Causal Inference

Divij Sinha 2025-04-17

Inference vs prediction

- Prediction -
  - Be correct about your outcome as often as possible
  - Be as close to the output as possible
- Inference
  - **Why** is a particular output what it is?
  - What is the **effect** of a particular input?

Inference vs prediction

- Prediction -
  - “What is the average sale price of a house in South Loop”
  - “Are sale prices in South Loop higher than Englewood”
- Inference
  - “What is the effect of square footage on sale price”
  - “Why are sale prices in South Loop higher than Englewood”

“What is the effect of square footage on sale price”

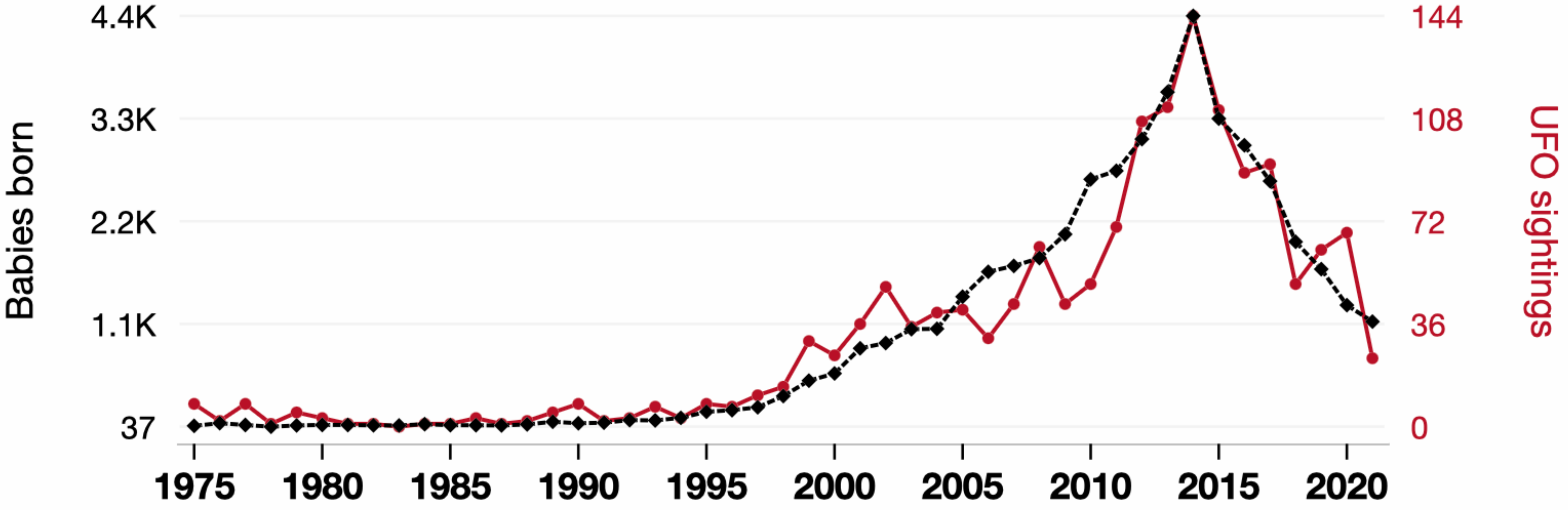
- Run a linear regression on it?
  - Why? Why not?

  - “What is the effect of ice cream sales on temperature”

Popularity of the first name Annabelle

correlates with

UFO sightings in Maryland



Babies of all sexes born in the US named Annabelle · Source: US Social Security Administration

UFO sightings reported in Maryland · Source: National UFO Reporting Center

1975-2021, r=0.962, r²=0.925, p<0.01 · tylervigen.com/spurious/correlation/3029

Plausible Explanation

Show GenAI's made-up explanation

As the number of Annabelles grew, so did the collective power of their positive energy, inadvertently attracting curious extraterrestrial beings to the skies above Maryland. It seems that the universe just couldn't resist the charm and magnetism that this particular name exuded, leading to a surge in close encounters of the Annabelle kind!

The Belle and the Beams: A Statistical Analysis of Annabelle's Popularity and UFO Sightings in Maryland

---

Caleb Horton, Anthony Travis, Gavin P Truman

Abstract

The present study examines the potential relationship between the popularity of the first name Annabelle and the occurrence of UFO sightings in the state of Maryland. By analyzing data from the US Social Security Administration and the National UFO Reporting Center

1. Introduction

Annabelle may be a name synonymous with beauty and grace, but could there be more to this moniker than meets the eye? The present study delves into the intriguing possibility of a connection between the popularity of the first name Annabelle and the

Correlation does NOT **mean causation**

Linear regressions test for correlation!

Causal Inference

The process of estimating the effect of a variable on an outcome, isolating that effect from other factors.

What if the topic is closer?

Absurdity masks the difficulty

- Number of electrical engineers vs Energy production
- Decrease in Fertility Rates vs Increase in education
- CO_2 emissions vs increase in global temperatures?

What is the effect of ice cream sales on temperature?

We know there is none, how do we prove it?

What is the actual question?

