UPPSALA UNIVERSITET

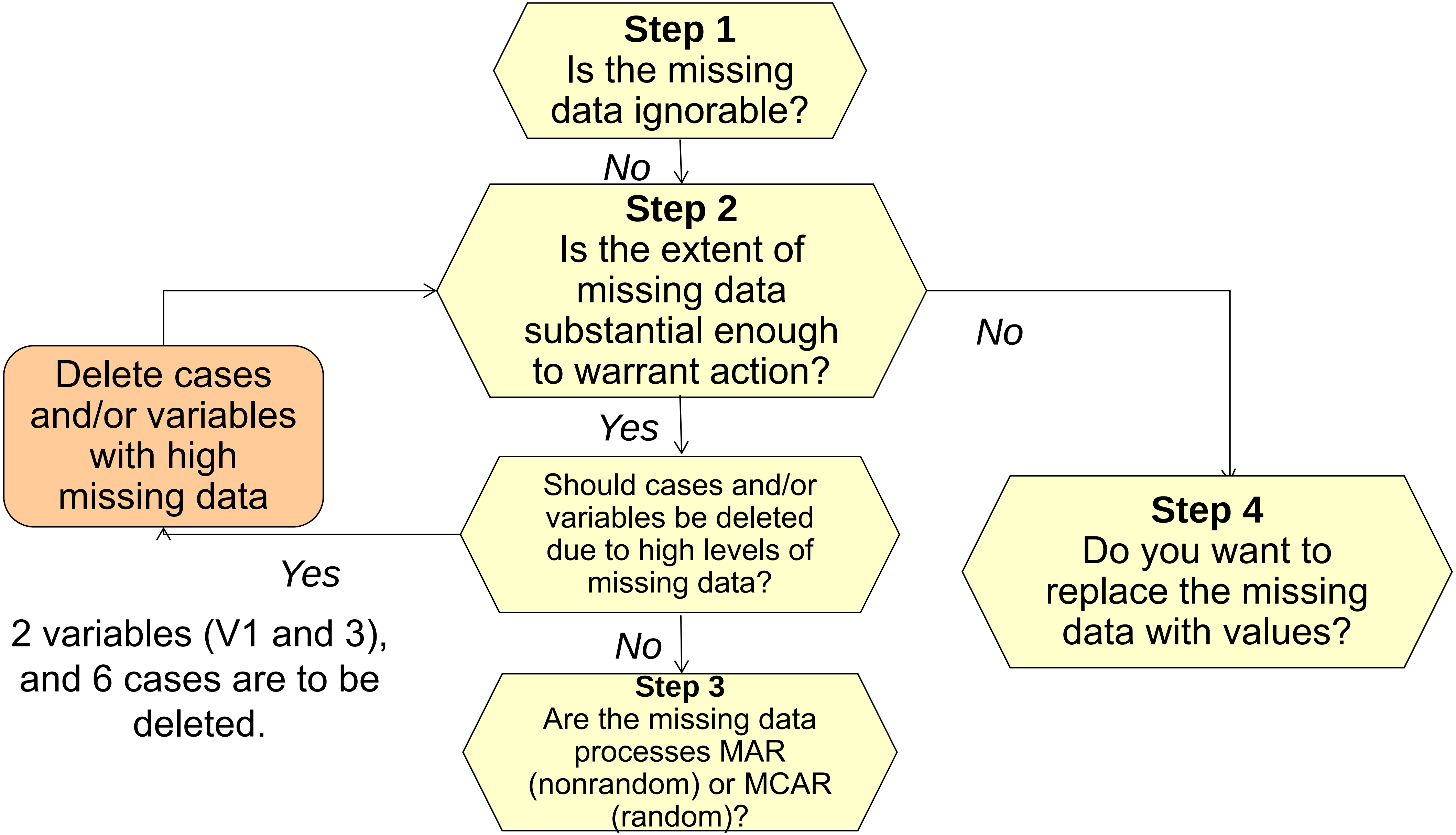
Quantitative Methods II

WELCOME!

Lecture 4

Thommy Perlinger

A four-step process for identifying missing data and applying remedies



2 variables (V1 and 3), and 6 cases are to be deleted.

A four-step process for identifying missing data and applying remedies

Step 3: diagnose the randomness of the missing data processes

If the extent of missing data is substantial enough to warrant action, the degree of randomness in the missing data has to be ascertained.

A nonrandom missing data process is present between the two variables X and Y when significant differences in the values of X occur between cases that have valid data for Y versus those cases with missing data on Y.

Levels of randomness of the missing data process

Two levels of randomness of missing data:

- Missing At Random (MAR), which requires special methods to accommodate a nonrandom component.
- Missing Completely At Random (MCAR), which is sufficiently random to accommodate any type of missing data remedy.

The distinction between these two levels is in the generalizability to the population.

Example: Missing at random (MAR)

X= gender of the respondents (assumed to be known)

Y = household income

Missing data are random for both males and females, but occur much more frequently for males.

The missing data is random within the gender variable, but the observed data is not generalizable to the population since it does not reflect the ultimate distribution of the household income values.

Example: Missing completely at random (MCAR)

X= gender of the respondents (assumed to be known)

Y = household income

Missing data are random for both males and females, and in equal proportions for both gender.

In this missing data process, any remedy can be applied without having to consider the impact of any other variable or missing data process.

Diagnostic tests for levels of randomness

There are two diagnostics tests that can be used to assess the level of randomness (MAR or MCAR):

1) Two groups of individuals are formed: one with missing values of Y, and another with valid values of Y. Then statistical tests (e.g. *t*-tests) are performed to see if differences exist between the two groups based on other variables of interest. Significant differences indicate the possibility of nonrandom missing data.

A number of variables should be examined to find any consistent pattern. Either a large number of differences or a systematic pattern may indicate a nonrandom component (MAR).

Diagnostic tests for levels of randomness

2) An overall test of randomness compares patterns of missing data on all variables with the pattern expected for random missing data. If no significant differences are found, the missing data can be classified as MCAR.

   If significant differences *are* found, the nonrandom missing data processes have to be investigated.

As a result of these tests, the missing data process is classified as either MAR or MCAR.

UPPSALA UNIVERSITET

Example: HBAT missing data

1) Two groups of individuals are formed: one with missing values of e.g. V2, and another with valid values of V2.

id	v2	Group_missingV2
201	,9	0
202	,4	0
203	.	1
204	1,5	0

Then, t-tests are performed to see if differences exist between the two groups based on all other numerical variables of interest.

