WikipediA

Data analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information.[1] In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.

Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination. The term *data analysis* is sometimes used as a synonym for data modeling.

In statistics, **exploratory data analysis** (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Explor

Contents

The process of data analysis
- Data requirements
- Data collection
- Data processing
- Data cleaning
- Exploratory data analysis
- Modeling and algorithms
- Data product
- Communication

Quantitative messages

Techniques for analyzing quantitative data

Analytical activities of data users

Barriers to effective analysis
- Confusing fact and opinion
- Cognitive biases

- Innumeracy

Other topics

- Smart buildings
- Analytics and business intelligence
- Education

Practitioner notes

- Initial data analysis
  - Quality of data
  - Quality of measurements
  - Initial transformations
  - Did the implementation of the study fulfill the intentions of the research design?
  - Characteristics of data sample
  - Final stage of the initial data analysis
  - Analysis
  - Nonlinear analysis
- Main data analysis
  - Exploratory and confirmatory approaches
  - Stability of results
  - Statistical methods

Free software for data analysis

International data analysis contests

See also

References

- Citations
- Bibliography

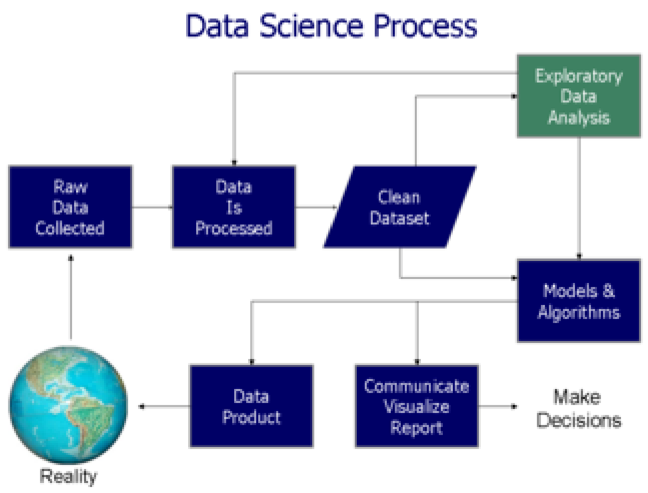
Further reading

The process of data analysis

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.[2]

Statistician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."[3]

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.[4]



Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013

Data requirements

The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).[4]

Data collection

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.[4]

Data processing

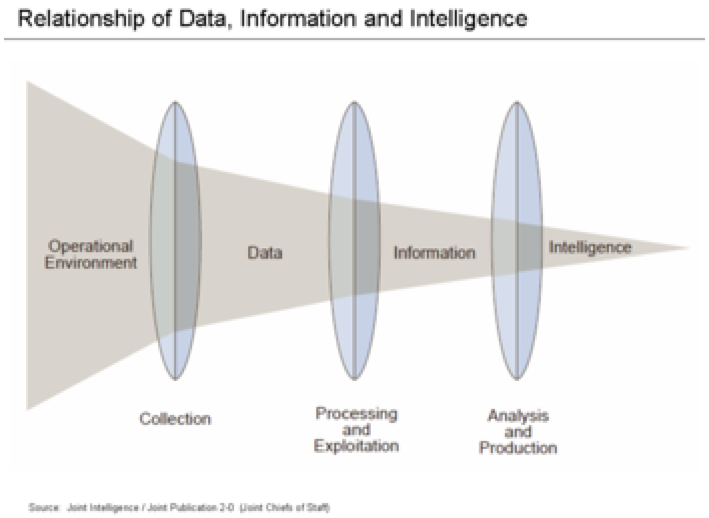
Data initially obtained must be processed or organised for analysis. For instance, these may involve placing data into rows and columns in a table format (i.e., structured data) for further analysis, such as within a spreadsheet or statistical software.[4]

Data cleaning

Once processed and organised, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data,[5] deduplication, and column segmentation.[6] Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable.[7] Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data such as phone numbers, email addresses, employers etc. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spell checkers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.[8]

Exploratory data analysis

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data.[9][10] The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics, such as the average or median, may be generated to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.[4]


Relationship of Data, Information and Intelligence
Operational Environment
Data
Information
Intelligence
Collection
Processing and Exploitation
Analysis and Production
Source: Joint Intelligence / Joint Publication 2-0 (Joint Chiefs of Staff)

The phases of the intelligence cycle used to convert raw information into actionable intelligence or knowledge are conceptually similar to the phases in data analysis.

