Irkutsk National Research Technical University
Baikal school of BRICS

09.04.02 Information systems and technology
Program: Information technologies, networks and big data

Course: Data analysis

Atalyan Alina Valerievna
alinaa@mail.ru

Lecture 2

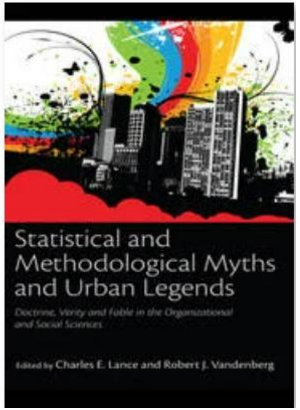
Data Manipulation and Cleaning

1. Missing Data
   - Identifying and dealing with NA values
   - Imputation techniques
2. Outliers
   - Definition.
   - Outlier Detection.
   - Applications And Techniques
3. Data Transformation
   - Reshaping and aggregating data
   - Combining datasets

Missing Data

Definition

The term ***missing data*** is defined here as a statistical problem characterized by an incomplete data matrix that results when one or more individuals in a sampling frame do not respond to one or more survey items.



Most missing data are due to survey nonresponse, which can vary from an intentional decision (discarding a survey or skipping sensitive items) to a rather unintentional act (forgetting a survey or being too busy to attend to a survey);
but missing data can also arise from technical errors on the part of the researcher or equipment (online survey programming errors or computer malfunction).

Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences (pp. 7-36). New York, NY: Routledge.

Missing Data

Three Levels of Missing Data: Item-, Construct-, and Person-Levels

Complete Data

	X_1	X_2	X_3	Y
*person*1	3	2	2	1
*person*2	2	2	2	3
*person*3	4	3	4	4
*person*4	3	3	3	3
*person*5	2	3	2	3
*person*6	4	4	4	3
*person*7	4	4	3	5
*person*8	3	2	3	5
*person*9	5	5	4	5
*person*10	2	3	2	3

Incomplete Data

	X_1	X_2	X_3	Y
*person*1	3	.	2	1
*person*2	.	.	.	3
*person*3	4	3	4	4
*person*4	.	.	.	.
*person*5	2	3	2	3
*person*6	.	.	.	.
*person*7	4	4	3	5
*person*8	3	2	.	5
*person*9	5	5	4	.
*person*10	2	3	2	3

Three Levels of Missingness

- Item-level missingness
- Construct-level missingness
- Person-level missingness

Missing Data

Three Levels of Missing Data and their Corresponding Missing Data Techniques

Level of Missing Data	Recommended Missing Data Technique	Favorable Condition for Technique
Item level	Use each person's $\text{mean}_{\text{(across available items)}}$ to represent the construct.	Parallel items[a]
Construct level	Use maximum likelihood (ML) or multiple imputation (MI), with auxiliary variables.	Missing at random (MAR) mechanism (probability of missingness is correlated with observed variables) or missing completely at random (MCAR) mechanism (completely random missingness)
Person level (i.e., as reflected in response rate)	Use sensitivity analysis.	Data are available from previous studies that compare respondents to nonrespondents on the constructs of interest (e.g., $r_{miss,x}$ can be estimated)

Missing Data

Three Mechanisms of Missing Data: Random Missingness (MCAR) and Systematic Missingness (MAR and MNAR)