UPPSALA UNIVERSITET

Separate Variance t Tests[a]

		v2	v4	v5	v6	v7	v8	v9
v2	t	.	-2,2	4,2	2,4	-1,2	-1,1	-1,2
	df	.	12,1	17,8	12,0	11,0	9,3	18,6
	P(2-tail)	.	,044	,001	,034	,260	,318	,233
	# Present	54	50	49	53	51	52	50
	# Missing	0	10	10	10	9	8	10
	Mean(Present)	1,896	4,988	2,704	2,506	6,682	45,462	4,754
	Mean(Missing)	.	5,940	3,500	3,110	7,400	49,250	5,020
v4	t	2,6	.	,2	1,4	1,5	,2	-2,4
	df	5,5	.	4,0	3,8	5,8	4,1	4,5
	P(2-tail)	,046	.	,888	,249	,197	,830	,064
	# Present	50	60	55	59	56	56	56
	# Missing	4	0	4	4	4	4	4
	Mean(Present)	1,942	5,147	2,842	2,625	6,832	46,018	4,757
	Mean(Missing)	1,325	.	2,800	2,250	6,200	45,250	5,375
v5	t	-,3	,4	.	-,9	-,4	,5	,6
	df	6,4	7,1	.	4,8	4,5	4,4	4,5
	P(2-tail)	,749	,734	.	,423	,696	,669	,605
	# Present	49	55	59	58	55	55	55
	# Missing	5	5	0	5	5	5	5
	Mean(Present)	1,888	5,156	2,839	2,579	6,758	46,182	4,820
	Mean(Missing)	1,980	5,040	.	2,860	7,140	43,600	4,560
v7	t	,9	-2,1	,9	-1,5	.	,5	,4
	df	2,3	3,6	3,6	4,8	.	2,1	4,5
	P(2-tail)	,440	,118	,441	,193	.	,658	,704
	# Present	51	56	55	59	60	57	56
	# Missing	3	4	4	4	0	3	4
	Mean(Present)	1,920	5,073	2,860	2,581	6,790	46,140	4,805
	Mean(Missing)	1,500	6,175	2,550	2,900	.	42,667	4,700
v8	t	-1,4	-1,1	-,9	-1,8	1,7	.	1,6
	df	1,0	3,9	4,1	4,0	9,1	.	5,7
	P(2-tail)	,384	,326	,401	,149	,128	.	,155
	# Present	52	56	55	59	57	60	56
	# Missing	2	4	4	4	3	0	4
	Mean(Present)	1,854	5,113	2,822	2,573	6,816	45,967	4,821
	Mean(Missing)	3,000	5,625	3,075	3,025	6,300	.	4,475
v9	t	,8	2,5	2,7	1,3	,9	2,4	.
	df	3,7	3,6	3,8	2,3	4,2	4,6	.
	P(2-tail)	,463	,076	,056	,302	,409	,066	.
	# Present	50	56	55	60	56	56	60
	# Missing	4	4	4	3	4	4	0
	Mean(Present)	1,920	5,232	2,895	2,623	6,825	46,429	4,798
	Mean(Missing)	1,600	3,950	2,075	2,167	6,300	39,500	.

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

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

Variables used to test for differences between the groups

Variable that the groups are based on

UPPSALA UNIVERSITET

Example: HBAT missing data

Separate Variance t Tests[a]

		V2	V4	V5	V6	V7	V8	V9
V2	t	.	-2,2	-4,2	-2,4	-1,2	-1,1	-1,2
	df	.	12,1	17,8	12,0	11,0	9,3	18,6
	P(2-tail)	.	,044	,001	,034	,260	,318	,233
	# Present	54	50	49	53	51	52	50
	# Missing	0	10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Mean(Present)	1,896	4,988	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Mean(Missing)	.	5,940	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
V4	t	2,6	.	,2	1,4	1,5	,2	-2,4
	df	5,5	.	[illegible]	[illegible]	[illegible]	[illegible]	4,5
	P(2-tail)	,046	.	[illegible]	[illegible]	[illegible]	[illegible]	,064
	# Present	50	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	56
	# Missing	4	0	4	4	4	4	4
	Mean(Present)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Mean(Missing)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	t	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	df	6,4	7,1		4,8	4,5	4,4	4,5

Three significant differences between groups based on V2.

Only one significant difference among the rest of the tests.

SPSS:
Analyze >> Missing Value Analysis. Click "Descriptives", mark "t tests with groups formed by indicator variables"

UPPSALA UNIVERSITET

Example: HBAT missing data

2) An overall test of randomness.

EM Means[a]

v2	v4	v5	v6	v7	v8	v9
1,989	5,137	2,826	2,583	6,835	46,029	4,758

a. Little's MCAR test: Chi-Square = 57,708, DF = 56, Sig. = ,412

P-value (two-sided)

H_0 : *The observed pattern of missing data does not differ from a random pattern.*

H_a : *The observed pattern of missing data differs from a random pattern.*

SPSS: Analyze >> Missing Value Analysis.
To the right under "Estimation", mark "EM" (for Little's MCAR test).

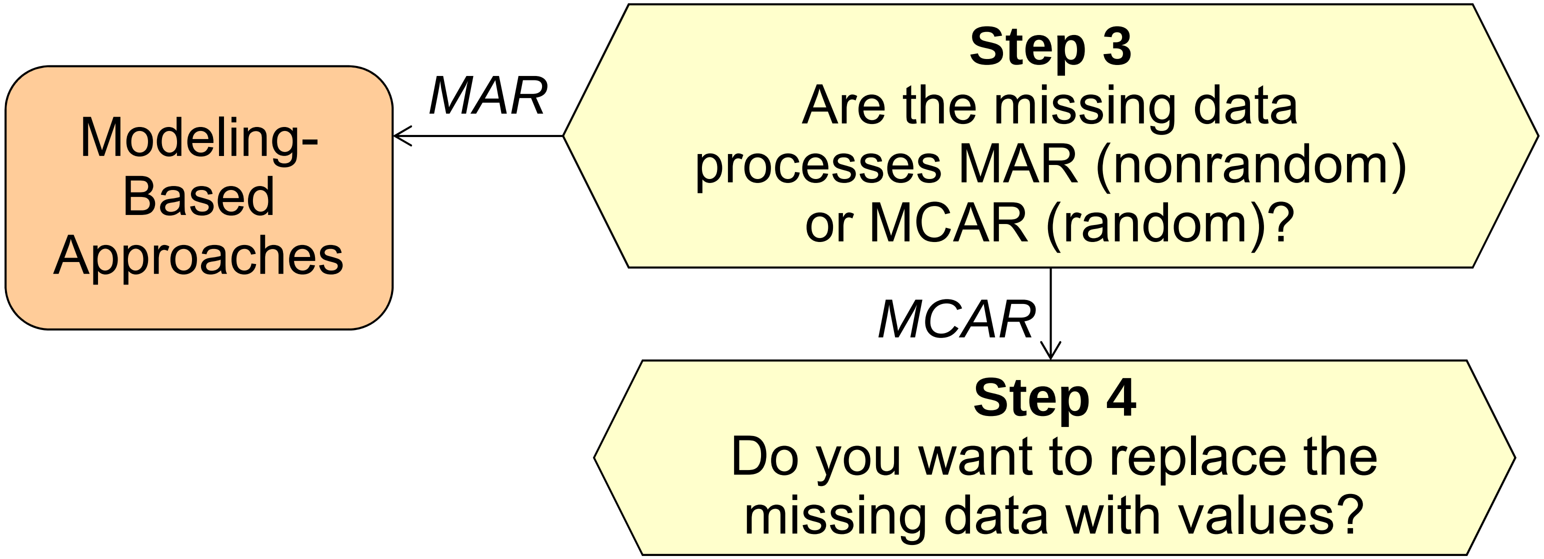
Example: HBAT missing data

This result, together with the analysis showing minimal differences in a nonrandom pattern, allows us to conclude that the missing data process is MCAR.

If the MCAR test had been significant, or a nonrandom pattern had been obvious in the previous analysis, the missing data process would have been concluded to be MAR.