Modeling and algorithms

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).[2]

Inferential statistics includes techniques to measure relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (independent variable X) explains the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising). It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X. Analysts may attempt to build models that are descriptive of the data to simplify analysis and communicate results.[2]

Data product

A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm. An example is an application that analyzes data about customer purchasing history and recommends other purchases the customer might enjoy.[4]

Communication

Once the data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.[4]

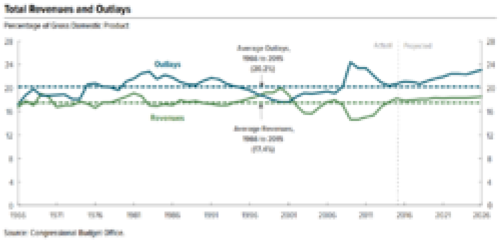
When determining how to communicate the results, the analyst may consider data visualization techniques to help clearly and efficiently communicate the message to the audience. Data visualization uses information displays (such as tables and charts) to help communicate key messages contained in the data. Tables are helpful to a user who might lookup specific numbers, while charts (e.g., bar charts or line charts) may help explain the quantitative messages contained in the data.

Quantitative messages

Data visualization to understand the results of a data analysis.[11]

Stephen Few described eight types of quantitative messages that users may attempt to understand or communicate from a set of data and the associated graphs used to help communicate the message. Customers specifying requirements and analysts performing the data analysis may consider these messages during the course of the process.

1. Time-series: A single variable is captured over a period of time, such as the unemployment rate over a 10-year period. A line chart may be used to demonstrate the trend.
2. Ranking: Categorical subdivisions are ranked in ascending or descending order, such as a ranking of sales performance (the *measure*) by sales persons (the *category*, with each sales person a *categorical subdivision*) during a single period. A bar chart may be used to show the comparison across the sales persons.
3. Part-to-whole: Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%). A pie chart or bar chart can show the comparison of ratios, such as the market share represented by competitors in a market.
4. Deviation: Categorical subdivisions are compared against a reference, such as a comparison of actual vs. budget expenses for several departments of a business for a given time period. A bar chart can show comparison of the actual versus the reference amount.
5. Frequency distribution: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0–10%, 11–20%, etc. A histogram, a type of bar chart, may be used for this analysis.
6. Correlation: Comparison between observations represented by two variables (X,Y) to determine if they tend to move in the same or opposite directions. For example, plotting unemployment (X) and inflation (Y) for a sample of months. A scatter plot is typically used for this message.
7. Nominal comparison: Comparing categorical subdivisions in no particular order, such as the sales volume by product code. A bar chart may be used for this comparison.
8. Geographic or geospatial: Comparison of a variable across a map or layout, such as the unemployment rate by state or the number of persons on the various floors of a building. A cartogram is a typical graphic used.[12][13]



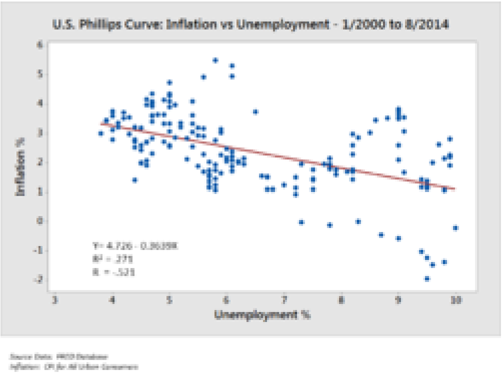
A time series illustrated with a line chart demonstrating trends in U.S. federal spending and revenue over time.

Techniques for analyzing quantitative data

Author Jonathan Koomey has recommended a series of best practices for understanding quantitative data. These include:

- Check raw data for anomalies prior to performing your analysis;
- Re-perform important calculations, such as verifying columns of data that are formula driven;
- Confirm main totals are the sum of subtotals;
- Check relationships between numbers that should be related in a predictable way, such as ratios over time;
- Normalize numbers to make comparisons easier, such as analyzing amounts per person or relative to GDP or as an index value relative to a base year;
- Break problems into component parts by analyzing factors that led to the results, such as DuPont analysis of return on equity.[7]

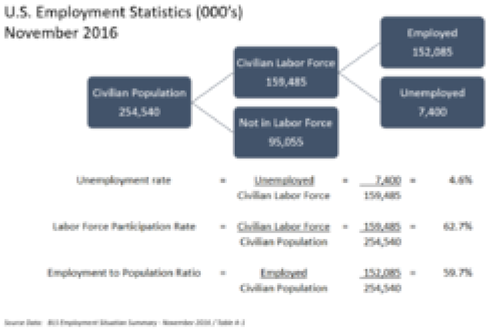
For the variables under examination, analysts typically obtain descriptive statistics for them, such as the mean (average), median, and standard deviation. They may also analyze the distribution of the key variables to see how the individual values cluster around the mean.



A scatterplot illustrating correlation between two variables (inflation and unemployment) measured at points in time.

