method Used to Deal with Missing Values



Maximum Likelihood Imputation



- This method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance - covariance matrix to estimate the regression model
- PROC MI uses the default algorithm (EM) to do maximum likelihood of the means and the covariance matrix and, then, it considers these estimates as starting values for multiple imputation algorithms.



Maximum Likelihood Imputation

PROC MI - Example



```
proc mi data=Sashelp.Heart nimpute=0;
em out=EMImputedHeart;
var AgeAtStart Height Weight Diastolic Systolic MRW
    Smoking Cholesterol;
run;
```

suppresses the multiple imputation

Estimates and writes the means and covariance matrix into SAS data set

The output data set from PROC MI (the EM covariance matrix) as input to PROC REG to estimate our model

Initial Parameter Estimates for EM

TYPE	_NAME_	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
MEAN		44.068727	64.813185	153.086681	85.358610	136.909580	119.957525	9.366518	227.417441
COV	AgeAtStart	73.529838	0	0	0	0	0	0	0
COV	Height	0	12.835792	0	0	0	0	0	0
COV	Weight	0	0	836.101866	0	0	0	0	0
COV	Diastolic	0	0	0	168.301098	0	0	0	0
COV	Systolic	0	0	0	0	563.568435	0	0	0
COV	MRW	0	0	0	0	0	399.336335	0	0
COV	Smoking	0	0	0	0	0	0	144.755816	0
COV	Cholesterol	0	0	0	0	0	0	0	2019.201302

EM (MLE) Parameter Estimates

TYPE	_NAME_	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
MEAN		44.068727	64.812354	153.072405	85.358610	136.909580	119.949917	9.369819	227.405950
COV	AgeAtStart	73.515722	-4.047092	23.354557	30.630758	77.213740	35.060972	-17.295472	105.390745
COV	Height	-4.047092	12.831202	53.597287	-0.662474	-6.070537	-9.750197	12.393431	-12.836556
COV	Weight	23.354557	53.597287	836.128050	123.013989	181.028016	443.321192	30.799915	95.621641
COV	Diastolic	30.630758	-0.662474	123.013989	168.268788	245.120542	99.821910	-10.148263	106.846696
COV	Systolic	77.213740	-6.070537	181.028016	245.120542	563.460244	171.972983	-26.582956	212.219624
COV	MRW	35.060972	-9.750197	443.321192	99.821910	171.972983	399.260989	-30.154493	123.324234
COV	Smoking	-17.295472	12.393431	30.799915	-10.148263	-26.582956	-30.154493	144.746562	-6.171769
COV	Cholesterol	105.390745	-12.836556	95.621641	106.846696	212.219624	123.324234	-6.171769	2018.939258



Conclusion





Conclusion

Missing Data Imputation and Its Effect on the Accuracy of Classification



- Missing data introduces complexity in statistical inference about unknown parameter or the population quantity.
- The complete data model assumption and missing data mechanism are needed to construct inference from incomplete data.
- The mean and median imputation have a similar mean percentage of observations correctly classified.
- The hot deck imputation has the highest mean percentage of observations correctly classified.
- Multiple imputation approach serves as the most general-purpose method, but not the best absolute methodology.



From a practical point of view, a carefully chosen imputation process and a decently efficient statistical method for analysis of complete data will not have many problems.



Thank You





1. Data Science - Innovative Developments in Data Analysis and Clustering by Francesco Palumbo, Angela Montanari, Maurizio Vichi and the topic from Lynette A. Hunt
2. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K Kapoor
3. SAS Blogs by Rick Wicklin
4. Missing Data Analysis in Practice, Trivellore Raghunathan
5. Multiple Imputation of Missing Data Using SAS by Patricia Berglund and Steven Heeringa
6. SAS/STAT 14.1 Online Documentation
7. Missing data: the hidden problem, SPSS White Paper
8. Dealing with missing data: Key assumptions and methods for applied analysis, Marina Soley-Bori



Edwin Ponraj Thangarajan & Giri Balasubramanian

PRA Health Sciences

40, II Main Road, R.A. Puram

Chennai - 600 028, Tamilnadu, India

Email: ThangarajanEdwin@prahs.com,
BalasubramanianGiri@prahs.com

Web: www.prahs.com



PRAHEALTHSCIENCES

NO MATTER WHERE YOUR
CLINICAL DEVELOPMENT TAKES YOU,
WE'RE ALREADY THERE.