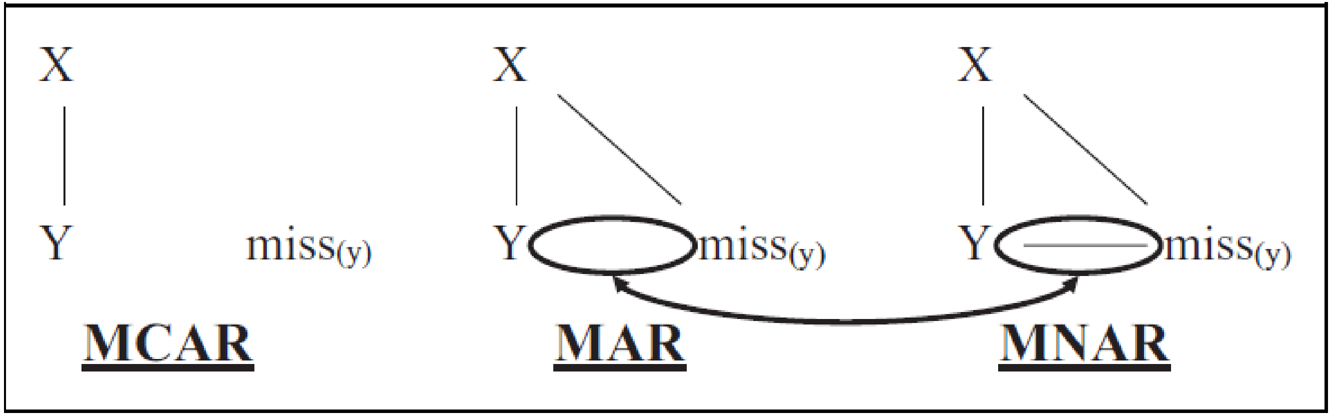
MCAR (missing completely at random) – the probability that a variable value is missing does not depend on the observed data values nor on the missing data values [i.e., p(missing|complete data) j p(missing)]. The missingness pattern results from a process completely unrelated to the variables in one's analyses, or from a completely random process (similar to flipping a coin or rolling a die).

MAR (missing at random) – the probability that a variable value is missing partly depends on other data that are observed in the dataset, but does not depend on any of the values that are missing [i.e., p(missing|complete data) j p(missing|observed data)].

MNAR (missing not at random) – the probability that a variable value is missing depends on the missing data values themselves [i.e., p(missing|complete data) 6j p(missing|observed data)].

Missing Data

Three missing data mechanisms (MCAR, MAR, MNAR) and the continuum between MAR and MNAR.



Note: Adapted from Schafer and Graham (2002, p. 152). Each line represents the relationship between two variables. Y is an incomplete variable (partly missing), and X is an observed variable. Miss(y) is a dummy variable that captures whether data are missing on variable Y. Notice that the difference between MAR and MNAR is simply the extent to which miss(y) is related to Y itself after X has been controlled.

MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random.

Missing Data

Two Missing Data Problems: Bias and Inaccurate Standard Errors/Hypothesis Tests

	Missingness Mechanism		
Missing Data Technique	MCAR	MAR	MNAR
Listwise Deletion	Unbiased; Large Std. Errors (Low Power)	Biased; Large Std. Errors (Low Power)	Biased; Large Std. Errors (Low Power)
Pairwise Deletion	Unbiased; Inaccurate Std. Errors	Biased; Inaccurate Std. Errors	Biased; Inaccurate Std. Errors
Single Imputation	Often Biased; Inaccurate Std. Errors	Often Biased; Inaccurate Std. Errors	Biased; Inaccurate Std. Errors
Maximum Likelihood (ML)	**Unbiased; Accurate Std. Errors**	**Unbiased; Accurate Std. Errors**	Biased; Accurate Std. Errors
Multiple Imputation (MI)	**Unbiased; Accurate Std. Errors**	**Unbiased; Accurate Std. Errors**	Biased; Accurate Std. Errors

Note. Recommended techniques are in boldface. Adapted from Newman (2009).

MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random

Missing Data

Several missing data considerations that must precede data analysis

Missing Data Are Partly Unavoidable, and Partly Avoidable

Define the Target Population of Interest

Missing Data

Missing data treatments

listwise deletion,
pairwise deletion,
single imputation/ad hoc approaches,
maximum likelihood (ML) approaches:
  full information maximum likelihood [FIML]
  the expectation-maximization [EM] algorithm
multiple imputation.