The consultants at McKinsey and Company named a technique for breaking a quantitative problem down into its component parts called the MECE principle. Each layer can be broken down into its components; each of the sub-components must be mutually exclusive of each other and collectively add up to the layer above them. The relationship is referred to as "Mutually Exclusive and Collectively Exhaustive" or MECE. For example, profit by definition can be broken down into total revenue and total cost. In turn, total revenue can be analyzed by its components, such as revenue of divisions A, B, and C (which are mutually exclusive of each other) and should add to the total revenue (collectively exhaustive).

Analysts may use robust statistical measurements to solve certain analytical problems. Hypothesis testing is used when a particular hypothesis about the true state of affairs is made by the analyst and data is gathered to determine whether that state of affairs is true or false. For example, the hypothesis might be that "Unemployment has no effect on inflation", which relates to an economics concept called the Phillips Curve. Hypothesis testing involves considering the likelihood of Type I and type II errors, which relate to whether the data supports accepting or rejecting the hypothesis.


U.S. Employment Statistics (000's)
November 2016
Civilian Population
254,540
Civilian Labor Force
159,485
Not in Labor Force
95,055
Employed
152,085
Unemployed
7,400
Unemployment rate = Unemployed / Civilian Labor Force = 7,400 / 159,485 = 4.6%
Labor Force Participation Rate = Civilian Labor Force / Civilian Population = 159,485 / 254,540 = 62.7%
Employment to Population Ratio = Employed / Civilian Population 152,085 / 254,540 = 59.7%
Source Data: BLS Employment Situation Summary - November 2016 / Table A-1

An illustration of the MECE principle used for data analysis.

Regression analysis may be used when the analyst is trying to determine the extent to which independent variable X affects dependent variable Y (e.g., "To what extent do changes in the unemployment rate (X) affect the inflation rate (Y)?"). This is an attempt to model or fit an equation line or curve to the data, such that Y is a function of X.

Necessary condition analysis (https://www.erim.eur.nl/centres/necessary-condition-analysis/) (NCA) may be used when the analyst is trying to determine the extent to which independent variable X allows variable Y (e.g., "To what extent is a certain unemployment rate (X) necessary for a certain inflation rate (Y)?"). Whereas (multiple) regression analysis uses additive logic where each X-variable can produce the outcome and the X's can compensate for each other (they are sufficient but not necessary), necessary condition analysis (NCA) uses necessity logic, where one or more X-variables allow the outcome to exist, but may not produce it (they are necessary but not sufficient). Each single necessary condition must be present and compensation is not possible.

Analytical activities of data users

Users may have particular data points of interest within a data set, as opposed to general messaging outlined above. Such low-level user analytic activities are presented in the following table. The taxonomy can also be organized by three poles of activities: retrieving values, finding data points, and arranging data points.[14][15][16][17]

#	Task	General Description	Pro Forma Abstract	Examples
1	**Retrieve Value**	Given a set of specific cases, find attributes of those cases.	What are the values of attributes {X, Y, Z, ...} in the data cases {A, B, C, ...}?	*- What is the mileage per gallon of the Ford Mondeo?* *- How long is the movie Gone with the Wind?*
2	**Filter**	Given some concrete conditions on attribute values, find data cases satisfying those conditions.	Which data cases satisfy conditions {A, B, C...}?	*- What Kellogg's cereals have high fiber?* *- What comedies have won awards?* *- Which funds underperformed the SP-500?*
3	**Compute Derived Value**	Given a set of data cases, compute an aggregate numeric representation of those data cases.	What is the value of aggregation function F over a given set S of data cases?	*- What is the average calorie content of Post cereals?* *- What is the gross income of all stores combined?* *- How many manufacturers of cars are there?*
4	**Find Extremum**	Find data cases possessing an extreme value of an attribute over its range within the data set.	What are the top/bottom N data cases with respect to attribute A?	*- What is the car with the highest MPG?* *- What director/film has won the most awards?* *- What Marvel Studios film has the most recent release date?*
5	**Sort**	Given a set of data cases, rank them according to some ordinal metric.	What is the sorted order of a set S of data cases according to their value of attribute A?	*- Order the cars by weight.* *- Rank the cereals by calories.*
6	**Determine Range**	Given a set of data cases and an attribute of interest, find the span of values within the set.	What is the range of values of attribute A in a set S of data cases?	*- What is the range of film lengths?* *- What is the range of car horsepowers?* *- What actresses are in the data set?*
7	**Characterize Distribution**	Given a set of data cases and a quantitative attribute of interest, characterize the distribution of that attribute's values over the set.	What is the distribution of values of attribute A in a set S of data cases?	*- What is the distribution of carbohydrates in cereals?* *- What is the age distribution of shoppers?*
8	**Find Anomalies**	Identify any anomalies within a given set of data cases with respect to a	Which data cases in a set S of data cases have	*- Are there exceptions to the relationship between horsepower and acceleration?* *- Are there any outliers in protein?*

