Lecture 6: Data Context and Quality; Pitfalls in Inferential Statistics

CMSC 25910

Spring 2022

The University of Chicago

THE UNIVERSITY OF CHICAGO

Data Challenges

Pitfalls of Data Repurposing

- Data is continuously repurposed
  - That’s one reason to keep accumulating it
  - ML Training datasets and much more
- Beware of the purpose for which you are repurposing data
  - Why was that dataset created? What was its intended purpose?

Data Quality

- Context
  - Documentation
  - Provenance
  - Assumptions
- Content
  - Errors
  - Missing Data
  - Data formats

Documenting Data

Why we need context

- Data acquisition
- Data stewards
- Data owners
- Data engineers
- Data analysts
- Data consumers

With multiple people involved in the process of transforming raw data into insights, some assumptions with downstream impact may end up buried in the complexity of organizations. Context is important!

Metadata Questions

"We have extracted 155 DGIC questions data workers often face that would be addressed with access to the right MIs. These questions illustrate the metadata landscape we consider in this paper. We have synthesized the 155 questions into 27"

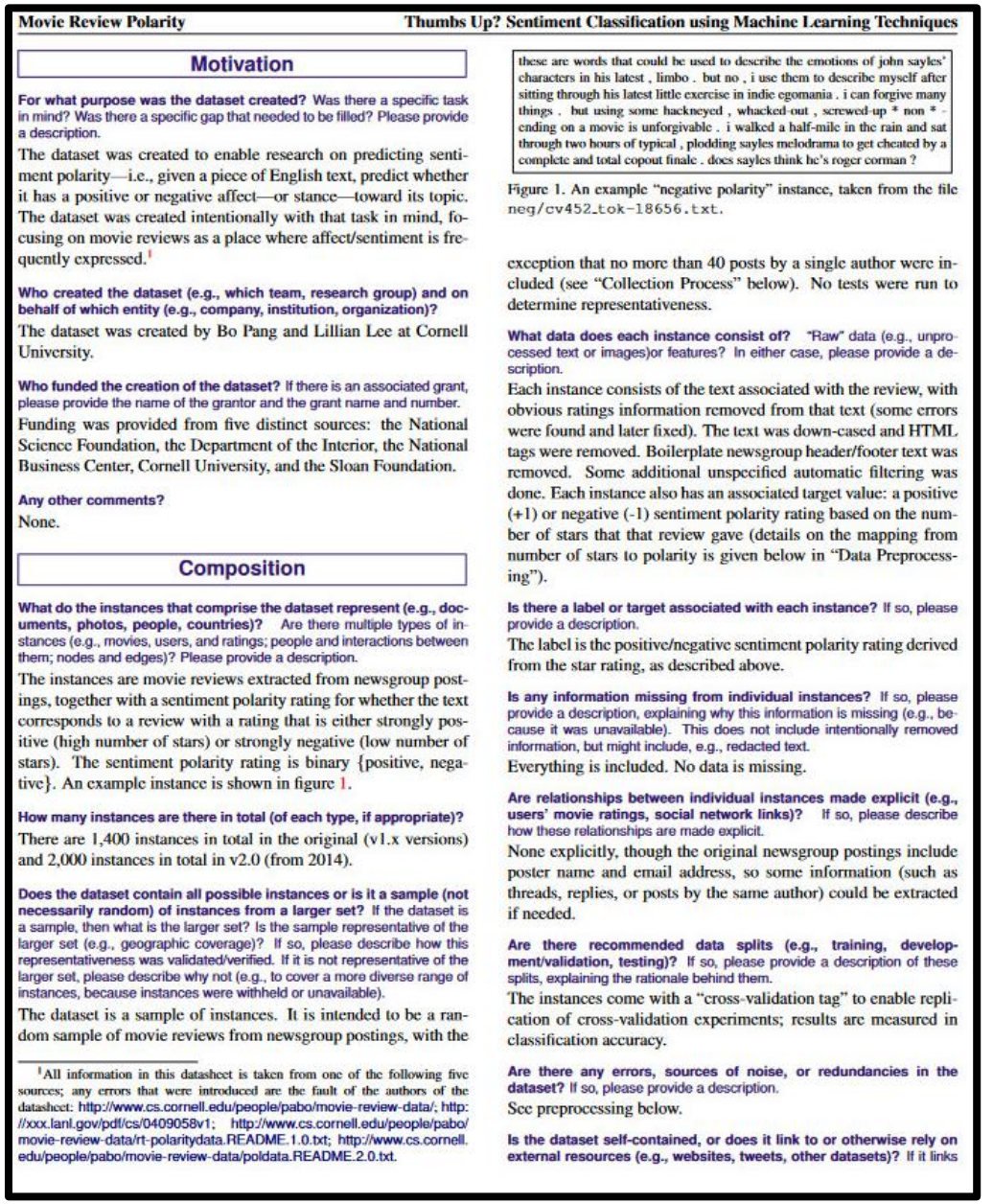
	Representatives of Common Data Questions	DGIC Category	5W1H+R Category
Q1	For what purpose was the dataset created?	G [21, 49]	Why
Q2	Are there tasks for which the dataset should not be used?	G,C [21]	Why
Q3	Who created the dataset?	G,C [21, 26]	Who
Q4	Who was involved in the data creation process?	G,C [21]	Who
Q5	How can the owner/curator/manager of the dataset be contacted?	G [21]	Who
Q6	What are the privacy and legal constraints on the accessibility of the dataset?	C [38]	Who
Q7	Is there an access control list for the dataset?	G,D [26]	Who
Q8	What is the reputation of the creator of a dataset?	G [24]	Who
Q9	What do the instances of the dataset represent?	D,G,I [21]	What
Q10	What is the size of the dataset?	D,G,I [26]	What
Q11	Are there errors in the dataset?	D,G,I [21, 24, 38]	What
Q12	Does the dataset have missing values?	D,G,I [24]	What
Q13	What is the domain of the values in this dataset?	D,G,I [30]	What
Q14	If the dataset is a sample of a larger dataset, what was the sampling strategy?	G,I [21]	How
Q15	Does the dataset contain personally identifiable information (PII)?	G,C [4, 49]	What
Q16	What is the quality of the dataset?	G [3, 4, 13, 39]	What
Q17	Was any preprocessing/cleaning/labeling of the dataset done?	G [21]	How
Q18	Was data collection randomized? Could it be biased in any way?	G [38]	How
Q19	Is there anything about dataset preprocessing/cleaning that could impact future uses?	G [21]	How
Q20	What is the dataset's release date?	D,G,I [30]	When
Q21	Is there an expiration date for this dataset?	D,G [3]	When
Q22	How often will the dataset be updated?	G,I [21]	When
Q23	When was the data last modified?	D,G,I [26]	When
Q24	How easy is it to download and explore this dataset?	D [24]	Where
Q25	What is the format of the dataset, and what type of repository is the dataset located in?	D [38]	Where
Q26	What is the provenance of this dataset?	I [54]	Relationship
Q27	What other datasets exist in this repository that are related to this dataset?	D,G,I [52]	Relationship

