STAT 1291: Data Science

Lecture 13 - Professional Ethics

Sungkyu Jung

Professional ethics

Professional ethics describe the special responsibilities not to take unfair advantage of your data-analytic skills.

How to Lie with Statistics

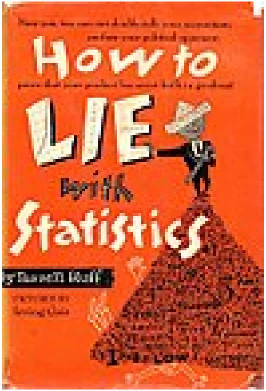
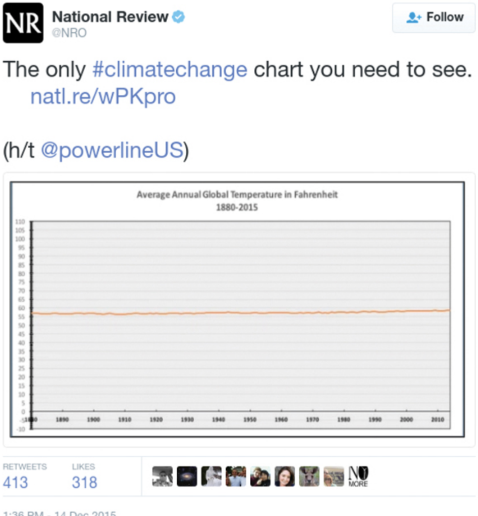


Figure 1: https://en.wikipedia.org/w/index.php?curid=4109999

- A best-seller, written by Darrell Huff in 1954.
- Used as an introductory textbook of statistics, “outlining errors when it comes to the interpretation of statistics, and how these errors may create incorrect conclusions”

Example

Some of the graphical tricks in “How to Lie with Statistics” are still in use. Consider the following example:



National Review
@NRO

Follow

The only #climatechange chart you need to see. natl.re/wPKpro

(h/t @powerlineUS)

RETWEETS 413

LIKES 318

1:36 PM - 14 Dec 2015

Is this unethical?

- There is a tacit graphical convention that the coordinate scales on which the data are plotted are relevant to an informed interpretation of the data.
- The x-axis is okay.
- The y-axis is a relevant scale for showing season-to-season variation in temperature, that is not the salient issue with respect to climate change.
- Global average temperature on the order of 5 degrees Fahrenheit can result in rising ocean levels, intensification of storms, ecological and agricultural disruption, etc.
- The National Review graphic has obscured the data by showing them on an irrelevant scale where the actual changes in temperature are practically invisible.

Evaluating ethics of a situation

- Common sense is a good starting point. Tell the truth. Don't steal. Don't harm innocent people.
- But professional ethics also require a neutral, unemotional, and informed assessment.
- A dramatic illustration:
  - A situation where the lawyers for an accused murderer found the bodies of two victims whose deaths were unknown to authorities.
  - The responsibility to confidentiality for their client precluded the lawyers from reporting the discovery.
- The lawyers' careers were destroyed by the public and political recriminations that followed, yet courts and legal scholars have confirmed that the lawyers were right to do what they did, and have even held them up as heroes for their ethical behavior.

Situation 1: CEO

- A statistical consultant for a client who wanted a model to predict commercial outcomes.
- Using public data, a multiple linear regression model was found appropriate and fit.
- The CEO asked the statistical consultant whether the coefficients in the model could be “tweaked” to reflect the specific values of his company.
- How should the consultant respond?

Situation 2: Employment discrimination

- United States Office of Federal Contract Compliance Programs (OFCCP) asks for hiring and salary data from a company.
- The company usually complies, sometimes unaware that the OFCCP applies a method to identify "discrimination" through a two-standard-deviation test, outlined in the Uniform Guidelines on Employee Selection Procedures (UGESP).
- A company that does not discriminate has some risk of being labeled as discriminating by the OFCCP method.
- By using a questionable statistical method, is the OFCCP acting unethically?