		given relationship or expectation, e.g. statistical outliers.	unexpected/exceptional values?	
9	**Cluster**	Given a set of data cases, find clusters of similar attribute values.	Which data cases in a set S of data cases are similar in value for attributes {X, Y, Z, ...}?	*- Are there groups of cereals w/ similar fat/calories/sugar?* *- Is there a cluster of typical film lengths?*
10	**Correlate**	Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.	What is the correlation between attributes X and Y over a given set S of data cases?	*- Is there a correlation between carbohydrates and fat?* *- Is there a correlation between country of origin and MPG?* *- Do different genders have a preferred payment method?* *- Is there a trend of increasing film length over the years?*
11	**Contextualization**[17]	Given a set of data cases, find contextual relevancy of the data to the users.	Which data cases in a set S of data cases are relevant to the current users' context?	*- Are there groups of restaurants that have foods based on my current caloric intake?*

Barriers to effective analysis

Barriers to effective analysis may exist among the analysts performing the data analysis or among the audience. Distinguishing fact from opinion, cognitive biases, and innumeracy are all challenges to sound data analysis.

Confusing fact and opinion

> You are entitled to your own opinion, but you are not entitled to your own facts.
>
> Daniel Patrick Moynihan

Effective analysis requires obtaining relevant facts to answer questions, support a conclusion or formal opinion, or test hypotheses. Facts by definition are irrefutable, meaning that any person involved in the analysis should be able to agree upon them. For example, in August 2010, the Congressional Budget Office (CBO) estimated that extending the Bush tax cuts of 2001 and 2003 for the 2011–2020 time period would add approximately $3.3 trillion to the national debt.[18] Everyone should be able to agree that indeed this is what CBO reported; they can all examine the report. This makes it a fact. Whether persons agree or disagree with the CBO is their own opinion.

As another example, the auditor of a public company must arrive at a formal opinion on whether financial statements of publicly traded corporations are "fairly stated, in all material respects." This requires extensive analysis of factual data and evidence to support their opinion. When making the leap from facts to opinions, there is always the possibility that the opinion is erroneous.

Cognitive biases

There are a variety of cognitive biases that can adversely affect analysis. For example, confirmation bias is the tendency to search for or interpret information in a way that confirms one's preconceptions. In addition, individuals may discredit information that does not support their views.

Analysts may be trained specifically to be aware of these biases and how to overcome them. In his book *Psychology of Intelligence Analysis*, retired CIA analyst Richards Heuer wrote that analysts should clearly delineate their assumptions and chains of inference and specify the degree and source of the uncertainty involved in the conclusions. He emphasized procedures to help surface and debate alternative points of view.[19]

Innumeracy

Effective analysts are generally adept with a variety of numerical techniques. However, audiences may not have such literacy with numbers or numeracy; they are said to be innumerate. Persons communicating the data may also be attempting to mislead or misinform, deliberately using bad numerical techniques.[20]

For example, whether a number is rising or falling may not be the key factor. More important may be the number relative to another number, such as the size of government revenue or spending relative to the size of the economy (GDP) or the amount of cost relative to revenue in corporate financial statements. This numerical technique is referred to as normalization[7] or common-sizing. There are many such techniques employed by analysts, whether adjusting for inflation (i.e., comparing real vs. nominal data) or considering population increases, demographics, etc. Analysts apply a variety of techniques to address the various quantitative messages described in the section above.

Analysts may also analyze data under different assumptions or scenarios. For example, when analysts perform financial statement analysis, they will often recast the financial statements under different assumptions to help arrive at an estimate of future cash flow, which they then discount to present value based on some interest rate, to determine the valuation of the company or its stock. Similarly, the CBO analyzes the effects of various policy options on the government's revenue, outlays and deficits, creating alternative future scenarios for key measures.

Other topics

Smart buildings

A data analytics approach can be used in order to predict energy consumption in buildings.[21] The different steps of the data analysis process are carried out in order to realise smart buildings, where the building management and control operations including heating, ventilation, air conditioning, lighting and security are realised automatically by miming the needs of the building users and optimising resources like energy and time.