https://arxiv.org/pdf/2103.07532.pdf

Documentation, Context, Semantics

- Provenance/Lineage
  - How was this data recorded/obtained/acquired/produced?
- Metadata
  - Is there documentation associated with the data?
  - What do the attributes mean?
  - What units do they use?
  - (If not) find out that information before using the data. Note the assumptions you had to make

Datasheets for Datasets



Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

[1] All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/ (page 1 of 4)

Metadata Management and Catalogs

- Data Catalog:
  - A database for metadata
  - Centralizes tribal knowledge
  - Many challenges to make this work well
- Cultural and socio-technical as much as a technical problem
  - Incentives to get people to insert metadata into the catalogs
  - Documenting datasets is not in their critical path except in regulated industries or domains with strong auditors

Berkeley's Ground [28]
Microsoft Azure Data Catalog [31]
Apache Atlas [3]
Denodo platform [17]
SAP Data Intelligence platform [46]
Boomi Data platform [7]
WeWork's Marquez [50]
Lyft's Amundsen [41]
Linkedin's Datahub [37]

Data Errors

Types of Data Errors

- Outliers
  - Values that deviate from the distribution (statistical sense)
  - 2, 3, 4, 5654545, 3, 2
- Duplicates
  - Distinct records that refer to the same real-world entity
  - e.g., (first name, last name), (last name, first name)
- Rule Violations
  - Records that violate *integrity constraints*: not null, uniqueness, etc.
- Pattern Violations
  - Violate syntactic and semantic constraints: alignment, misspelling, semantic data types, etc.
    - ZIP code -> State

