Introduction to Data Analysis

Data Analysis for Journalism and Political Communication (Fall 2024)

Prof. Bell

Chalabi: 3 Ways to Spot a Bad Statistic

Chalabi: 3 Ways to Spot a Bad Statistic

1. Can you see uncertainty?
2. Can we look beyond the averages?
3. How was the data collected?

CAN YOU SEE THE UNCERTAINTY?

- We rarely get a complete count of everything

Can you see the uncertainty?

- We rarely get a complete count of everything
- Humans do not do well with probability

CAN YOU SEE THE UNCERTAINTY?

- We rarely get a complete count of everything
- Humans do not do well with probability

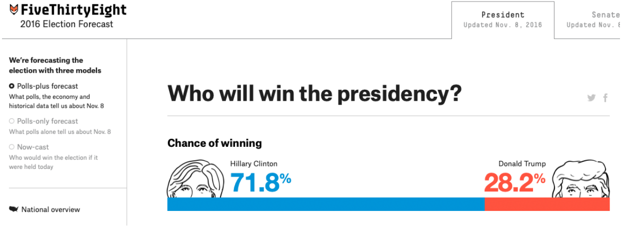
> “There are only five probabilities the average human can handle: 99 percent, one percent, 100 percent, zero, and 50-50. That’s it.”
> - Richard Thaler, Nobel Laureate in Economics

Can you see the uncertainty?

- We rarely get a complete count of everything
- Humans do not do well with probability

Can you see the uncertainty?

- We rarely get a complete count of everything
- Humans do not do well with probability



FiveThirtyEight
2016 Election Forecast

President
Updated Nov. 8, 2016

Senate
Updated Nov.

We're forecasting the election with three models

● Polls-plus forecast
What polls, the economy and historical data tell us about Nov. 8

○ Polls-only forecast
What polls alone tell us about Nov. 8

○ Now-cast
Who would win the election if it were held today

National overview

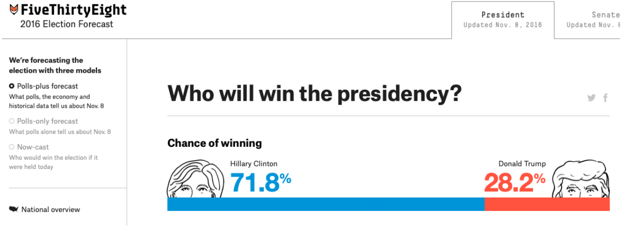
Who will win the presidency?

Chance of winning

Hillary Clinton	Donald Trump
71.8%	28.2%

Can you see the uncertainty?

- We rarely get a complete count of everything
- Humans do not do well with probability



FiveThirtyEight
2016 Election Forecast

President
Updated Nov. 8, 2016

Senate
Updated Nov. 8

We're forecasting the election with three models

● Polls-plus forecast
What polls, the economy and historical data tell us about Nov. 8

○ Polls-only forecast
What polls alone tell us about Nov. 8

○ Now-cast
Who would win the election if it were held today

National overview

Who will win the presidency?

Chance of winning

Hillary Clinton	Donald Trump
71.8%	28.2%

- We will learn about how to measure and communicate about uncertainty

Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data

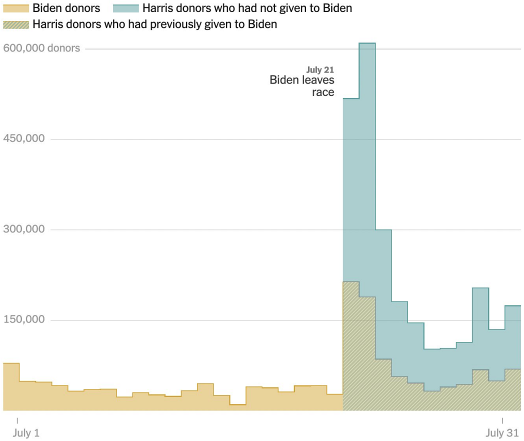
Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context

Can we look beyond the averages?

