Data Integrity

Freitag, 11. Juli 2025 10:29

Data Integrity
L> accuracy, completeness, consistency, and trustworthiness of data throughout it's lifecycle
L> sometimes one missing thing can make whole data useless

Data Replication
L> process of storing data in multiple locations
L> if not all people use the same data
L> might cause problems

Data Transfer
L> process of copying data from a storage device to memory
L> or from one computer to another

Data Manipulation
L> process of changing data to make it more organised and easier to read

Other Threads
L> Human error
L> Viruses
L> Malware
L> Hacking
L> System Failures

Example
L> Data analyst working worldwide
L> not checking on the different types of time formats
L> might cause big problems in dating
L> Data Replication: Wrong dates lead to copying of incomplete data
L> Data Transfer: Incorrect classifying makes imports fail
L> Data Manipulation: analyst deletes "duplicated" reports, which are not

Data Constrains *(criteria that determine validity)*

Data constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid
Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid
Mandatory	Values can't be left blank or empty	If age is mandatory, that value must be filled in
Unique	Values can't have a duplicate	Two people can't have the same mobile phone number within the same service area
Regular expression (regex) patterns	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)
Cross-field validation	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
Primary-key	(Databases only) value must be unique per column	A database table can't have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.
Set-membership	(Databases only) values for a column must come from a set of discrete values	Value for a column must be set to Yes, No, or Not Applicable
Foreign-key	(Databases only) values for a column must be unique values coming from a column in another table	In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table
Accuracy	The degree to which the data conforms to the actual entity being measured or described	If values for zip codes are validated by street location, the accuracy of the data goes up.
Completeness	The degree to which the data contains all desired components or measures	If data for personal profiles required hair and eye color, and both are collected, the data is complete.
Consistency	The degree to which the data is repeatable from different points of entry or collection	If a customer has the same address in the sales and repair databases, the data is consistent.

Datei auswählen Keine ausgewählt

End

Insufficient Data

Freitag, 11. Juli 2025 10:29

Types of insufficient data
L> data from only one source
L> data keeps updating *(not complete)*
L> outdated data *(find new dataset to work with)*
L> geographically-limited data *(no local data for global company)*

Ways to address insufficient data
L> identify trends with the available data *(and then qualify findings)*
L> wait for more data if time allows
L> talk with stakeholders and adjust your objective
L> look for a new dataset

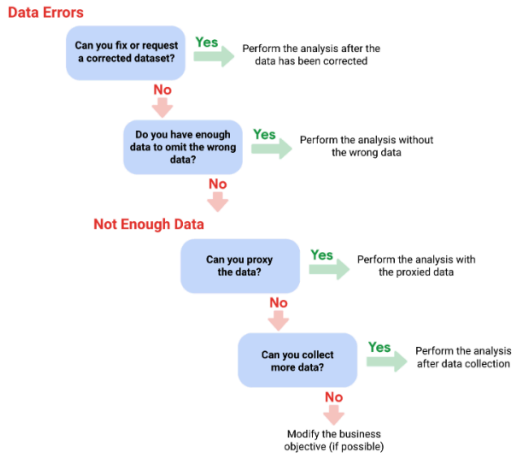
Proxy Data
L> data from other datasets, which can perform the same job
L> the most common workaround if there is no data

Mixing Proxy Data with Actual Data
L> including labrador data, when golden retriever is not enough
L> increase the range of age for an analysis
L> lots of possibilities

Having the wrong data
L> restate your needs - redefine what you settled at the beginning
L> you can try to correct the error or look for new patterns
L> you can try to ignore bad data and continue with good data
L> but you have to make sure that the data won't cause systematic bias

Sometimes data with errors can be a warning sign that it's not reliable!

Use the following decision tree as a reminder of how to deal with data errors or not enough data:



End

Datei auswählen Keine ausgewählt

Sample Size Intro

Freitag, 11. Juli 2025 10:29

Key Information

The Minimum number of sample size
L> don't use a sample size less than 30!
L> it's proven! -> by the Central Limit Theorem (CLT)
L> the optimal confidence level is at 95%

Use A Larger Sample Size
L> for a higher confidence level
L> for decreasing the margin of error
L> for greater statistical significance
L> but it can be **more expensive!**

Sample Size - What To Know
L> Sampling Sizes actually takes place before we even get the data
but it's good to know the backgrounds and be aware
L> the degree to which we can be confident
L> but can lead to uncertainty
L> can lead to sampling bias

Term Definitions

Population
L> all possible data values in a certain dataset

Sample
L> subset of a population

Sample Size
L> a part of a population that is representative of the population
L> goal is to get enough information from a small group in the population
L> so that we can make predictions for the whole population

Sampling Bias
L> a sample is not representative of the population as a whole
L> some members of population might be over- or underrepresented
L> e.g. cat owners only with smartphones represented, without not

Random Sampling
L> way of selecting a sample from a population
L> so that every possible type of the sample
L> has an equal chance of being chosen