Types of Data Errors

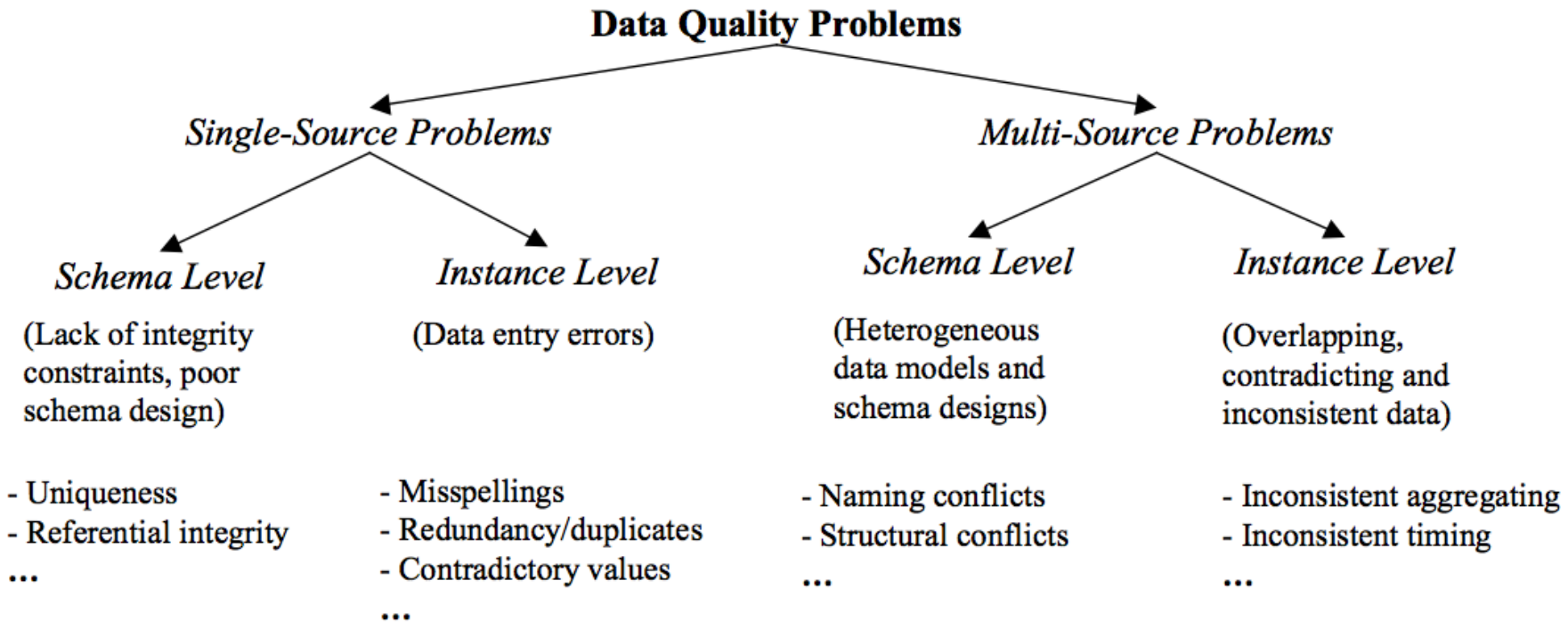


Figure 2. Classification of data quality problems in data sources

Rahm and Do: Data Cleaning: Problems and Current Approaches.
IEEE Data Engineering Bulletin (23): 3-13 (2000)

How Do Errors Affect Analysis?

Dep. Name	Num Employees
Computer Science	41
Economics	112
Statistics	26
CS	41
Physics	33
Chemistry	31

- Duplicates and outliers affect descriptive stats / aggregation
- Outliers are not always errors
  - They may indicate different measurement standards or methods, or particular distributions (e.g., long-tailed)
- Error or not? Can depend on what we are trying to achieve

The Art of Data Cleaning

“As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an ‘objective truth’ or is any interpretation necessarily biased by some subjective filter or the way that data is ‘cleaned?’”
The Promise and Peril of Big Data. Bollier, 2010, p. 13

Tooling for Data Cleaning

- OpenRefine
- Ad-hoc tools
- Most data cleaning is accomplished using ad-hoc scripts prepared by data engineers and stewards
  - This is at odds with good documentation of datasets!