Analytics and business intelligence

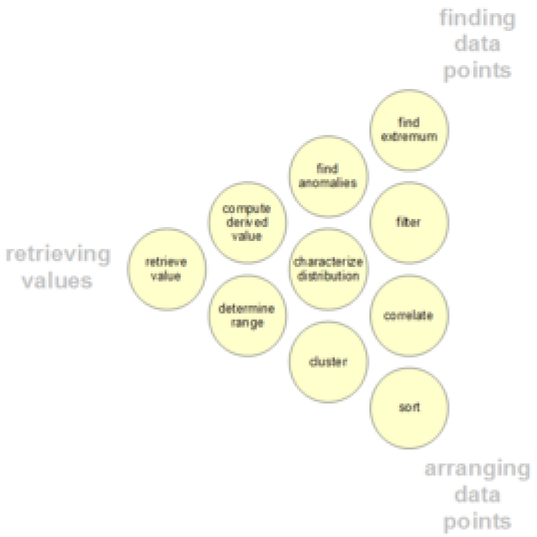
Analytics is the "extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions." It is a subset of business intelligence, which is a set of technologies and processes that use data to understand and analyze business performance.[22]

Education

In education, most educators have access to a data system for the purpose of analyzing student data.[23] These data systems present data to educators in an over-the-counter data format (embedding labels, supplemental documentation, and a help system and making key package/display and content decisions) to improve the accuracy of educators' data analyses.[24]

Practitioner notes

This section contains rather technical explanations that may assist practitioners but are beyond the typical scope of a Wikipedia article.



Analytic activities of data visualization users

Initial data analysis

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that is aimed at answering the original research question. The initial data analysis phase is guided by the following four questions:[25]

Quality of data

The quality of the data should be checked as early as possible. Data quality can be assessed in several ways, using different types of analysis: frequency counts, descriptive statistics (mean, standard deviation, median), normality (skewness, kurtosis, frequency histograms, n: variables are compared with coding schemes of variables external to the data set, and possibly corrected if coding schemes are not comparable.

- Test for common-method variance.

The choice of analyses to assess the data quality during the initial data analysis phase depends on the analyses that will be conducted in the main analysis phase.[26]

Quality of measurements

The quality of the measurement instruments should only be checked during the initial data analysis phase when this is not the focus or research question of the study. One should check whether structure of measurement instruments corresponds to structure reported in the literature.

There are two ways to assess measurement: [NOTE: only one way seems to be listed]

- Analysis of homogeneity (internal consistency), which gives an indication of the reliability of a measurement instrument. During this analysis, one inspects the variances of the items and the scales, the Cronbach's α of the scales, and the change in the Cronbach's alpha when an item would be deleted from a scale[27]

Initial transformations

After assessing the quality of the data and of the measurements, one might decide to impute missing data, or to perform initial transformations of one or more variables, although this can also be done during the main analysis phase.[28]
Possible transformations of variables are:[29]

- Square root transformation (if the distribution differs moderately from normal)
- Log-transformation (if the distribution differs substantially from normal)
- Inverse transformation (if the distribution differs severely from normal)
- Make categorical (ordinal / dichotomous) (if the distribution differs severely from normal, and no transformations help)

Did the implementation of the study fulfill the intentions of the research design?

One should check the success of the randomization procedure, for instance by checking whether background and substantive variables are equally distributed within and across groups.
If the study did not need or use a randomization procedure, one should check the success of the non-random sampling, for instance by checking whether all subgroups of the population of interest are represented in sample.
Other possible data distortions that should be checked are:

- dropout (this should be identified during the initial data analysis phase)
- Item nonresponse (whether this is random or not should be assessed during the initial data analysis phase)
- Treatment quality (using manipulation checks).[30]

Characteristics of data sample

In any report or article, the structure of the sample must be accurately described. It is especially important to exactly determine the structure of the sample (and specifically the size of the subgroups) when subgroup analyses will be performed during the main analysis phase.
The characteristics of the data sample can be assessed by looking at:

- Basic statistics of important variables
- Scatter plots
- Correlations and associations
- Cross-tabulations[31]

Final stage of the initial data analysis

During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken.

Also, the original plan for the main data analyses can and should be specified in more detail or rewritten.

In order to do this, several decisions about the main data analyses can and should be made:

- In the case of non-normals: should one transform variables; make variables categorical (ordinal/dichotomous); adapt the analysis method?
- In the case of missing data: should one neglect or impute the missing data; which imputation technique should be used?
- In the case of outliers: should one use robust analysis techniques?
- In case items do not fit the scale: should one adapt the measurement instrument by omitting items, or rather ensure comparability with other (uses of the) measurement instrument(s)?
- In the case of (too) small subgroups: should one drop the hypothesis about inter-group differences, or use small sample techniques, like exact tests or bootstrapping?
- In case the randomization procedure seems to be defective: can and should one calculate propensity scores and include them as covariates in the main analyses?[32]

Analysis

Several analyses can be used during the initial data analysis phase:[33]

- Univariate statistics (single variable)
- Bivariate associations (correlations)
- Graphical techniques (scatter plots)