UPPSALA UNIVERSITET

A four-step process for identifying missing data and applying remedies



A four-step process for identifying missing data and applying remedies

Step 4: select the imputation method

The potential impact of imputation on the analysis must be considered:

- Imputation can lure the user into believing that the data are complete after all.
- Imputation is dangerous since it lumps together situations where the problem is sufficiently minor to be legitimately handled in this way, and situations where standard estimators applied to the real and imputed data have substantial biases.

Imputation of a MAR missing data process

If a nonrandom or MAR missing data process is found, there is only one remedy to be applied (any other method introduces bias into the results):

Use a Modeling Based Approach, i.e. incorporate the missing data into the analysis .

Two approaches:

1) Techniques that attempt to model the processes underlying the missing data and to make the most accurate and reasonable estimates possible. E.g. Structural Equation Modeling (SEM), or the co-called EM approach (an iterative two-stage method).

Imputation of a MAR missing data process

2) Inclusion of missing data directly into the analysis, defining observations with missing data as a select subset of the sample. Most applicable for dealing with missing values on the explanatory variables of a dependent relationship.

   E.g. in regression analysis, observations with missing data are coded by using a dummy variable (1=missing, 0=valid values).

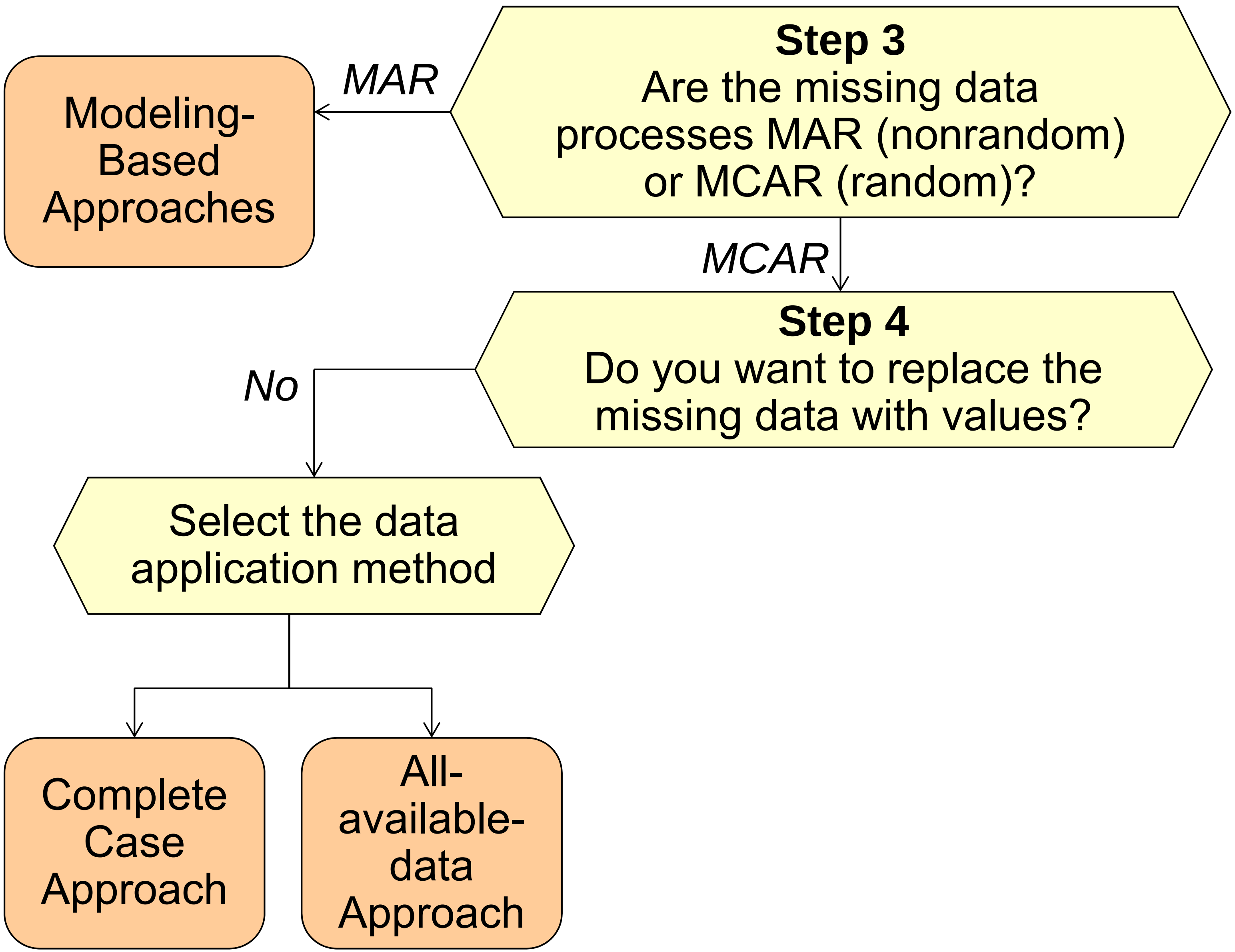
   This method enables you to retain all observations in the analysis and thus maintains the sample size.

Imputation of a MCAR missing data process

If an MCAR missing data process is found, there are two basic approaches to be used:

- **Use only valid data**
- **Define replacement values for the missing data**

A four-step process for identifying missing data and applying remedies



UPPSALA UNIVERSITET

Complete Case Approach

Include only those observations with complete data

id	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14
~~205~~	~~5,1~~	~~1,4~~	.	~~4,8~~	~~3,3~~	~~2,6~~	~~3,8~~	~~49,0~~	~~4,9~~	~~0~~	~~1~~	~~0~~	~~0~~	~~2~~
206	4,6	2,1	7,9	5,8	3,4	2,8	4,7	49,0	5,9	0	1	0	1	3
~~207~~	.	~~1,5~~	.	~~4,8~~	~~1,9~~	~~2,5~~	~~7,2~~	~~36,0~~	.	~~1~~	~~0~~	~~1~~	~~0~~	~~1~~
208	5,2	1,3	9,7	6,1	3,2	3,9	6,7	54,0	5,8	0	1	0	1	3
209	3,5	2,8	9,9	3,5	3,1	1,7	5,4	49,0	5,4	0	1	0	1	3
211	3,0	2,8	7,8	7,1	3,0	3,8	7,9	49,0	4,4	0	1	1	1	2
212	4,8	1,7	7,6	4,2	3,3	1,4	5,8	39,0	5,5	0	1	0	0	2
~~213~~	~~3,1~~	.	.	~~7,8~~	~~3,6~~	~~4,0~~	~~5,9~~	~~43,0~~	~~5,2~~	~~0~~	~~1~~	~~1~~	~~1~~	~~2~~

Available in all statistical programs and in many programs the default method.

SPSS: "Listwise deletion" (used for each statistical method).
E.g. Analyze >> Regression >> Linear
Click "Options" and choose "Exclude cases listwise"

Complete Case Approach

Disadvantages:

1) Highly affected by any nonrandom missing data processes. Results are not generalizable to the population.
2) Large reduction in sample size (missing data on *any* variable eliminates the entire case).

UPPSALA UNIVERSITET

Complete Case Approach

The complete case approach is best suited when:

- the extent of missing data is small,
- the sample is large enough to allow for deletion of cases, and
- the relationships in the data are so strong that they are not affected by any missing data process.

Example: HBAT missing data

We saw before that after deleting V1 and V3, there are only 37 cases with complete data (out of n = 70).