Missing Data

Missing Data

- NULL values
  - Many representations: NULL, null, “NULL”, “”, 0, -1, No, “nil”, , NA, Nope…
- NULL values can result from collection or data cleaning
- How do people ‘repair’ dirty data?
  - A default strategy is to set the value = NULL
- Common approach: Drop rows and hope for the best! ☹

Missing Data

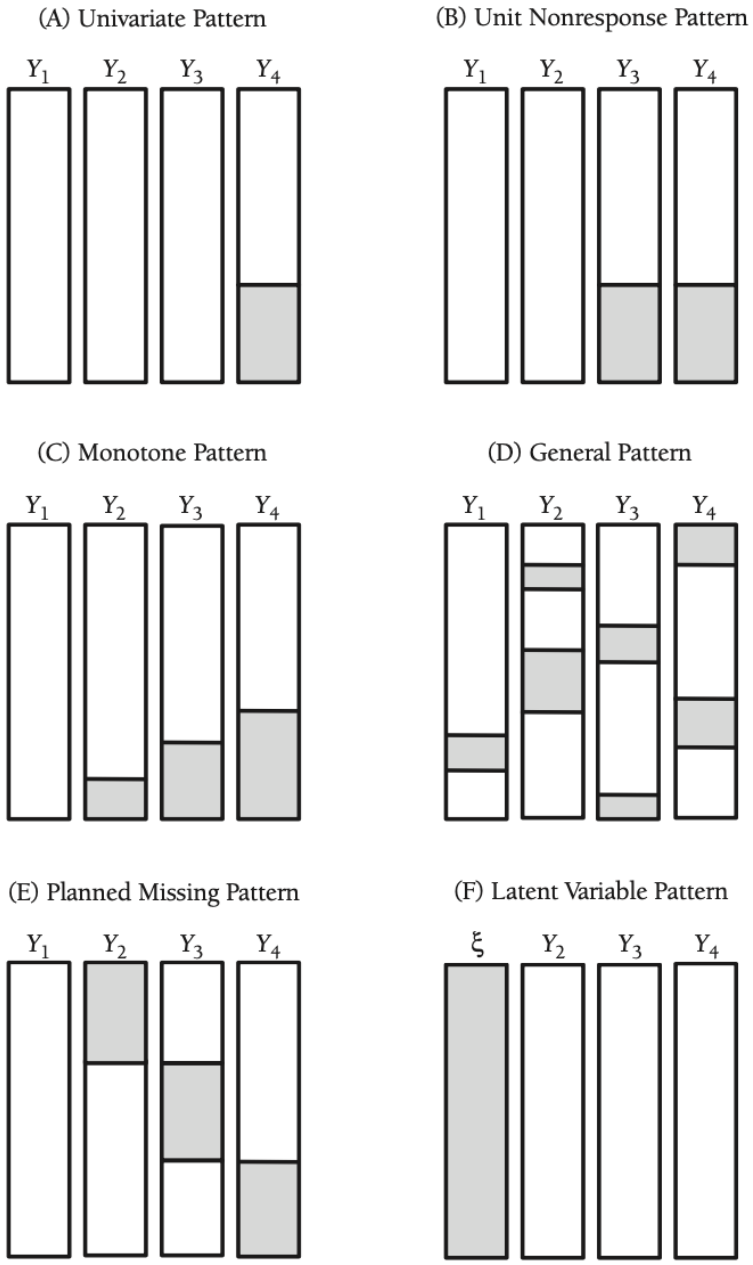
“…for most of our scientific history, we have approached missing data much like a doctor from the ancient world might use bloodletting to cure disease or amputation to stem infection (e.g, removing the infected parts of one’s data by using list-wise or pair-wise deletion). My metaphor should make you feel a bit squeamish, just as you should feel if you deal with missing data using the antediluvian and ill-advised approaches of old.“ Todd Little. Preface to Applied Missing Data Analysis, Craig Enders.

Missing Data

- Missing data patterns
  - What data is missing (e.g., which cells in a table)
- Missing data mechanisms
  - Aims to find relationships between observed variables and missing data (not necessarily explain why data is missing)

Missing Data Patterns

- Patterns: locations of missing values
- Does not explain why data is missing
- Certain patterns associated with reasons
  - e.g., attrition in multi-phase study



Example

- Consider a company’s hiring procedure to consist of two stages:
  - IQ test to determine whom to hire
  - Job performance review by a manager 6 months in

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	MAR
78	—
84	—
84	—
85	—
87	—
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 1
  - Why are those values missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	MAR
78	—
84	—
84	—
85	—
87	—
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 1
  - Why are those values missing?

- **Missing at Random (MAR)**. Probability of missing data in attribute X **depends on some other attribute Y**, but **not the values of X**.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	Complete
78	9
84	13
84	10
85	8
87	7
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

Ideal Scenario; ground truth

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MCAR
78	9	—
84	13	13
84	10	—
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	—
113	12	12
115	14	14
118	16	16
134	12	—

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 2
  - Why are those values missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MCAR
78	9	—
84	13	13
84	10	—
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	—
113	12	12
115	14	14
118	16	16
134	12	—

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 2
  - Why are those values missing?