It is important to take the measurement levels of the variables into account for the analyses, as special statistical techniques are available for each level:[34]

- Nominal and ordinal variables
  - Frequency counts (numbers and percentages)
  - Associations
    - circumambulations (crosstabulations)
    - hierarchical loglinear analysis (restricted to a maximum of 8 variables)
    - loglinear analysis (to identify relevant/important variables and possible confounders)
  - Exact tests or bootstrapping (in case subgroups are small)
  - Computation of new variables
- Continuous variables
  - Distribution
    - Statistics (M, SD, variance, skewness, kurtosis)
    - Stem-and-leaf displays
    - Box plots

Nonlinear analysis

Nonlinear analysis will be necessary when the data is recorded from a nonlinear system. Nonlinear systems can exhibit complex dynamic effects including bifurcations, chaos, harmonics and subharmonics that cannot be analyzed using simple linear methods. Nonlinear data analysis is closely related to nonlinear system identification.[35]

Main data analysis

In the main analysis phase analyses aimed at answering the research question are performed as well as any other relevant analysis needed to write the first draft of the research report.[36]

Exploratory and confirmatory approaches

In the main analysis phase either an exploratory or confirmatory approach can be adopted. Usually the approach is decided before data is collected. In an exploratory analysis no clear hypothesis is stated before analysing the data, and the data is searched for models that describe the data well. In a confirmatory analysis clear hypotheses about the data are tested.

Exploratory data analysis should be interpreted carefully. When testing multiple models at once there is a high chance on finding at least one of them to be significant, but this can be due to a type 1 error. It is important to always adjust the significance level when testing multiple models with, for example, a Bonferroni correction. Also, one should not follow up an exploratory analysis with a confirmatory analysis in the same dataset. An exploratory analysis is used to find ideas for a theory, but not to test that theory as well. When a model is found exploratory in a dataset, then following up that analysis with a confirmatory analysis in the same dataset could simply mean that the results of the confirmatory analysis are due to the same type 1 error that resulted in the exploratory model in the first place. The confirmatory analysis therefore will not be more informative than the original exploratory analysis.[37]

Stability of results

It is important to obtain some indication about how generalizable the results are.[38] While this is hard to check, one can look at the stability of the results. Are the results reliable and reproducible? There are two main ways of doing this:

- Cross-validation: By splitting the data in multiple parts we can check if an analysis (like a fitted model) based on one part of the data generalizes to another part of the data as well.
- Sensitivity analysis: A procedure to study the behavior of a system or model when global parameters are (systematically) varied. One way to do this is with bootstrapping.

Statistical methods

Many statistical methods have been used for statistical analyses. A very brief list of four of the more popular methods is:

- General linear model: A widely used model on which various methods are based (e.g. t test, ANOVA, ANCOVA, MANOVA). Usable for assessing the effect of several predictors on one or more continuous dependent variables.
- Generalized linear model: An extension of the general linear model for discrete dependent variables.
- Structural equation modelling: Usable for assessing latent structures from measured manifest variables.
- Item response theory: Models for (mostly) assessing one latent variable from several binary measured variables (e.g. an exam).

Free software for data analysis

- DevInfo – a database system endorsed by the United Nations Development Group for monitoring and analyzing human development.
- ELKI – data mining framework in Java with data mining oriented visualization functions.
- KNIME – the Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- Orange – A visual programming tool featuring interactive data visualization and methods for statistical data analysis, data mining, and machine learning.
- PAST (https://folk.uio.no/ohammer/past/) – free software for scientific data analysis
- PAW – FORTRAN/C data analysis framework developed at CERN
- R – a programming language and software environment for statistical computing and graphics.
- ROOT – C++ data analysis framework developed at CERN
- SciPy and Pandas – Python libraries for data analysis

International data analysis contests

Different companies or organizations hold a data analysis contests to encourage researchers utilize their data or to solve a particular question using data analysis. A few examples of well-known international data analysis contests are as follows.

- Kaggle competition held by Kaggle[39]
- LTPP data analysis contest held by FHWA and ASCE.[40][41]

See also

- Actuarial science
- Analytics
- Big data
- Business intelligence
- Censoring (statistics)
- Computational physics
- Data acquisition
- Data blending
- Data governance
- Data mining
- Data Presentation Architecture
- Data science
- Digital signal processing
- Dimension reduction
- Early case assessment
- Exploratory data analysis
- Fourier analysis
- Machine learning
- Multilinear PCA
- Multilinear subspace learning

- Multiway data analysis
- Nearest neighbor search
- Nonlinear system identification
- Predictive analytics
- Principal component analysis
- Qualitative research
- Scientific computing
- Structured data analysis (statistics)
- System identification
- Test method
- Text analytics
- Unstructured data
- Wavelet