- Can more people buying more ice-creams lead to a change in temperature?

- What do you need to answer this question?

Ideal Situation

- Clone the world into 2
  - In one version you sell ice-cream
  - In another version you don’t sell ice-cream

- The idea at play here is the **counterfactual**

Counterfactual

YOU CANNOT HAVE BOTH!

- **The outcome that would have occurred under an alternative scenario — one that didn’t happen**
- We cannot clone the world unfortunately

Fundamental problem of causal inference

- **We cannot observe both what happens and what does not happen at the same time!**
  - We can either sell ice-cream, or not sell ice-cream
  - We cannot both sell AND not sell ice-cream

How to approach this problem?

In the absence of a counterfactual

- Let us say the unit of study is a city. We cannot clone a city but we can have two cities
  - Pick say 2 cities - Chicago and ??
    - Sell ice cream in Chicago
    - Do not sell ice cream in ??

- What if you sell ice-cream in summer in Chicago and do not sell ice-cream in winter in NY?
  - Positive correlation!
- What if you sell ice-cream in winter in Chicago and do not sell ice-cream in summer in NY?
  - Negative correlation!

- *Correct answer is **0** correlation*

- Sell ice-cream all year in Chicago, and do not sell ice-cream all year in NY
  - What if Chicago just hates ice-cream and no one buys it?

- Sell ice-cream all year in Chicago, and do not sell ice-cream all year in *Dallas*?
  - Dallas is on average hotter than Chicago!

What we really need are two cities that are exactly the same in all behaviors

This would mean that our understanding is only limited to these two cities!!

Caveat

Solution

- What we do instead is get lots of different cities, some that like ice-cream some that don't, some big, some small etc. etc.

- We RANDOMLY assign some of them as getting ice-cream, and some of them as not getting ice-cream

- We then measure the temperature across them!

Average treatment effect

- $ATE = E[Y(\text{ice-cream}) - Y(\text{no ice-cream})]$

- Average Difference between temps across cities that get ice-cream, and cities that do not get ice-creams

City	Randomized Ice Cream Sold?	Avg Temp (°F)
Chicago	Yes	70
New York	No	70
Dallas	Yes	90
Boston	No	90

Where are regressions?

- $temp_i = \beta_0 + icecream_i\ \beta_1 + season_i\ \beta_2 + \epsilon$
  - We're trying to isolate β_1 as the causal effect of ice cream on temperature
  - IceCream is the treatment (binary: yes/no)
  - Season is a confounder

  - Is Season important? What about population, racial make-up etc?

Causation and models

- As we said earlier, correlation does not mean causation
- HOWEVER, correlation is important in the process to establish causation
- We need to consider how the process is occurring, what sources of variation might exist, and how we can account for them
- Fundamentally we run very similar linear regression models as before, however the big difference is in the process not the math

Caveat!

- These results are consistent and applicable only WITHIN the set of units you pick, or the general population from where they belong
    - If all cities you pick are from the USA, you can say that ice-cream does not effect temperature *<u>in the USA</u>*!
    - If you only pick cities with over a million people, you can say that ice-cream does not effect temperature *<u>in large cities</u>*!
    - If you pick big and small cities from around the world, being generally representative of the population you care about, then you can speak more generally!

How to prove X

Randomized Controlled Trials - *"gold standard"*

- Create an experiment where that is the only thing that varies across all units
  - This ensures treatment is uncorrelated with all confounders — both observed and unobserved
- Pick a set of diverse units representative of the population you care about
- Then divide into two halves randomly
- In one half do X, in the other do not do X
- Measure the outcome variable in both halves, and look at their difference

This is not easy!!

Selection Bias

- This is what happens when assignment is not random
  - That is, your assignment is correlated with some other variable
    - Eg. - You sell all ice-creams in summer, and none in winter
- Randomization avoids **selection bias**

Natural Experiments
When we cannot run our own experiments

- A natural experiment is when there are real-world changes that create random selection without being directly correlated to the outcome
  - Examples
    - Draft Lottery
    - Policy Shocks - minimum wage increase
    - Cutoffs - Medicare age
    - Lotteries - school, visa etc.

Extensions

- Matching
  - You have some set of units where “treatment” is observed
  - Find very similar units where treatment is not observed
    - Relies on rich data on other variables to find good matches
    - If there is variation on something you cannot observe, it fails

Differences in Differences

- Useful in cases when interested in outcomes over time

	Year Before	**Year After**
City A	10,000 trees	18,000 trees
City B	7,000 trees	11,000 trees

- Still relies on units being similar, as the assumption is that the trend will be the same across years in the absence of treatment!
  - ***"parallel trends assumption"***

Regression Discontinuity Design

- Useful in cases of arbitrary cutoffs
- You want to see the effects of Medicare on people going to hospitals.
  - Since you cannot deny people Medicare, cannot run your own experiment
  - Since older people are more likely to go to the doctor, there is no random assignment as age and Medicare access are closely related
- You look at people in the age group 64-66, where 64-65 do not have access, and 65-66 do have access
- The smaller the window the better, but small window -> smaller samples