- **Missing Completely at Random (MCAR)**. Probability of missing data in X is **unrelated** to values of X and **unrelated** to other attributes.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MNAR
78	9	9
84	13	13
84	10	10
85	8	—
87	7	—
91	7	—
92	9	9
94	9	9
94	11	11
96	7	—
99	7	—
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 3
  - Why are those values missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MNAR
78	9	9
84	13	13
84	10	10
85	8	—
87	7	—
91	7	—
92	9	9
94	9	9
94	11	11
96	7	—
99	7	—
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 3
  - Why are those values missing?

- **Missing Not at Random (MNAR)**. Probability of missing data in attribute X is related to the **values** of X.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

From “Applied Missing Data Analysis”. Craig Enders

- **Missing at Random (MAR)**. Probability of missing data in attribute X depends on some other attribute, Y, but not the values of X.
- **Missing Completely at Random (MCAR)**. Probability of missing data in X is unrelated to values of X and unrelated to other attributes.
- **Missing Not at Random (MNAR)**. Probability of missing data in attribute X is related to the values of X.

Handling Missing Data

- Drop rows with any missing data
- Fill in the blanks with the mean (or 0, or a random value)
- Maximum Likelihood Estimation
  - See https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
- Multiple Imputation
  - See https://en.wikipedia.org/wiki/Imputation_(statistics)

Handling Missing Data: Deletion

- Remove tuples that have at least one missing value
  - Assumes MCAR. Otherwise this will bias the data!
  - May reduce sample size a lot!
- Widespread because it’s very easy to implement
  - Lots of software packages include a ‘drop_null’ function
  - See Pandas documentation on ‘working with missing data’

Handling Missing Data: Imputation

- Generates a value for each missing data point
- Yields a complete dataset (unlike deletion methods)
- Can produce biased datasets (sometimes even when data is MCAR)
- Arithmetic Mean Imputation/Mean substitution:
  - Reduces the variability of the data -> attenuates standard error/deviation
- Regression Imputation: regression line fit using other (correlated) variable
  - Overestimates correlations
- Stochastic Regression Imputation: Augments regression imputation with a normally distributed residual term (i.e., adds normal noise)
  - Gives unbiased parameter estimates under MAR
- Often requires numerical data; there are advanced techniques for filling categorical data (augmentation, enrichment techniques)

Handling Missing Data: Other Techniques

- Hot-Deck Imputation: Fill missing value with non-missing value
  - Variation: cluster other observations based on variables first
  - Think about the many assumptions this method is making!
- Many other methods:
  - Similar response pattern imputation (similar to hot-deck)
  - Averaging available items
  - Last observation carried forward

Best Practices

- All previous methods assume MCAR and will bias data when data is MAR or MNAR (sometimes even when it’s MCAR)
- **Maximum Likelihood Estimation** (MLE) and **multiple imputation** can in some cases produce unbiased estimates with MCAR and MAR data, but not with MNAR

Disguised Missing Values

- Phone number:
  - (999)999-9999
- Email address:
  - nope@nope.com
- Age:
  - 666

Source	Table	Column	DMVs
UCI ML	Diabetes	Blood Pressurse	0
	adult	workclass	?
		education	Some College
U.S. FDA	Even Reports	EVENT_DT	20010101, 20030101
data.gov	Vendor Location	Ref_ID	-1
data.gov	Graduation	Regents Num	s, -
data.gov.uk	Accidents 2015	Junction Control	-1

Table 1: Sample DMVs.

From “FAHES: A Disguised Missing Value Detector.” KDD 2018

Disguised Missing Values

The Washington Post
Democracy Dies in Darkness

Lawsuit: How a quiet Kansas home wound up with 600 million IP addresses and a world of trouble

By Travis M. Andrews
August 10, 2016

https://www.washingtonpost.com/news/morning-mix/wp/2016/08/10/lawsuit-how-a-quiet-kansas-home-wound-up-with-600-million-ip-addresses-and-a-world-of-trouble/

Disguised Missing Values

“The first time Taylor realized something was amiss was when she received a call in 2011 from a small business owner who angrily blamed her for his customers’ email problems…After that initial strange call to Taylor, complaints started pouring in, often with distressing and sometimes criminal accusations aimed at the Arnolds, the Wichita Eagle reported… Officers would show up, accusing them of harboring runaway children. Of keeping girls in the house to make pornographic films. Ambulances appeared, prepared to save suicidal persons. FBI agents, federal marshals and IRS collectors have all appeared on their doorstep.”