References

Citations

1. Exploring Data Analysis (https://spotlessdata.com/blog/exploring-data-analysis)
2. Judd, Charles and, McCleland, Gary (1989). *Data Analysis*. Harcourt Brace Jovanovich. ISBN 0-15-516765-0.
3. John Tukey-The Future of Data Analysis-July 1961 (http://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711)
4. O'Neil, Cathy and, Schutt, Rachel (2013). *Doing Data Science*. O'Reilly. ISBN 978-1-449-35865-5.
5. Clean Data in CRM: The Key to Generate Sales-Ready Leads and Boost Your Revenue Pool (https://www.suntecindia.com/blog/clean-data-in-crm-the-key-to-generate-sales-ready-leads-and-boost-your-revenue-pool/) Retrieved 29th July, 2016
6. "Data Cleaning" (http://research.microsoft.com/en-us/projects/datacleaning/). Microsoft Research. Retrieved 26 October 2013.
7. Perceptual Edge-Jonathan Koomey-Best practices for understanding quantitative data-February 14, 2006 (http://www.perceptualedge.com/articles/b-eye/quantitative_data.pdf)
8. Hellerstein, Joseph (27 February 2008). "Quantitative Data Cleaning for Large Databases" (http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf) (PDF). *EECS Computer Science Division*: 3. Retrieved 26 October 2013.
9. Stephen Few-Perceptual Edge-Selecting the Right Graph For Your Message-September 2004 (http://www.perceptualedge.com/articles/ie/the_right_graph.pdf)
10. Behrens-Principles and Procedures of Exploratory Data Analysis-American Psychological Association-1997 (http://cll.stanford.edu/~willb/course/behrens97pm.pdf)
11. Grandjean, Martin (2014). "La connaissance est un réseau" (http://www.martingrandjean.ch/wp-content/uploads/2015/02/Grandjean-2014-Connaissance-reseau.pdf) (PDF). *Les Cahiers du Numérique*. **10** (3): 37–54. doi:10.3166/lcn.10.3.37-54 (https://doi.org/10.3166/lcn.10.3.37-54).
12. Stephen Few-Perceptual Edge-Selecting the Right Graph for Your Message-2004 (http://www.perceptualedge.com/articles/ie/the_right_graph.pdf)
13. Stephen Few-Perceptual Edge-Graph Selection Matrix (http://www.perceptualedge.com/articles/misc/Graph_Selection_Matrix.pdf)
14. Robert Amar, James Eagan, and John Stasko (2005) "Low-Level Components of Analytic Activity in Information Visualization" (http://www.cc.gatech.edu/~stasko/papers/infovis05.pdf)
15. William Newman (1994) "A Preliminary Analysis of the Products of HCI Research, Using Pro Forma Abstracts" (http://www.mdnpress.com/wmn/pdfs/chi94-pro-formas-2.pdf)
16. Mary Shaw (2002) "What Makes Good Research in Software Engineering?" (http://www.cs.cmu.edu/~Compose/ftp/shaw-fin-etaps.pdf)

17. "ConTaaS: An Approach to Internet-Scale Contextualisation for Developing Efficient Internet of Things Applications" (https://scholarspace.manoa.hawaii.edu/handle/10125/41879). *ScholarSpace*. HICSS50. Retrieved May 24, 2017.
18. "Congressional Budget Office-The Budget and Economic Outlook-August 2010-Table 1.7 on Page 24" (http://www.cbo.gov/publication/21670) (PDF). Retrieved 2011-03-31.
19. "Introduction" (https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/art3.html). *cia.gov*.
20. Bloomberg-Barry Ritholz-Bad Math that Passes for Insight-October 28, 2014 (http://www.bloombergview.com/articles/2014-10-28/bad-math-that-passes-for-insight)
21. González-Vidal, Aurora; Moreno-Cano, Victoria (2016). "Towards energy efficiency smart buildings models based on intelligent data analytics". *Procedia Computer Science*. **83** (Elsevier): 994–999. doi:10.1016/j.procs.2016.04.213 (https://doi.org/10.1016/j.procs.2016.04.213).
22. Davenport, Thomas and, Harris, Jeanne (2007). *Competing on Analytics*. O'Reilly. ISBN 978-1-4221-0332-6.
23. Aarons, D. (2009). Report finds states on course to build pupil-data systems. (http://search.proquest.com/docview/202710770?accountid=28180) *Education Week, 29*(13), 6.
24. Rankin, J. (2013, March 28). How data Systems & reports can either fight or propagate the data analysis error epidemic, and how educator leaders can help. (https://sas.elluminate.com/site/external/recording/playback/link/table/dropin?sid=2008350&suid=D.4DF60C7117D5A77FE3AED546909ED2) *Presentation conducted from Technology Information Center for Administrative Leadership (TICAL) School Leadership Summit.*
25. Adèr 2008a, p. 337.
26. Adèr 2008a, pp. 338-341.
27. Adèr 2008a, pp. 341-342.
28. Adèr 2008a, p. 344.
29. Tabachnick & Fidell, 2007, p. 87-88.
30. Adèr 2008a, pp. 344-345.
31. Adèr 2008a, p. 345.
32. Adèr 2008a, pp. 345-346.
33. Adèr 2008a, pp. 346-347.
34. Adèr 2008a, pp. 349-353.
35. Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013
36. Adèr 2008b, p. 363.
37. Adèr 2008b, pp. 361-362.
38. Adèr 2008b, pp. 361-371.
39. "The machine learning community takes on the Higgs" (http://www.symmetrymagazine.org/article/july-2014/the-machine-learning-community-takes-on-the-higgs/). *Symmetry Magazine*. July 15, 2014. Retrieved 14 January 2015.
40. Nehme, Jean (September 29, 2016). "LTPP International Data Analysis Contest" (https://www.fhwa.dot.gov/research/tfhrc/programs/infrastructure/pavements/ltpp/2016_2017_asce_ltpp_contest_guidelines.cfm). Federal Highway Administration. Retrieved October 22, 2017.