Margin of error
L> the difference between the population and the sample

Confidence level
L> in percent
L> how confident you are about the results of a survey
L> you run a survey 100 times, and get 95 similar results
L> the level is targeted before you start your study

Confidence interval
L> the range of possible values
L> that the population's result would be at the predicted confidence level
L> simply: the sample result +/- the margin of error

Statistical significance
L> determination whether result could be due to random chance or not
L> the greater significance, the less due to chance

Datei auswählen Keine ausgewählt

Testing Data

Freitag, 11. Juli 2025 10:29

Statistical Power
L> probability of getting meaningful results from a test
L> 0.6 = 60% that the result is statistically significant
L> if a test is statistically significant, the results are real
L> and not an error caused by random chance
L> usually we need a statistical power of at least 0.8 *(80%)*
L> then we can consider our result statistically significant

Hypothesis Testing
L> a way to see if a survey or experiment has meaningful results
L> testing an ad on a sample group before spreading worldwide

End

Datei auswählen Keine ausgewählt

Proxy Data

Freitag, 11. Juli 2025 10:29

When Data isn't readily available

Examples for Usage of Proxy Data

Business scenario	How proxy data can be used
A new car model was just launched a few days ago and the auto dealership can't wait until the end of the month for sales data to come in. They want sales projections now.	The analyst proxies the number of clicks to the car specifications on the dealership's website as an estimate of potential sales at the dealership.
A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years.	The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years.
The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren't publicly available yet.	The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier.

We can also use Open (public) Datasets

CSV = Credit Card Customers Dataset

JSON - Trending YouTube Videos

SQLite - U.S. Wildfire Data

Big Query - Dataset from the Google Merchandise Store

Kaggle - Community Datasets

Be cautious when using proxy data and ensure that it is well-suited for the intended purpose.

Datei auswählen Keine ausgewählt

Sample Size Calculation

Freitag, 11. Juli 2025 10:29

Confidence Level
L> probability that sample size accurately reflects the greater population
L> most industries hope for at least 90% or 95% confidence level

Margin Level Of Error
L> how close results are to what it would be with the entire population

Example - Candy Company
+ approaching candy preference of a middle school
+ population - 500 students
+ they want a confidence level of 95%
+ they want a Margin Level Of Error of 5%
+ put all the numbers into the calculator
L> The Result is ~218

Repetition

- **Confidence level**: The probability that your sample size accurately reflects the greater population.
- **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population.
- **Population**: This is the total number you hope to pull your sample from.
- **Sample**: A part of a population that is representative of the population.
- **Estimated response rate**: If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

Sample Size Calculator
https://www.surveymonkey.com/mp/sample-size-calculator/

Consider!
L> when you need a sample size of 100 individuals
L> but your response rate is 10%
L> you need to send your survey to 1,000 individuals to get the 100

Datei auswählen Keine ausgewählt

Margin Of Error

Freitag, 11. Juli 2025 10:29

Margin Of Error
L> maximum amount that sample results are expected to differ
L> from those of the actual population

Example
L> you make a survey about if people would like a 4-day-week
L> the result shows that **60% say yes**
L> while **Margin Of Error = 10%**
L> this means - the **true result is between 50% and 70%**
L> if we set the Confidence Level to 95%
L> then there's a **Chance of 95%**, that the **Result is between 50% & 70%**
L> since the numbers cross 50%
L> you can conclude that the population likes the idea of a 4-day-week

When I want to lower the Margin Of Error
L> then I need to increase the sample size
L> the same effect happens when decreasing the confidence level

Margin Of Error Calculator
L> we need the population size and sample size
L> we have to set a confidence level *(typically 90% or 95%)*
L> depending on the confidence level it will tell us the Margin

Datei auswählen Keine ausgewählt

Glossary

Freitag, 11. Juli 2025 10:29

Terms and definitions for Course 4, Module 1

Accuracy: The degree to which the data conforms to the actual entity being measured or described

Completeness: The degree to which the data contains all desired components or measures

Confidence interval: A range of values that conveys how likely a statistical estimate reflects the population

Confidence level: The probability that a sample size accurately reflects the greater population

Consistency: The degree to which data is repeatable from different points of entry or collection

Cross-field validation: A process that ensures certain conditions for multiple data fields are satisfied

Data constraints: The criteria that determine whether a piece of a data is clean and valid

Data integrity: The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

Data manipulation: The process of changing data to make it more organized and easier to read

Data range: Numerical values that fall between predefined maximum and minimum values

Data replication: The process of storing data in multiple locations

DATEDIF: A spreadsheet function that calculates the number of days, months, or years between two dates

Estimated response rate: The average number of people who typically complete a survey

Hypothesis testing: A process to determine if a survey or experiment has meaningful results

Mandatory: A data value that cannot be left blank or empty

Margin of error: The maximum amount that the sample results are expected to differ from those of the actual population

Random sampling: A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

Regular expression (RegEx): A rule that says the values in a table must match a prescribed pattern

Datei auswählen Keine ausgewählt