https://www.washingtonpost.com/news/morning-mix/wp/2016/08/10/lawsuit-how-a-quiet-kansas-home-wound-up-with-600-million-ip-addresses-and-a-world-of-trouble/

Pitfalls of Statistical Hypothesis Testing

Section Outline

- **Population vs Samples**
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Population vs. Sample

- **Population**: Set of objects/events of interest for a question
  - Sometimes we don’t have access to, or don’t know, the population
- **Sample**: a subset of the population
  - We can create/collect the sample
  - We may be given the sample

Inferential Statistics

- Draw conclusions about the *population* from a *sample* of data
- Each conclusion will have an associated **sample error**
- A lot of inferential statistics is about characterizing sample error

Population – Sample Mismatch

- **Overgeneralization**: Using the sample to claim something about a *broader* population than what the sample represents
- **Bias**: We fail to obtain a *representative* sample of the population
- **Faulty generalization**: Anecdotal evidence / sample too small
- **Correlation is not causation**

Section Outline

- Population vs Samples
- **Quick recap of Statistical Inference**
- Hypothesis testing, p-values
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Goal of Statistical Inference

- To understand and quantify uncertainty of parameter estimates
- Parameter: what we are interested in learning (population)
  - Average, proportion, etc.
- *Sample, obtain point estimate, assume point estimate comes from a distribution so we can characterize its quality*

Key Terminology

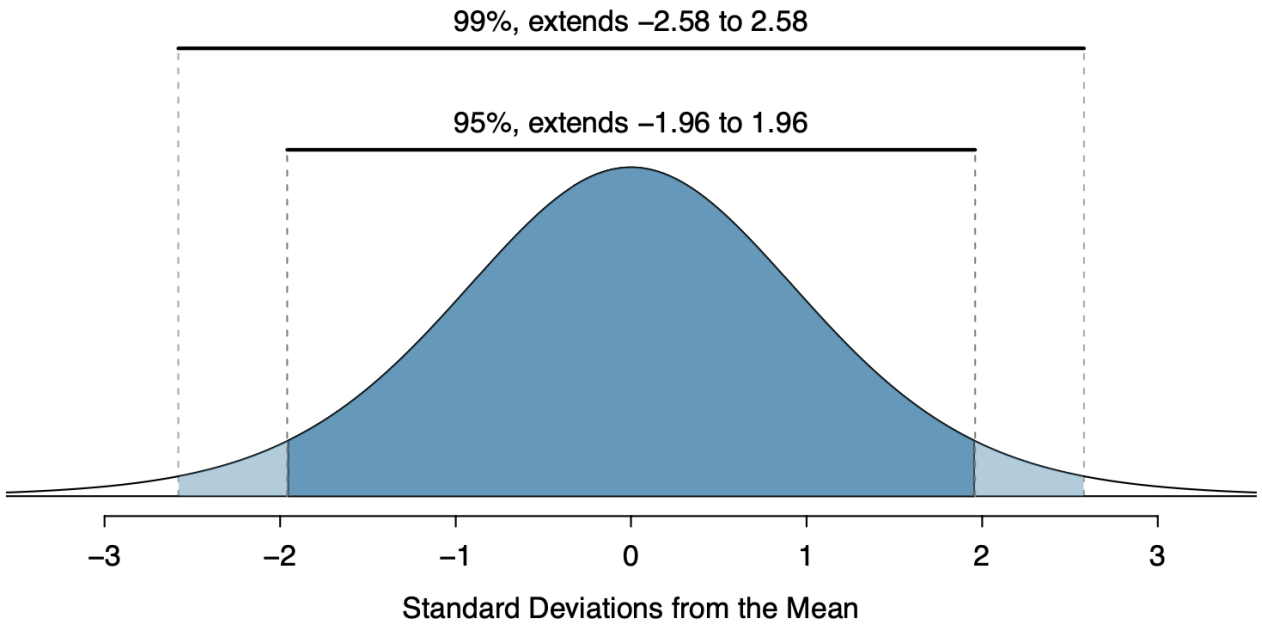
- **Population**: set of objects/events of interest for a question
- **Sample**: a subset of objects/events from the population
- **Parameter**: statistic of interest computed over the population
- **Point estimate**: statistic of interest computed over a sample of the population
- **Error**: difference between the estimate and ground truth
- **Sampling error**: How much estimate changes across samples
  - There will be some natural variation
  - Our goal is to characterize and understand this sampling error