The complete case approach is not an option, since the sample size reduces too much.

Using All-Available Data

No missing data are actually replaced, but the distribution characteristics (e.g. means or standard deviations) or relationships (e.g. correlations) are imputed from every valid value.

E.g. the mean of V1 is calculated from all available values of V1.

E.g. the correlation between V3 and V4 is calculated from all cases with available values on V3 and V4.

The number of observations used in calculations will vary for each correlation.

v1	v2	v3	v4
5,1	1,4	.	4,8
4,6	2,1	7,9	5,8
.	1,5	.	4,8
5,2	1,3	9,7	6,1
3,5	2,8	9,9	3,5
3,0	2,8	7,8	7,1
4,8	1,7	7,6	4,2
3,1	.	.	7,8
4,0	,5	6,7	4,5
.	1,6	6,4	5,0
6,1	,5	9,2	4,8
.	2,8	5,2	5,0
3,1	2,2	6,7	6,8
6,5	.	9,0	7,0
.	1,6	.	4,8
3,9	2,2	.	4,6
2,8	1,4	8,1	3,8
.	.	8,6	5,7
4,7	1,3	.	.
3,4	2,0	9,7	4,7

UPPSALA UNIVERSITET

Using All-Available Data

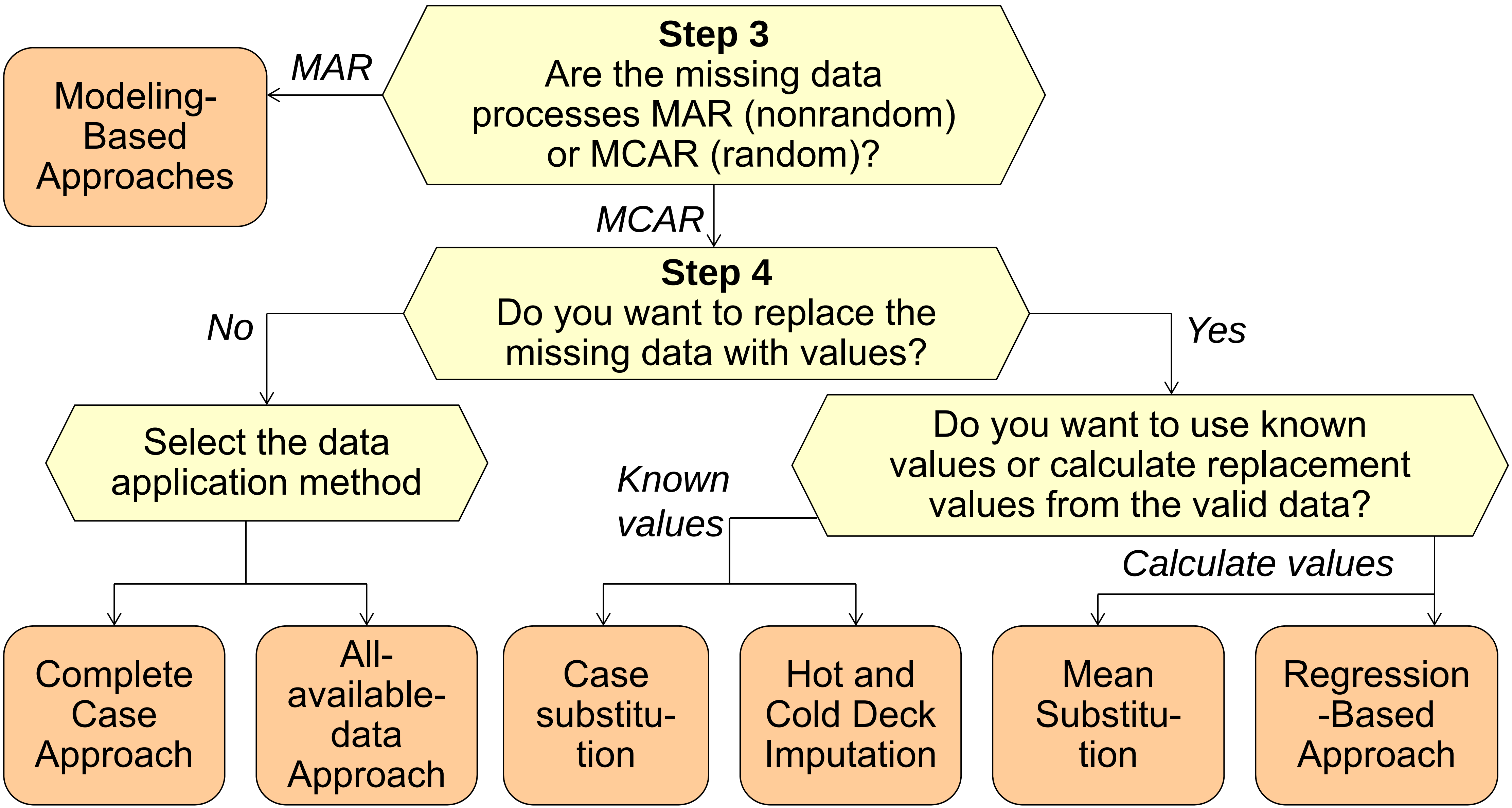
Disadvantage:

Correlations may be calculated that don’t fit the relationship that exists between different correlations.

SPSS: “Pairwise deletion” (used for each statistical method).
E.g. Analyze >> Regression >> Linear
Click “Options” and choose “Exclude cases pairwise”

UPPSALA UNIVERSITET

A four-step process for identifying missing data and applying remedies



Using known replacement values

A known value is identified, most often from a single observation, that is used to replace the missing data.

The observation with missing data is “matched” to a similar case, which provides the replacement value for the missing data.

Hot or Cold Deck Imputation

Hot Deck Imputation:

Each observation with missing data is paired with another case in the sample that is similar on some variable(s) that you specify.

Missing data are then replaced with valid values from that similar observation.

Cold Deck Imputation:

The replacement value is derived from an external source, such as a prior study.

Ensure that the replacement value from an external source is more valid than a value from the same sample.

Example: HBAT missing data

Say that we want to analyze the variables v5, v6, v7 and v8.

We want to replace the missing values for v5, using Hot Deck Imputation

Each observation with missing values for v5 is to be paired with another case with similar values on the other variables.

v5	v6	v7	v8
3,6	4,0	5,9	43,0
2,2	2,1	5,0	31,0
.	2,1	8,4	25,0
3,3	2,8	7,1	60,0
.	2,7	8,4	38,0
2,6	2,9	.	.
3,2	3,7	8,0	33,0
2,0	2,8	.	32,0
.	2,5	8,3	47,0
2,1	1,4	6,6	39,0

Example: HBAT missing data

Similar observations are easier found if the data set is sorted (by v6, v7 and v8)

Then, for every missing value of v5, find cases with similar values on the other variables.

v5	v6	v7	v8
3,0	3,8	7,9	49,0
3,2	3,9	6,7	54,0
.	3,9	6,8	54,0
3,3	3,9	7,3	59,0
3,6	4,0	5,9	43,0

Similar cases

The value 3.2 can be used to replace the missing value

SPSS: Data >> Sort Cases

Example: HBAT missing data

Sometimes it is not easy to find similar observations.

v5	v6	v7	v8
4,0	3,0	7,7	65,0
3,1	3,0	8,0	43,0
.	3,1	3,8	54,0
1,5	3,1	9,9	39,0
3,3	3,2	8,2	53,0

Similar cases ?