Situation 3: Data scraping

- In May 2016, the online OpenPsych Forum published a paper titled “The OkCupid data set: A very large public data set of dating site users”. The resulting data set contained 2,620 variables—including usernames, gender, and dating preferences—from 68,371 people scraped from the OkCupid dating website.
- The purpose was to provide an interesting open public data set to fellow researchers.
- The data scraping did not involve any illicit technology such as breaking passwords.
- Is this an ethical breach in doxing people by releasing personal data?

Situation 4: Reproducible spreadsheet analysis

- In 2010, Harvard economists Carmen Reinhart and Kenneth Rogoff published a report entitled Growth in a Time of Debt. These ideas influenced the thinking of policymakers during the time of the European debt crisis.
- Graduate student Thomas Herndon requested access to the data and analysis contained in the paper. After receiving the original spreadsheet from Reinhart, Herndon found several errors.

"I clicked on cell L51, and saw that they had only averaged rows 30 through 44, instead of rows 30 through 49."—Thomas Herndon

- There were coding errors, selective inclusion of data, and odd weighting of summary statistics that shaped the conclusions of the Reinhart/Rogoff paper.
- Does publishing a flawed analysis raise ethical questions?

Some principles to guide ethical action

1. Common sense: lying, cheating, and stealing are unethical.
2. Do not take advantage of your professional skills
   - As a professional, you possess skills that are not widely available.
   - Avoid using those skills in a way that is effectively lying
   - In every professional action you take, use *appropriate* methods and draw *appropriate* conclusions
3. "Appropriateness"? Draw on generally recognized professional standards
   - Use software systems that have been vetted by the community (e.g. SAS is the only tool that Food and Drug Administration approves).
   - Check that your data are what you believe them to be.
   - Don't use analytical methods that would not pass scrutiny by professional colleagues.

4. Be open and honest
    - Make sure your methods and conclusions are indeed appropriate
    - Don't overstate your confidence in results.
    - Inform possible risk and methodological limitation.
5. Have a professional responsibility to particular stakeholders
    - These vary depending on the circumstances
    - Sometimes, your main responsibility is simply to your employer or your client;
    - or, to the general public or to subjects in your study or individuals represented in your data;
    - or, to the research community or to your profession itself.

Principles to guide ethical action: A general rule

1. Do your work well by your own standards and by the standards of your profession.
2. Recognize the parties to whom you have a special professional obligation.
3. Report results and methods honestly and respect your responsibility to identify and report flaws and shortcomings in your work.

Now, back to the four situations.

Situation 1: CEO

- A statistical consultant for a client who wanted a model to predict commercial outcomes.
- Using public data, a multiple linear regression model was found appropriate and fit.
- The CEO asked the statistical consultant whether the coefficients in the model could be “tweaked” to reflect the specific values of his company.
- How should the consultant respond?

The stakeholder is the company, but manipulating the "coefficient estimates" is not accepted by your professional community.

However, statistical purity is not the issue (again, the stakeholder is the company). Your work is a tool for your client to use; they can use it as they want.

Sometimes, it's important to realize that your client's needs may not map well onto a particular statistical methodology. Often the problem that clients identify is not really the problem that needs to be solved when seen from an expert statistical perspective.

Situation 2: Employment discrimination

- United States Office of Federal Contract Compliance Programs (OFCCP) asks for hiring and salary data from a company.
- The company usually complies, sometimes unaware that the OFCCP applies a method to identify “discrimination” through a two-standard-deviation test, outlined in the Uniform Guidelines on Employee Selection Procedures (UGESP).
- A company that does not discriminate has some risk of being labeled as discriminating by the OFCCP method.
- By using a questionable statistical method, is the OFCCP acting unethically?

The methods used by OFCCP raise significant questions, since by construction they will sometimes label a company that is not discriminating as a discriminator.

Situation 3: Data scraping

- In May 2016, the online OpenPsych Forum published a paper titled "The OkCupid data set: A very large public data set of dating site users". The resulting data set contained 2,620 variables—including usernames, gender, and dating preferences—from 68,371 people scraped from the OkCupid dating website.
- The purpose was to provide an interesting open public data set to fellow researchers.
- The data scraping did not involve any illicit technology such as breaking passwords.
- Is this an ethical breach in doxing people by releasing personal data?