41. "Data.Gov:Long-Term Pavement Performance (LTPP)" (https://www.fhwa.dot.gov/research/tfhrc/programs/infrastructure/pavements/ltpp/). May 26, 2016. Retrieved November 10, 2017.

Bibliography

- Adèr, Herman J. (2008a). "Chapter 14: Phases and initial steps in data analysis". In Adèr, Herman J.; Mellenbergh, Gideon J.; Hand, David J. *Advising on research methods : a consultant's companion* (http://www.worldcat.org/title/advising-on-research-methods-a-consultants-companion/oclc/905799857/viewport). Huizen, Netherlands: Johannes van Kessel Pub. pp. 333–356. ISBN 9789079418015. OCLC 905799857 (https://www.worldcat.org/oclc/905799857).
- Adèr, Herman J. (2008b). "Chapter 15: The main analysis phase". In Adèr, Herman J.; Mellenbergh, Gideon J.; Hand, David J. *Advising on research methods : a consultant's companion* (http://www.worldcat.org/title/advising-on-research-methods-a-consultants-companion/oclc/905799857/viewport). Huizen, Netherlands: Johannes van Kessel Pub. pp. 357–386. ISBN 9789079418015. OCLC 905799857 (https://www.worldcat.org/oclc/905799857).
- Tabachnick, B.G. & Fidell, L.S. (2007). Chapter 4: Cleaning up your act. Screening data prior to analysis. In B.G. Tabachnick & L.S. Fidell (Eds.), Using Multivariate Statistics, Fifth Edition (pp. 60–116). Boston: Pearson Education, Inc. / Allyn and Bacon.

Further reading

- Adèr, H.J. & Mellenbergh, G.J. (with contributions by D.J. Hand) (2008). *Advising on Research Methods: A Consultant's Companion*. Huizen, the Netherlands: Johannes van Kessel Publishing.
- Chambers, John M.; Cleveland, William S.; Kleiner, Beat; Tukey, Paul A. (1983). *Graphical Methods for Data Analysis*, Wadsworth/Duxbury Press. ISBN 0-534-98052-X
- Fandango, Armando (2008). *Python Data Analysis, 2nd Edition*. Packt Publishers.
- Juran, Joseph M.; Godfrey, A. Blanton (1999). *Juran's Quality Handbook, 5th Edition.* New York: McGraw Hill. ISBN 0-07-034003-X
- Lewis-Beck, Michael S. (1995). *Data Analysis: an Introduction*, Sage Publications Inc, ISBN 0-8039-5772-6
- NIST/SEMATECH (2008) *Handbook of Statistical Methods* (http://www.itl.nist.gov/div898/handbook/),
- Pyzdek, T, (2003). *Quality Engineering Handbook*, ISBN 0-8247-4614-7
- Richard Veryard (1984). *Pragmatic Data Analysis*. Oxford : Blackwell Scientific Publications. ISBN 0-632-01311-7
- Tabachnick, B.G.; Fidell, L.S. (2007). *Using Multivariate Statistics, 5th Edition*. Boston: Pearson Education, Inc. / Allyn and Bacon, ISBN 978-0-205-45938-4

Retrieved from "https://en.wikipedia.org/w/index.php?title=Data_analysis&oldid=853179160"

This page was last edited on 2 August 2018, at 23:32 (UTC).