Sort the data set differently, by e.g. v7 first, and see what you can find.

SPSS: Data >> Sort Cases

Example: HBAT missing data

Sorted by v7, v6, v8

v5	v6	v7	v8
1,3	1,2	1,7	50,0
3,3	2,6	3,8	49,0
.	3,1	3,8	54,0
3,1	2,5	4,4	46,0
3,6	2,3	4,5	60,0

Similar cases ?

The definition of “similar cases” is up to you.

SPSS: Data >> Sort Cases

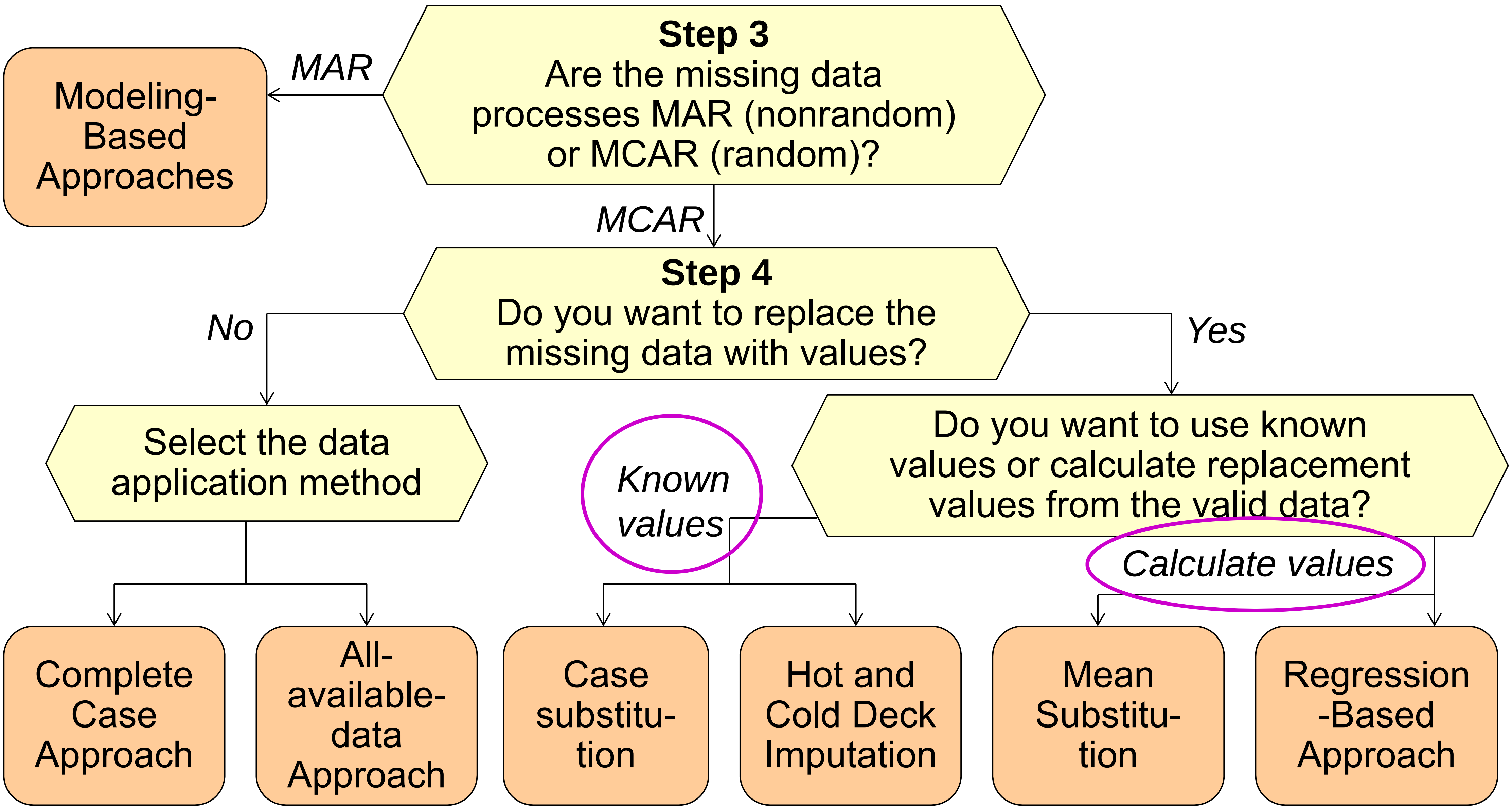
UPPSALA UNIVERSITET

Case Substitution

Entire cases with missing data are replaced by choosing another nonsampled observation (i.e., increasing the sample by another case)

E.g. a household with extensive missing data (or one that cannot be contacted) is replaced with another household not in the sample, preferably similar to the original observation.

Recap: A four-step process for identifying missing data and applying remedies



Mean Substitution

The missing values for a variable are substituted with the mean value of that variable calculated from all the valid values.

One of the most widely used methods (doesn't mean that it is the best method)

<u>Disadvantages:</u>

1) The variance of the variable is understated (more values equal to the mean decreases the average spread).
2) The actual distribution of values is distorted (there will be a larger concentration around the mean value).
3) The observed correlation is depressed since all missing data will have a single constant value.

Example: HBAT missing data

All missing values for V2 are substituted (imputed) by the mean of V2, which is 1.9

v2	v2_mean_subst
2,2	2,2
1,4	1,4
.	1,9
1,3	1,3
2,0	2,0
.	1,9
1,8	1,8
1,4	1,4
1,3	1,3
,7	,7
.	1,9
2,4	2,4

SPSS: Transform >> Replace Missing Values
Choose which variable(s) to impute missing values for, and choose Method "Series mean"

Regression Imputation

The missing values are predicted using regression analysis based on the relationship to other variables in the data set.

First, a regression equation e.g. for the variable V2 is estimated from the cases with valid data.

$$\widehat{V2} = b_0 + b_1 \cdot V4 + b_2 \cdot V5 + \ldots$$

Then, the missing values of V2 are estimated using this regression equation.

SPSS: Transform >> Replace Missing Values
Choose which variable(s) to impute missing values for, and choose Method "Linear trend at point"

Regression Imputation

Disadvantages:

1) The relationships already in the data are reinforced, the data become more characteristic of the sample and less generalizable.
2) The variance of the distribution is understated.
3) This approach assumes that the variable with missing data is highly correlated to the other variables. If this is not true, other methods are preferable.
4) A large sample size is needed for these calculations.
5) The predicted values may not fall in the valid ranges for variables.

Imputation or not?

There is little agreement about whether or not to conduct imputation.

If you do choose to conduct imputation of missing data, there are some guidelines that might be of support for your choice of imputation technique.

UPPSALA UNIVERSITET

Imputation techniques for missing data

Each method has advantages and disadvantages.
No single method is best in all situations.

It can be advantageous to combine several methods.
You can e.g. use two ore more imputation techniques and then use the mean of the different estimated values to substitute the missing values. That way you minimize the concerns with any single method.