Donors both old and new gave to the newly renamed Harris campaign

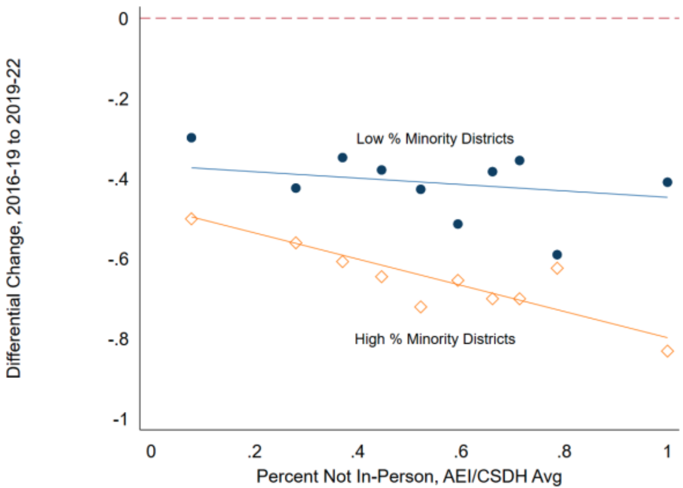
Both donors who had given to the Biden re-election campaign and new people who had not previously contributed rushed to donate to the Harris campaign.



Source: Federal Election Commission • The New York Times

Can we look beyond the averages?

Figure 5. Math achievement losses vs percent not-in-person, by percent minority



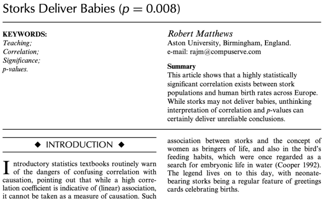
Source: Fahle, et al. (2023). “School District and Community Factors Associated With Learning Loss During the COVID-19 Pandemic.”

Can we look beyond the averages?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context
- We will talk about using data visualization to communicate about data, as well as researcher choices and biases

CAN WE LOOK BEYOND THE AVERAGES?

- There is always a trade-off between accuracy and simplicity when working with data
- We aggregate data to make it easier to comprehend, but we may also lose important context
- We will talk about using data visualization to communicate about data, as well as researcher choices and biases
- We will also talk about the importance of theory in understanding data, especially correlation vs. causation



Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary

This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

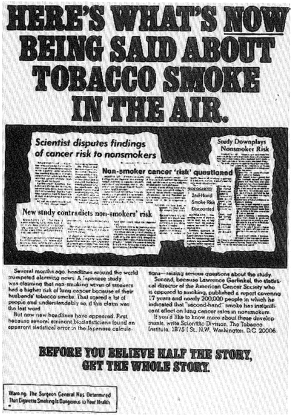
Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

How was the data collected?

- Data is not objective – it is generated by humans

How was the data collected?

- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



HERE'S WHAT'S NOW BEING SAID ABOUT TOBACCO SMOKE IN THE AIR.

Scientist disputes findings of cancer risk to nonsmokers

Study Downplays Nonsmoker Risk

Non-smoker cancer 'risk' questioned

2nd-Hand Smoke Risk Discounted

New study contradicts non-smokers' risk

Several months ago, headlines around the world trumpeted alarming news. A Japanese study was claiming that non-smoking wives of smokers had a higher risk of lung cancer because of their husbands' tobacco smoke. That scared a lot of people and understandably so, if this claim was the last word.

But now new headlines have appeared. First, because several eminent biostatisticians found an apparent statistical error in the Japanese calculations—raising serious questions about the study.

Second, because Lawrence Garfinkel, the statistical director of the American Cancer Society who is opposed to smoking, published a report covering 17 years and nearly 200,000 people in which he indicated that "second-hand" smoke has insignificant effect on lung cancer rates in nonsmokers.

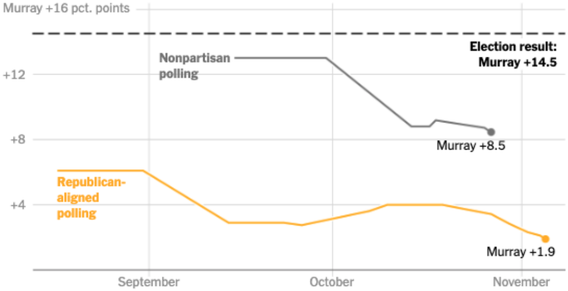
If you'd like to know more about these developments, write Scientific Division, The Tobacco Institute, 1875 I St., N.W., Washington, D.C. 20006.

BEFORE YOU BELIEVE HALF THE STORY, GET THE WHOLE STORY.

Warning: The Surgeon General Has Determined That Cigarette Smoking Is Dangerous to Your Health.

How was the data collected?

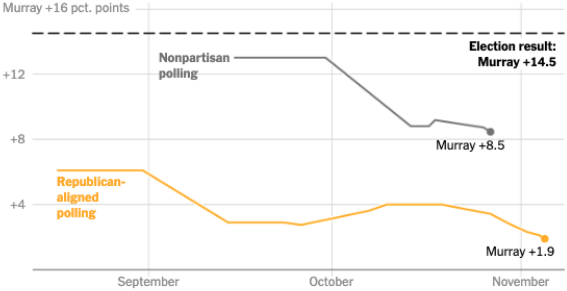
- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



Source: New York Times analysis of Washington Senate race polls aggregated by FiveThirtyEight · Notes: Trends are calculated with a 14-day average. Polling groups considered Republican-aligned include those identified by The New York Times and FiveThirtyEight. Polling groups considered nonpartisan are those not known to be aligned with or funded by a political party. · By Jason Kao

How was the data collected?