OkCupid provides public access to data. A researcher uses legitimate means to acquire those data. What could be wrong?

The matter of the stakeholders (OkCupid members, OkCupid itself).

- Human should not be exposed to any risk for which consent has not been explicitly given. The OkCupid members did not provide such consent.
- Since the data contain information that makes it possible to **identify individual humans**, there is a risk of the release of embarrassing information, or worse, information that jeopardizes the physical safety.
- **terms of use** that restrict how the data may be legitimately used.

Situation 4: Reproducible spreadsheet analysis

- In 2010, Harvard economists Carmen Reinhart and Kenneth Rogoff published a report entitled Growth in a Time of Debt. These ideas influenced the thinking of policymakers during the time of the European debt crisis.
- Graduate student Thomas Herndon requested access to the data and analysis contained in the paper. After receiving the original spreadsheet from Reinhart, Herndon found several errors.

"I clicked on cell L51, and saw that they had only averaged rows 30 through 44, instead of rows 30 through 49."—Thomas Herndon

- There were coding errors, selective inclusion of data, and odd weighting of summary statistics that shaped the conclusions of the Reinhart/Rogoff paper.
- Does publishing a flawed analysis raise ethical questions?

Reinhart and Rogoff recognized the parties to whom you have a special professional obligation, by providing their data and analysis to Graduate student Thomas Herndon.

It is not unethical to reach incorrect scientific conclusion.

MS Excel, the tool used by Reinhart and Rogoff, may be recongnized as standard by their community (economics), but is not in statistics.

On “Non-reproducible” spreadsheet analysis

Microsoft Excel is an unfortunate choice.

- It mixes the data with the analysis.
- It works at a low level of abstraction, so it???s difficult to program in a concise and readable way.
- Commands are customized to a particular size and organization of data, so it???s hard to apply to a new or modified data set.
- Programming and revision in Excel generally involves lots of click-and-drag copying, which is itself an error-prone operation.
- Data science professionals have an ethical obligation to use tools that are reliable, verifiable, and conducive to reproducible data analysis. (Using a scriptable language, such as R, is one way for reproducible data analysis.)

Data and disclosure

1. Can you (should you) identify this person from partial data?

- Protected health information (PHI)
- Health Insurance Portability and Accountability Act (HIPAA) dveloped procedures to ensure that *individually identifiable PHI is protected* when it is transferred, received, handled, analyzed, or shared.
- As an example, detailed geographic information (e.g., home or office location) is not allowed to be shared

2. Safe data storage?

- E.g. Equifax breach, where 143 million people's social security numbers are compromised.
- Each individual and organization needs to practice safe computing.

Data and disclosure, continued

3. How to scrape data on the Internet?

- Zillow.com has made access to their database available through an API.
- The *terms of use* for Zillow require that users of the API not replicate functionality of the Zillow website or mobile app, not retain any copies of the Zillow data, etc.

The Garden of Forking Paths

- Professor Gelman references “The Garden of Forking Paths” to describe how scientists can make false discoveries when they do not pre-specify a data analysis plan and instead choose “one analysis for the particular data they saw.”
- The “Garden of Forking Paths” refers to the near infinite number of choices facing researchers in cleaning and analyzing data, and emphasizes the need for pre-analysis planning and independent replication, an especially relevant consideration in social psychology’s recent replication crisis.

The Garden of Forking Paths, continued

- Imagine 100 parallel research efforts to investigate the effect of a drug that in reality has no effect at all.
- Any statistical test with significance level 0.05 has the 5% chance of making the type I error (detecting some effect when there is none).
- Roughly five of those 100 efforts are expected to culminate in a misleadingly “statistically significant”" ($p < 0.05$) result.
- The American Statistical Association’s ethical guidelines state, “Selecting the one *significant* result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading.”

Homework

Read Chapter 6 of MDSR.

1. Exercise 6.1
2. Exercise 6.2
3. Exercise 6.3
4. Exercise 6.10