There is an excellent comparison of different imputation techniques for missing data in the book!
(Table 8, p. 60)

Example: Combination of Imputation techniques

All missing values for V2 are substituted (imputed) by the mean of V2, which is 1.9

Then missing values for V2 are substituted (imputed) by a combination of variables, which is 2.3

And in the last step we take the mean

V2	v2_mean_subs	v2_reg_subs	v2_combined_subs
2,2	2,2	2,2	2,2
1,4	1,4	1,4	1,4
.	1,9	2,3	2,1
1,3	1,3	1,3	1,3
2,0	2,0	2,0	2,0
.	1,9	2,3	2,1
1,8	1,8	1,8	1,8
1,4	1,4	1,4	1,4
1,3	1,3	1,3	1,3
,7	,7	,7	,7
.	1,9	2,3	2,1
2,4	2,4	2,4	2,4

UPPSALA UNIVERSITET

Imputation of missing data

Rules of thumb

- Under 10% Any of the imputation methods can be applied, although the complete case method has been shown to be the least preferred.
- 10% to 20% The all-available, hot deck, case substitution, and regression methods are most preferred for MCAR data, and model-based methods are necessary with MAR data.
- Over 20% If deemed necessary to impute missing data, the preferred methods are:
  - The regression methods for MCAR
  - Model-based methods for MAR data

Program L4

- **Cleaning and transforming data**
  - Missing data, cont'd
  - Outliers
  - Assumptions of multivariate analysis

Program L4

- **Cleaning and transforming data**
  - Missing data, cont’d
  - Outliers
  - Assumptions of multivariate analysis

Outliers

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations.

It is typically an unusually high or low value, or a combination of values that make the observation stand out from the others.

Outliers can have a substantial effect on any type of analysis, and must therefore be investigated further.

Why outliers occur

Outliers can be:

- arising from a **procedural error**, such as a data entry error or a mistake in coding.
- the result of an **extraordinary event**, not comparable to anything normally seen.
- **extraordinary observations**, for which there is no explanation.
- observations that fall within the ordinary range of values on each of the variables, but are **unique in their combination of values** across the variables.

Detecting outliers

Outliers can be identified from a univariate, bivariate, or multivariate perspective based on the number of variables considered.

Look for a consistent pattern across these three perspectives to identify outliers.

When candidates for outlier designation are found, they must be examined, and you decide whether to keep or delete them (or correct them if the value is caused by mistake).

Univariate detection

Examine the distribution of observations for each variable in the analysis.

Identify as outliers those cases falling at the outer ranges (high or low) of the distribution, by looking at standardized values (standard scores or z-scores) that will tell you how many standard deviations from the mean each observation is located.

This way you can establish the threshold for designation of an outlier.

Outlier detection

Rules of thumb

- **Univariate methods**: Examine all metric variables to identify unique or extreme observations.
  - For small samples ($n \leq 80$), outliers are typically defined as cases with standardized values of ≥ 2.5 or ≤ -2.5.
  - For larger samples, increase the threshold value of standardized values up to ±4.

Example: HBAT (recap.)

Variable Description		Variable Type
Data Warehouse Classification Variables		
X_1	Customer Type	Nonmetric
X_2	Industry Type	Nonmetric
X_3	Firm Size	Nonmetric
X_4	Region	Nonmetric
X_5	Distribution System	Nonmetric

Example: HBAT (recap.)

Variable Description	Variable Type
Performance Perceptions Variables	
X_6 Product Quality	Metric
X_7 E-Commerce Activities/Web Site	Metric
X_8 Technical Support	Metric
X_9 Complaint Resolution	Metric
X_{10} Advertising	Metric
X_{11} Product Line	Metric
X_{12} Salesforce Image	Metric
X_{13} Competitive Pricing	Metric
X_{14} Warranty and Claims	Metric
X_{15} New Products	Metric
X_{16} Ordering and Billing	Metric
X_{17} Price Flexibility	Metric
X_{18} Delivery Speed	Metric

UPPSALA UNIVERSITET

Example: HBAT (recap.)

Variable Description	Variable Type
Outcome/Relationship Measures	
X_{19} Satisfaction	Metric
X_{20} Likelihood of Recommendation	Metric
X_{21} Likelihood of Future Purchase	Metric
X_{22} Current Purchase/Usage Level	Metric
X_{23} Consider Strategic Alliance/Partnership in Future	Nonmetric

Example: HBAT

Say that we want to investigate a relationship between the following variables:

- Perceived product quality (0-10), X6,
- Competitive pricing (0-10), X13
- Delivery speed (0-10), X18, and
- Satisfaction with past purchases (0-10), X19

Independent (explanatory) variables

Dependent/response variable

n = 100, we can use the rule of thumb for small samples (100 is close to 80)

UPPSALA UNIVERSITET

Example: HBAT (univariate detection)

id	x6	Zx6	Zx6_2_5	x13	Zx13	Zx13_2_5	x18	Zx18	Zx18_2_5	x19	Zx19	Zx19_2_5
22	9,6	1,282	.	4,5	-1,601	.	4,3	,564	.	9,9	2,502	1
7	6,9	-,652	.	8,9	1,247	.	2,0	-2,568	1	5,7	-1,022	.
84	6,4	-1,010	.	8,4	,923	.	1,6	-3,113	1	5,0	-1,609	.
1	8,5	,494	.	6,8	-,113	.	3,7	-,253	.	8,2	1,076	.

X6 & X13: No cases with standardized values exceeding ± 2.5 (sorted data set)

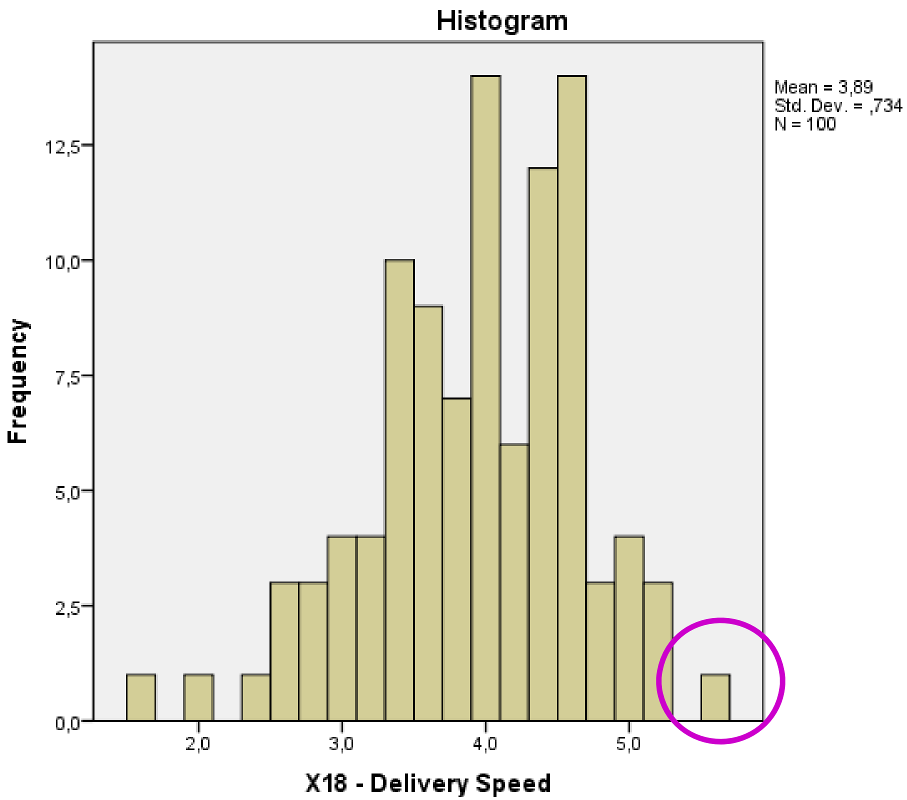
X18: two cases (id 7 & 84)