Inferential Stats 101

- **Confidence intervals**
  - Sampling distribution
  - Central Limit Theorem (CLT)
  - Interpreting confidence intervals
- Hypothesis Testing
  - Spell out **null hypothesis (H_0)** and **alternative hypothesis (H_A)**
  - Either we reject H_0, or we fail to reject H_0

Example Confidence Interval (CI)

- Construct normal distribution with mean = point estimate
- X% confidence interval is the range that encompasses X% of the distribution
- In the case of 95%, that's 1.96 standard deviations around the mean



Hypothesis Testing Using CI

- The evidence (sample) will give us a proportion
- Build a confidence interval around that proportion
- Check if the null hypothesis fall inside or outside the interval
  - Conclude if we have enough evidence to reject it
  - If we don’t have enough evidence, all we can say is that H_0 is not implausible and we thus fail to reject H_0

Example

- Random guessing on some example quiz would lead to 33.3%
- Random sample of 50 college-educated students
  - Note the sample determines for what population are we testing Ha
- 24% of students got the response correct
- Is the deviation (24% vs. 33.3%) due to sampling error?
  - Construct confidence interval around 24%: 12% to 35%
  - 33% falls within the confidence interval, so H_0 is not implausible
- This sample doesn't provide evidence for rejecting the idea that students do better than random guessing; we cannot reject H_0.

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- **Hypothesis testing, p-values**
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

p-values

- p-value: (informally) quantifies the strength of the evidence against H_0 and in favor of H_A
- p-value: probability of observing data at least as favorable to the alternative hypothesis as the current evidence if the null hypothesis were true
- We use a summary statistic of the data to compute the p-value

Example 1/6

- Do you support moving class to the metaverse?
  - Sample: 1000 American adults.

Example 2/6

- Do you support moving class to the metaverse?
  - Sample: 1000 American adults.
- H_0 : 50% support it. **Null value:** $p_0 = 0.5$
- H_A: Significantly more/less than half support it

Example 3/6

- Do you support moving class to the metaverse?
  - Sample: 1000 American adults.
- H_0 : 50% support it. **Null value:** $p_0 = 0.5$
- H_A: Significantly more/less than half support it
- 37% support moving class to the metaverse

Example 4/6

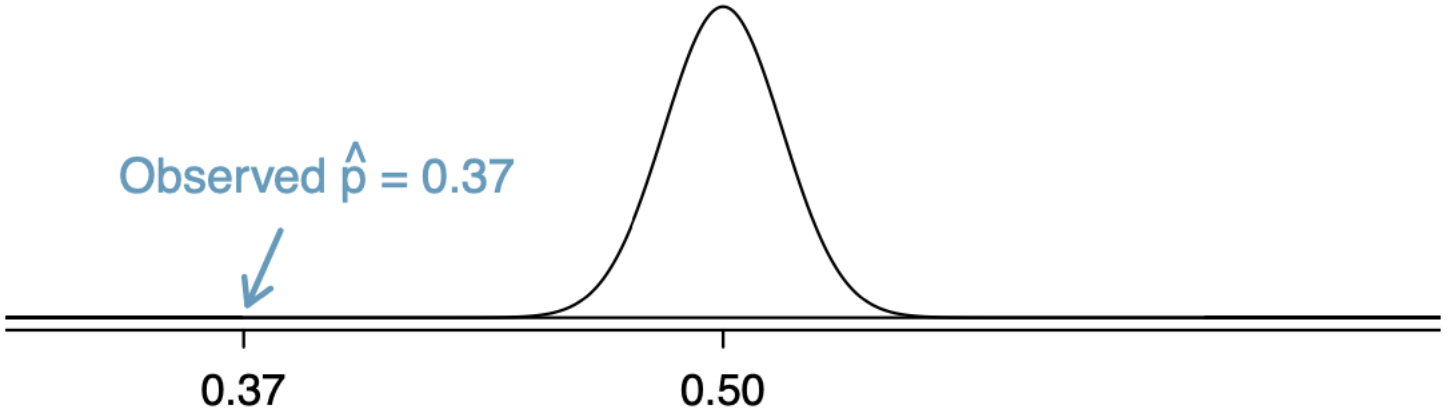
- Does 37% represent a real difference with respect to 50%? Or is it just sampling error?