Missing Data

Missing data treatments

Missing Data Treatment	Definition	Major Issues
Single Imputation (ad hoc techniques)	Fill in each missing value [e.g., using mean (across persons) imputation, regression imputation, hot deck imputation, etc.], then proceed with analysis based on partially-imputed 'complete' dataset.	Mean (across persons) imputation and regression imputation are both biased under MCAR! No single *n* makes sense for whole correlation matrix (SEs inaccurate). SEs underestimated if you treat dataset as complete.
Maximum Likelihood	Directly estimate parameters of interest from incomplete data matrix (e.g., FIML); or Compute summary estimates [means, SDs, correlations] (e.g., EM algorithm), then proceed with analysis based on these summary estimates.	Unbiased under MCAR and MAR. Improves as you add more variables to the imputation model. Number of variables should be < 100. Accurate SEs for FIML. For EM algorithm, no single *n* makes sense for whole correlation matrix (SEs inaccurate).

MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random

Missing Data

Missing data treatments

Missing Data Treatment	Definition	Major Issues
Listwise Deletion	Delete all cases (persons) for whom any data are missing, then proceed with the analysis.	Discards real data from partial respondents. Smallest *n*, lowest power. Biased under MAR and MNAR.
Pairwise Deletion	Calculate summary estimates (means, SDs, correlations) using all available cases (persons) who provide data relevant to each estimate, then proceed with analysis based on these estimates.	Different correlations represent different subpopulation mixtures. Sometimes covariance matrix is not positive definite. Biased under MAR and MNAR. No single *n* makes sense for whole correlation matrix (SEs inaccurate).

MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random

Missing Data

Missing data treatments

Missing Data Treatment	Definition	Major Issues
Multiple Imputation	Impute missing values multiple times, to create 40, partially-imputed datasets. Run the analysis on each imputed dataset. Combine the 40 results to get parameter estimates and standard errors.	Unbiased under MCAR and MAR. Improves as you add more variables to the imputation model. Number of variables should be < 100. Accurate SEs. Gives slightly different estimates each time. When used with SEM, suffers more nonconvergences.

MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random

Missing Data

Five practical guidelines for handling missing data

(1) Use all the available data (e.g., do not use listwise deletion).
(2) Do not use single imputation.
(3) For construct-level missingness that exceeds 10% of the sample, ML and multiple imputation (MI) techniques should be used under a strategy that includes auxiliary variables and any hypothesized interaction terms as part of the imputation/estimation model.
(4) For item-level missingness, one item is enough to represent a construct (i.e., do not discard a participant's responses simply because he or she failed to complete some of the items from a multi-item scale).
(5) For person-level missingness that yields a response rate below 30%, simple missing data sensitivity analyses should be conducted

Outliers

Definition

Outliers are patterns in data that do not conform to a well defined notion of normal behavior.

The "***interestingness***" or real life relevance of outliers is a key feature of outlier detection.

Outlier detection is related to

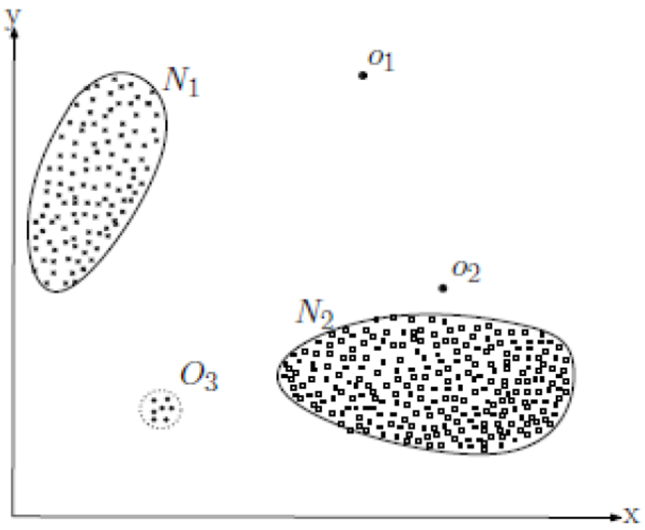
noise removal and noise accommodation

Noise can be defined as a phenomenon in data which is not of interest to the analyst, but acts as a hindrance to data analysis.

Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data

outlier detection is novelty detection which aims at detecting previously unobserved (emergent, novel) patterns in the data

The distinction between novel patterns and outliers is that the novel patterns are typically incorporated into the normal model after being detected.



A simple example of outliers in a 2-dimensional data set.

two normal regions: N1 and N2

points o1 and o2, and points in region O3: outliers

Outliers might be induced in the data for a variety of reasons

Outliers

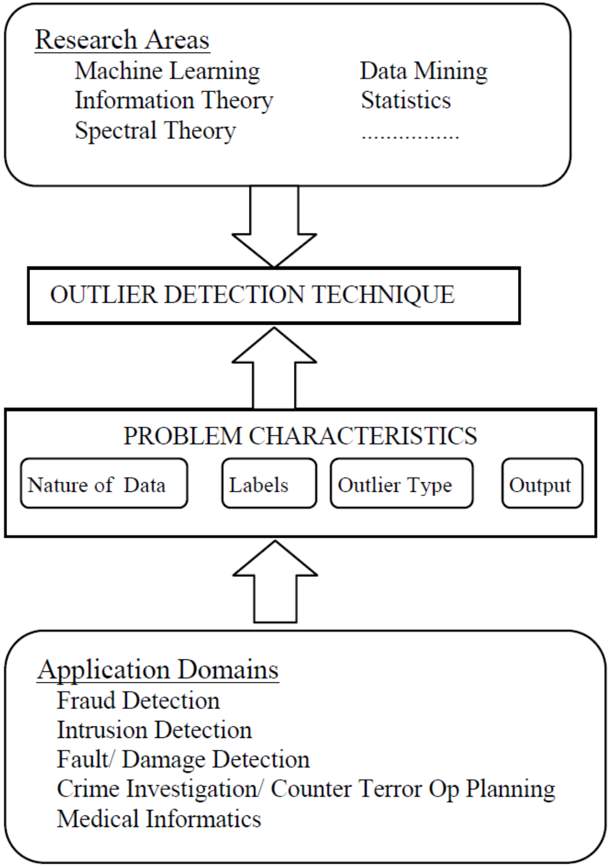
Difficulties in Outlier Detection

- Encompassing of every possible normal behavior in the region.
- Imprecise boundary between normal and outlier behavior since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.
- Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from malicious actions.
- In many domains normal behavior keeps evolving and may not be current to be a
- representative in the future.
- Differing notion of outliers in different application domains makes it difficult to apply technique developed in one domain to another.
- Availability of labeled data for training/validation of models used by outlier detection techniques.
- Noise in the data which tends to be similar to the actual outliers and hence difficult to distinguish and remove.

Outliers

Key components associated with outlier detection technique

The outlier detection problem, in its most general form, is not easy to solve. In fact, most of the existing outlier detection techniques solve a specific problem formulation which is induced by various factors such as nature of the data, availability of labeled data, type of outliers to be detected, etc. Often, these factors are determined by the application domain in which the outliers need to be detected.



Outliers

Aspects Determining the Formulation of Problem

A specific formulation of the problem is determined by several different factors:

Nature of Input Data
Type of Outlier – Point, Contextual, Collective
Data Labels
Output of Outlier Detection.