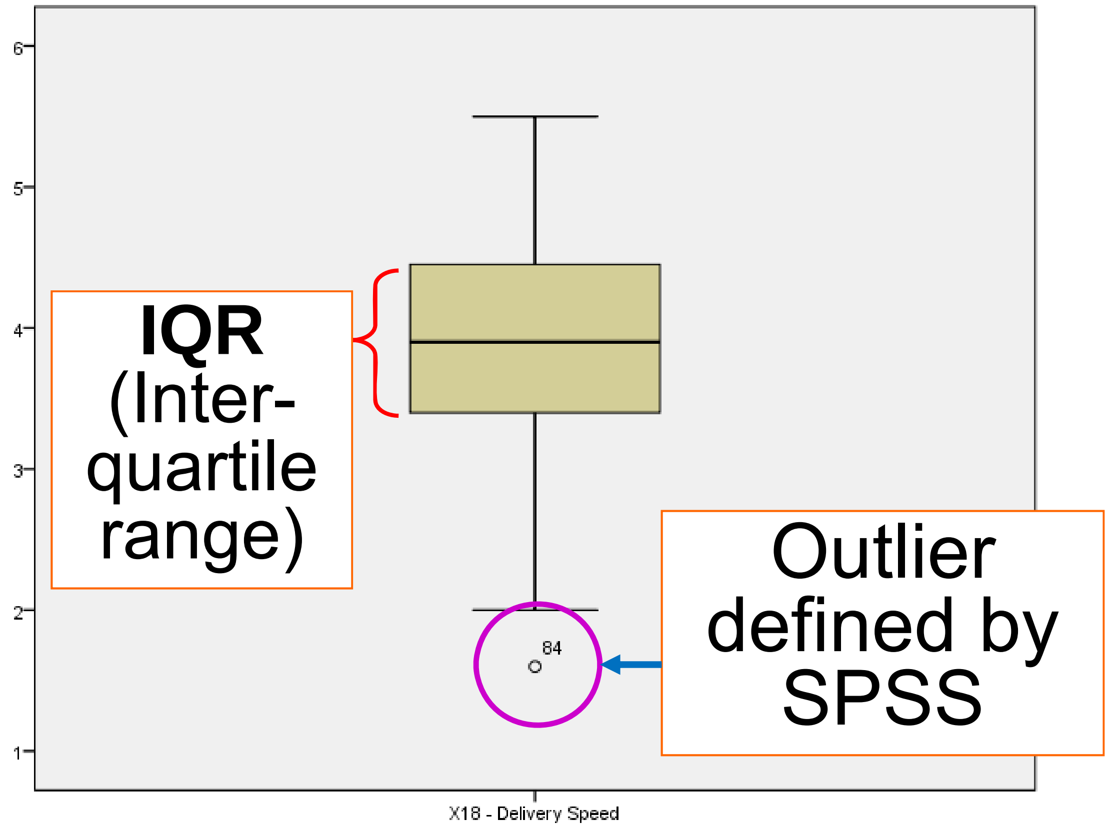
X19: one case (id 22)

No cases exceed the threshold on more than a single variable.

SPSS: Analyze >> Descriptive Statistics >> Descriptive
Mark “Save standardized values as variables”

Example: HBAT (univariate detection)





Outlier defined by SPSS: an observation more than 1.5 IQR away from Q1 or Q3.

SPSS: Analyze >> Descriptive Statistics >> Explore

Example: HBAT (univariate detection)

Descriptives

			Statistic	Std. Error
X18 - Delivery Speed	Mean		3,886	,0734
	95% Confidence Interval for Mean	Lower Bound	3,740	
		Upper Bound	4,032	
	5% Trimmed Mean		3,907	
	Median		3,900	
	Variance		,539	
	Std. Deviation		,7344	
	Minimum		1,6	
	Maximum		5,5	
	Range		3,9	
	Interquartile Range		1,1	
	Skewness		-,463	,241
	Kurtosis		,218	,478

Trimmed mean

If the trimmed mean is close to the mean, the 5% observations with the highest and lowest values have no major impact on the mean.

Bivariate detection

Examine pairs of variables through scatterplots.

Cases that fall markedly outside the range of the other observations will be seen as isolated points in the scatterplot.

