Lecture 8: Data ethics

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

Interlude: Correlation vs. Dependence vs. Sampling bias

Correlation: Two variables are correlated if their correlation coefficient is nonzero:

$$\text{Correlation coefficient: } \rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

where $E(X)$ is the expected value of a random variable X, and Var(X) is the variance of X.

Independence: Two variables X and Y are independent when their joint probability distribution is the product of their marginal probability distributions: for all x and y,

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

Two variables are independent when the value of one gives no information about the other.

- **Independence implies uncorrelation**: If two variables are independent, then they are uncorrelated.
- **Uncorrelation does not imply independence**: But it does not work the other way around.

Correlation vs. Dependence

Dependence is causal while correlation is associational.

Classic example of dependent data

Suppose there is a trial to test whether a vaccine against covid works.

If an individual is in a group where everyone else is vaccinated (and the vaccine works) then they are less likely to get covid regardless of their vaccination status (i.e., herd immunity.) This makes it harder to test whether the vaccine works for the individual because the outcome is being obscured by the peers.

The status of the peers affect the outcome of the individual.

Sampling bias

The simplest sample for making inferences about a population is a simple random sample.

Simple random sample: A subset of individuals chosen from a larger set (called the population) with the same probability.

- Randomization ensures that this happens.
- Example: taking every 5th person out of the phone book.

There are other types of samples that can be used to make inferences.

But, if a sample is biased (e.g. convenience sample, or one that was collected out of convenience), then we will make an inference that is biased.

Data ethics: Definition

“Data ethics is a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).” - Oxford professors and philosophers Luciano Floridi and Mariarosaria Taddeo

Nice article on Medium: https://medium.com/big-data-at-berkeley/things-you-need-to-know-before-you-become-a-data-scientist-a-beginners-guide-to-data-ethics-8f9aa21af742

Goal of your training

“produce graduates who not only have deep technical expertise, but who also know how to responsibly collect and manage data, and use it to inform decisions and advance innovation to benefit the rapidly evolving world they’re graduating into”.

Statistics really begins before the data are collected

Questions you should consider if you are collecting data, or if you are using someone else's data:

- Population: What is the population of interest?
- Sample selection: Which people should I sample, and which should I not sample? (Cost constraints?)
- Missing values: How should I deal with the missing observations? What do they represent? (e.g. did not want to respond, was not available, their answer was not one of the options, etc.)
- Meta analyzes: If you are using a dataset that was created by compiling several datasets, what is the methodology for each one?
- Consent: Have you informed your respondents that their participation is voluntary? Are the incentives in place coercive?
- Biased options: Even the questions and answers you make available can be biased. Have you checked that these are inclusive and recognize different points of view?

Privacy vs. confidentiality

Data ethics checklist:

- Have we listed how this technology can be attacked or abused? [SECURITY]
- Have we tested our training data to ensure it is fair and representative? [FAIRNESS]
- Have we studied and understood possible sources of bias in our data? [FAIRNESS]
- Does our team reflect diversity of opinions, backgrounds, and kinds of thought? [FAIRNESS]
- What kind of user consent do we need to collect to use the data? [PRIVACY/TRANSPARENCY]
- Do we have a mechanism for gathering consent from users? [TRANSPARENCY]

Data ethics checklist: (continued)

- Have we explained clearly what users are consenting to? [TRANSPARENCY]
- Do we have a mechanism for redress if people are harmed by the results? [TRANSPARENCY]
- Can we shut down this software in production if it is behaving badly?
- Have we tested for fairness with respect to different user groups? [FAIRNESS]
- Have we tested for disparate error rates among different user groups? [FAIRNESS]
- Do we test and monitor for model drift to ensure our software remains fair over time? [FAIRNESS]
- Do we have a plan to protect and secure user data? [SECURITY]

(Loukides, Mason, Patil)

Global trends: Where will data ethics be relevant?

- Chief Privacy Officers can expect ethics to become an explicit part of their role.
- Technology companies will lead the way for U.S. Federal Privacy legislation.
- Sustainable ethics codes will evolve to better address the challenges of a digital world.
- Product excellence and privacy by design will become synonymous.
- Companies will drive to educate policy-makers and regulators about their technologies.

By Barbara Lawler

Given the profound shift in the digital network globally, policymakers must consider:

- What harms are they trying to protect people from?
- What rights do they want to guarantee?
- What problems are they trying to solve?
- What are the privacy outcomes they hope to achieve for their citizens?

(In)famous example in data ethics: Predictive policing

See other slides.