Outliers

Nature of Input Data

Data instances (object, record, point, vector, pattern, event, case, sample, observation)

↓

Set of attributes (variable, characteristic, feature, field, dimension)

↓

different types (binary, categorical or continuous).

data instance → one attribute (univariate)
→ multiple attributes (multivariate). → The same type
→ A mixture of different data types

The nature of attributes determines the applicability of outlier detection techniques

Outliers

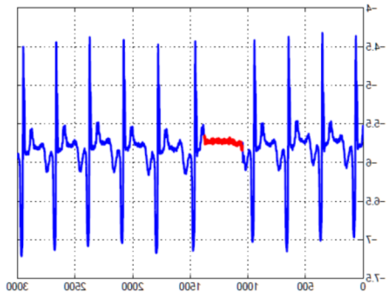
Type of Outlier – Point, Contextual, Collective

Point Outlier - the simplest type of outlier and is the focus of majority of research on outlier detection.

Contextual Outlier - If a data instance is anomalous in a specific con-text (but not otherwise), then it is termed as a contextual outlier.

> **Contextual attributes.** The contextual attributes are used to determine the context (or neighborhood) for that instance (the longitude and latitude of a location are the contextual attributes).
>
> **Behavioral attributes.** The behavioral attributes define the non-contextual characteristics of an instance (the amount of rainfall at any location is a behavioral attribute)

Collective Outlier - If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier (a human electrocardiogram output)



Outliers

Data Labels

The labels associated with a data instance denote if that instance is normal or anomalous.

Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set.

Output of Outlier Detection

An important aspect for any outlier detection technique is the manner in which the outliers are reported.

Scores: Scoring techniques assign an outlier score to each instance in the test data depending on the degree to which that instance is considered an outlier.

Labels: Techniques in this category assign a label (normal or anomalous) to each test instance.

Outliers

Applications of Outlier Detection

Fraud Detection

Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies insurance agencies, cell phone companies, stock market, etc.

Mobile Phone Fraud Detection

In this activity monitoring problem the calling behavior of each account is scanned to issue an alarm when an account appears to have been misused.

Medical and Public Health Outlier Detection

Data Transformation

Reshaping data. Conceptual framework

We think about data in terms of a matrix or data frame, where we have observations in the rows and variables in the columns.

For the purposes of reshaping, we can divide the variables into two groups:

- Identier (id) variables identify the unit that measurements take place on. Id variables are usually discrete, and are typically fixed by design.
- Measured variables represent what is measured on that unit (Y)

Data Transformation

Reshaping data. Conceptual framework

We think about data in terms of a matrix or data frame, where we have observations in the rows and variables in the columns.
For the purposes of reshaping, we can divide the variables into two groups:

- Identier (id) variables identify the unit that measurements take place on. Id variables are usually discrete, and are typically fixed by design.
- Measured variables represent what is measured on that unit (Y)

	subject	time	age	weight	height
1	John Smith	1	33	90	2
2	Mary Smith	1			2

as:

	subject	time	variable	value
1	John Smith	1	age	33
2	John Smith	1	weight	90
3	John Smith	1	height	2
4	Mary Smith	1	height	2

Data Transformation

Reshaping data.

Reshaping data refers to the process of transforming data from one structure or format to another, often to make it more suitable for analysis, visualization, or modeling.
This is a common task in data preprocessing and can involve operations like pivoting, melting, stacking, unstacking, or converting between wide and long formats.

Melting data

- Melting data with id variables encoded in column names
- Already molten data

High-dimensional arrays

Missing Data

Little's missing completely at random (MCAR) test

Description

Use Little's (1988) test statistic to assess if data is missing completely at random (MCAR). The null hypothesis in this test is that the data is MCAR, and the test statistic is a chi-squared value.

R-package **naniar**

Missing values, plotted with the vis_miss function

R-package **mice** (Multivariate Imputation via Chained Equations) - uses chained equations starting with the least missing

R-package **Amelia**

Outliers

R-packages

DMwR
lofactor() function Local Outlier Factor (LOF) is an algorithm used to identify outliers by comparing the local density of a point with that of its neighbors :
car
outlierTest() function gives the most extreme observation based on the given model and allows to test whether it is an outlier
OutlierDetection
Outliers
Mvoutlier
aq.plot() function