UPPSALA UNIVERSITET

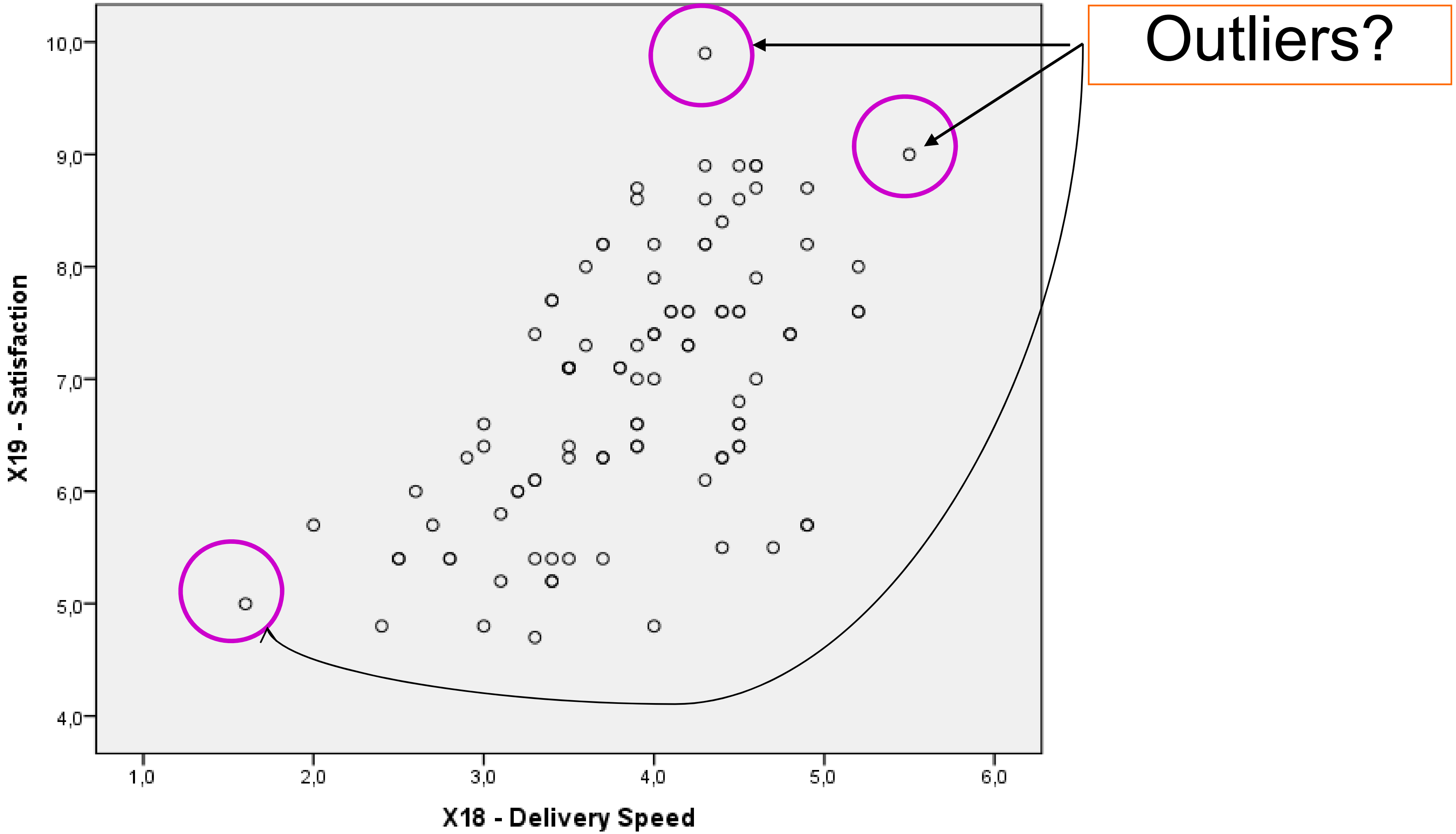
Outlier detection

Rules of thumb

- **Bivariate methods**: Focus on specific variable relation-ships, such as the independent vs. dependent variables.
  - Use scatterplots (with confidence intervals at a specified confidence level)

UPPSALA UNIVERSITET

Example: HBAT



Multivariate detection

The **Mahalanobis D^2 measure** objectively measures the multidimensional position of each observation relative to the mean center of all observations.

Higher D^2 values represent observations farther away from the general distribution of observations.

D^2/df (df=number of independent/explanatory variables) is approximately *t* distributed.

Outlier detection

Rules of thumb

- **Multivariate methods**: Best suited for examining a complete variate, such as the independent variables in regression or the variables in factor analysis.
  - Significance levels when testing the D^2/df measure should be conservative, 0.005 or 0.001. This leads to threshold values of 2.5 for small samples, vs. 3 or 4 in larger samples.

Example: HBAT

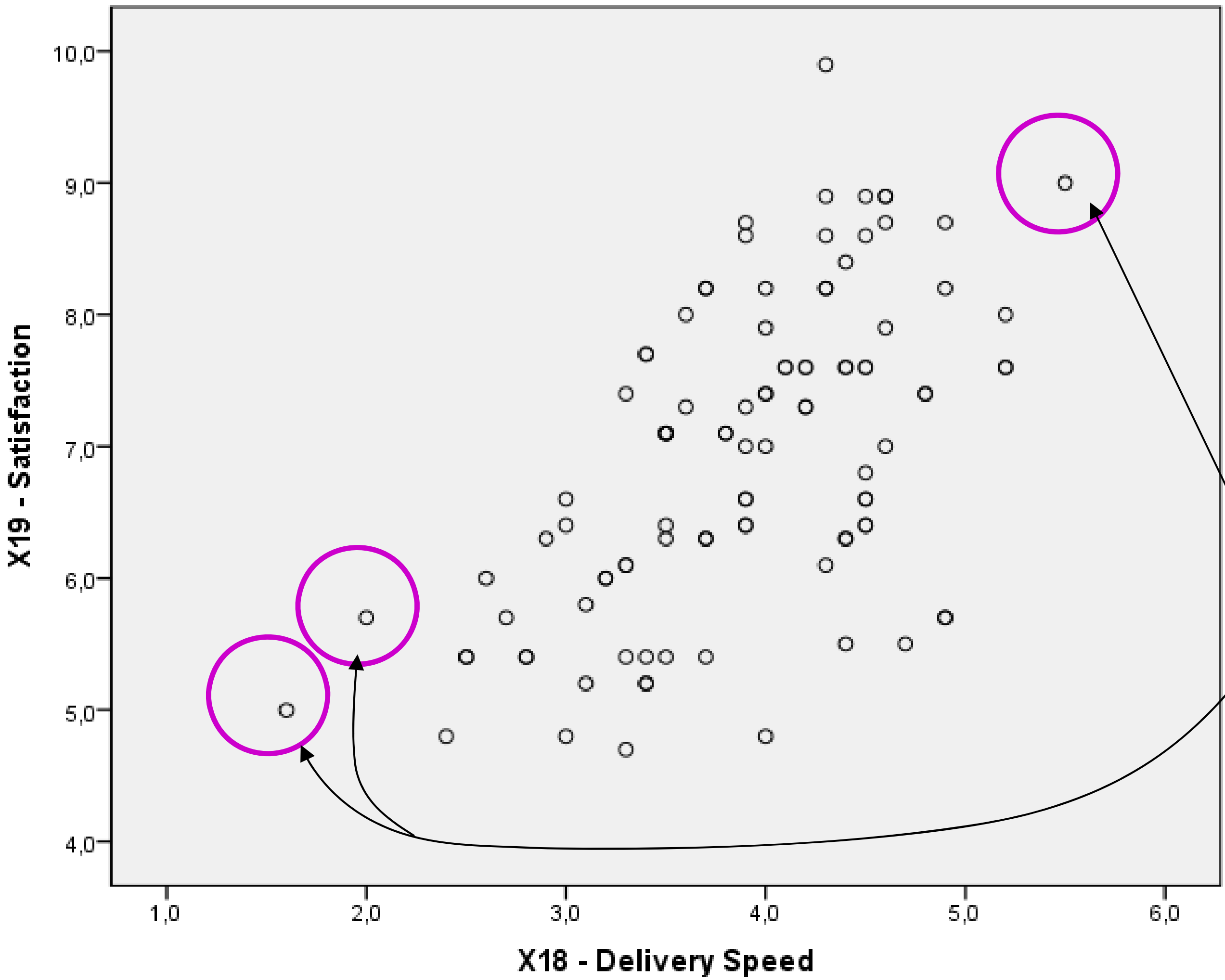
id	x18	Zx18	Zx18 2 5	x19	Zx19	Zx19 2 5	MAH 1	MAH df	MAH 2 5
7	2,0	-2,568	1	5,7	-1,022	.	7,75151	2,58	1
84	1,6	-3,113	1	5,0	-1,609	.	10,66992	3,56	1
57	5,5	2,198	.	9,0	1,747	.	8,48972	2,83	1
22	4,3	,564	.	9,9	2,502	1	3,25324	1,08	.
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]		[illegible]	[illegible]	

Mahalanobis D^2 $D^2/3$

3 cases with D^2/df measures exceeding 2.5 (sorted data set). Two of them are potential outliers according to the univariate detection as well (variable x18 exceeds ±2.5 standard deviations), while the third case's variable values are unique only in combination.

SPSS: Analyze >> Regression >> Linear
Click "Save", mark "Mahalanobis"

Example: HBAT



Possible outliers according to Mahalanobis' D^2

Outlier description and profiling

Generate profiles of each identified outlier observation, and identify the variable(s) responsible for its being an outlier.

Select only observations that demonstrate real uniqueness in comparison with the remainder of the population across as many perspectives as possible.

Refrain from designating too many observations as outliers, do not eliminate cases not consistent with the remaining cases just because they are different.

Retention or deletion of the outlier

UPPSALA UNIVERSITET

Retain possible outliers unless demonstrable proof indicates that they are truly deviant and not representative of any observations in the population.

If they do represent any part of the population, no matter how uncommon they are, they should be retained to ensure generalizability to the entire population.

When deleting outliers, the multivariate analysis may improve, but the generalizability is limited.