Example 5/6

- Does 37% represent a real difference with respect to 50%? Or is it just sampling error?
  - What would the sampling distribution of p look like if H_0 were true?
    - If H_0 is true then population proportion is $p_0 = 0.5$
    - Is the sampling distribution normal? Independent sample
      - We check success failure condition using p_0 (we are assuming H_0 is true).
      - We compute the standard error (relies on the size of the sample)
  - If H_0 is true, distribution follows a normal with mean 0.5 and se = 0.016

Example 6/6

- Now we know the shape of the distribution (called **null distribution**) we can place the point estimate we have



- The p-value represents the probability of observing p_hat, if the null hypothesis were true.
  - If the value is smaller than the chosen significance level, we reject H_0

Constructing the p-value 1/2

- Assume the null hypothesis is true
- Check if the null hypothesis comes from a normal distribution:
  - Need to check conditions of independence and success-failure ratio
  - This is called the null distribution
- Check the sample proportion with respect to the null hypothesis distribution

Constructing the p-value 2/2

- We construct the null distribution, we then find the tail area given by our sample proportion
- **p-value represents the probability of observing the sample proportion (or more extreme) by chance if the null hypothesis were true**
- We compare p-value to alpha. If p-value < alpha we reject H_0
  - We say data provide strong evidence against H_0
  - Alpha is often 0.05 or 0.01 in user-centered CS research
- Finally, describe conclusion in the context of the data

Interpreting Hypothesis Tests

- There are two outcomes after conducting a hypothesis test:
  - We reject the null hypothesis
  - We do not reject the null hypothesis
- The following are **not** valid outcomes:
  - We accept the null hypothesis
  - We prove the null hypothesis

Other Tests

- Different statistics of interest have different sampling distributions
  - There are different tests and different ways of computing confidence intervals for those
- The principles are the same. Hypothesis testing is a framework

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- **Errors**
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Errors

- **Type 1 error / false positive**
  - Reject null hypothesis when it’s true
- **Type 2 error / false negative**
  - Fail to reject null hypothesis when alternative is true
- Consider how changing confidence interval affects these errors
  - When do we consider we have enough evidence to reject H_0
    - The threshold we choose is the **significance level**
    - How many false negatives will we have then?

Practical vs. Statistical Significance

- Statistical significance, $p < alpha$
  - We can increase statistical significance by using larger samples
- What is practical significance?
  - This depends on the application. Sometimes, it is more subjective
- Statistical Power: probability of true positive
  - P(reject H_0) when H_A is true.
  - This depends on the test you use

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- Errors
- **Multiple comparisons**
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Remember

- A *p*-value says there's *p* chance of getting the observed result *if* the null hypothesis were true
  - It **does not** mean there's a *p* chance the null hypothesis is true

Multiple Comparisons Problem

- Test multiple hypothesis on same dataset
  - False positives add to each other. Probability of false positive (false discovery) increases
- Test same hypothesis on multiple datasets
  - Same as above
- Assume H_0 is true. alpha=5%. 20 hypotheses to test
  - What's the probability of obtaining 1 discovery (rejecting H_0)?
    - 1 – P(no significant results)
    - P(no significant results) = (1 - alpha)^num_hypotheses
    - $1 – (1 – (0.05)^{20})$ -> 64% of making a 'discovery' (even if no individual hypothesis is true)

Bonferroni Method

- *If the sum of the Type I error rates for different tests is less than α, then the overall Type I error rate (FWER) for the combined tests will be at most* α*.*
- The Bonferroni method is conservative
  - See https://en.wikipedia.org/wiki/Bonferroni_correction
  - Bonferroni's conservativeness means it reduces statistical power (i.e., it reduces the probability of true positives)
- Better: use the Bonferroni-Holm adjustment
  - See https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method
- But how do you choose/define the *family* of tests?

False Discovery Rate Methods

- Alternative: control false discovery rate
  - See https://en.wikipedia.org/wiki/False_discovery_rate
  - Proportion of discoveries that are false positives
- Set the maximum allowed # of false positives (Q)
  - You choose this based on the application, like alpha
- Sort p-values from low to high, rank i=0 to i=m, for m tests
  - Compare each p-value to (i/m)Q
- Find largest p-value, p* s.t. p* < (i/m)Q
  - p* and any p s.t. p < p* are significant

False Discovery Rate Methods

- How to choose Q?
  - What's the cost of an additional experiment? And of a false negative?
    - If low and high, then you should tend to choose a higher Q
- FDR is less sensitive to the test family than Bonferroni
- Are tests really independent?
  - There are more advanced methods for when there's some dependence

P-Hacking, Data Dredging / Snooping

- This is what happens when we disregard the multiple comparisons problem
- Identify statistically significant patterns (discoveries) while increasing and understating the risk of false positives.
  - Run many statistical tests and only report those with significant results