- Data is not objective – it is generated by humans
- Some data is produced by unscrupulous actors



Source: New York Times analysis of Washington Senate race polls aggregated by FiveThirtyEight • Notes: Trends are calculated with a 14-day average. Polling groups considered Republican-aligned include those identified by The New York Times and FiveThirtyEight. Polling groups considered nonpartisan are those not known to be aligned with or funded by a political party. • By Jason Kao

- But most of the time, poor analysis is not nefarious – humans are imperfect

How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data

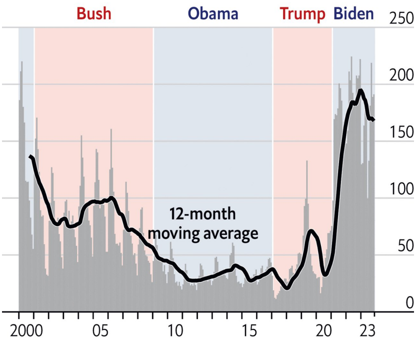
How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data
- We will spend a lot of time thinking about the **data generating process** and how it can bias our results

How was the data collected?

Illegal immigration 1

Monthly encounters at the south-west land border*, '000



*Only encounters between ports of entry. Since March 2020 monthly totals include apprehensions & expulsions. Prior totals include apprehensions only

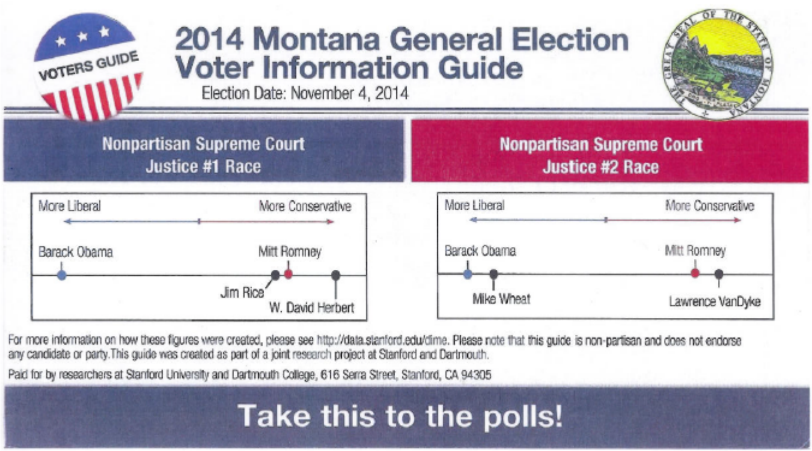
Source: US Customs and Border Protection

The Economist

How was the data collected?

- Garbage in = garbage out: no amount of statistical wizardry can compensate for bad data
- We will spend a lot of time thinking about the **data generating process** and how it can bias our results
- We will also discuss our ethical responsibilities around data

How was the data collected?



VOTERS GUIDE

2014 Montana General Election Voter Information Guide

Election Date: November 4, 2014

THE GREAT SEAL OF THE STATE OF MONTANA

Nonpartisan Supreme Court Justice #1 Race

More Liberal
More Conservative
Barack Obama
Mitt Romney
Jim Rice
W. David Herbert

Nonpartisan Supreme Court Justice #2 Race

More Liberal
More Conservative
Barack Obama
Mitt Romney
Mike Wheat
Lawrence VanDyke

For more information on how these figures were created, please see http://data.stanford.edu/dime. Please note that this guide is non-partisan and does not endorse any candidate or party.This guide was created as part of a joint research project at Stanford and Dartmouth.

Paid for by researchers at Stanford University and Dartmouth College, 616 Serra Street, Stanford, CA 94305

Take this to the polls!

Group Discussion

Introduce yourself to your neighbor(s) and take a few minutes to review these additional graphs from Mona Chalabi. Do any of these stand out to you as being good (or bad) examples of our three questions for spotting a bad statistic?

1. Can you see uncertainty?
2. Can we look beyond the averages?
3. How was the data collected?

On your notecard, please write:

1. Preferred name
2. Preferred pronouns
3. Year in school and major
4. Your background in coding and/or statistics
5. One thing you hope to get out of this class