ACC 547: DATA ANALYTICS IN ACCOUNTING

CAUSAL MODELING AND STATISTICAL SIGNIFICANCE

Culverhouse College of Business

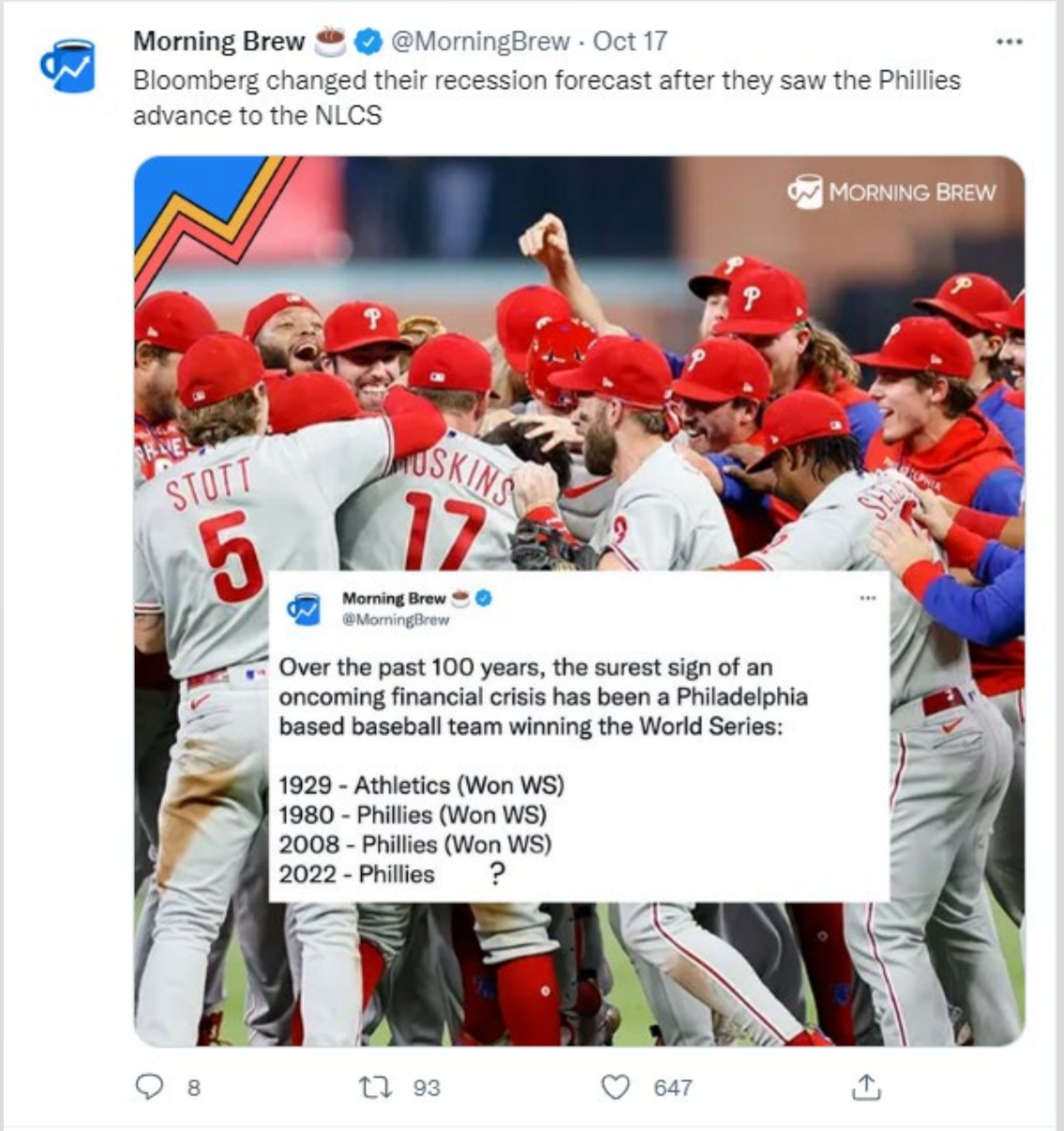
CORRELATION OR CAUSATION?

You have probably heard some variation of the phrase “correlation does not imply causation.”

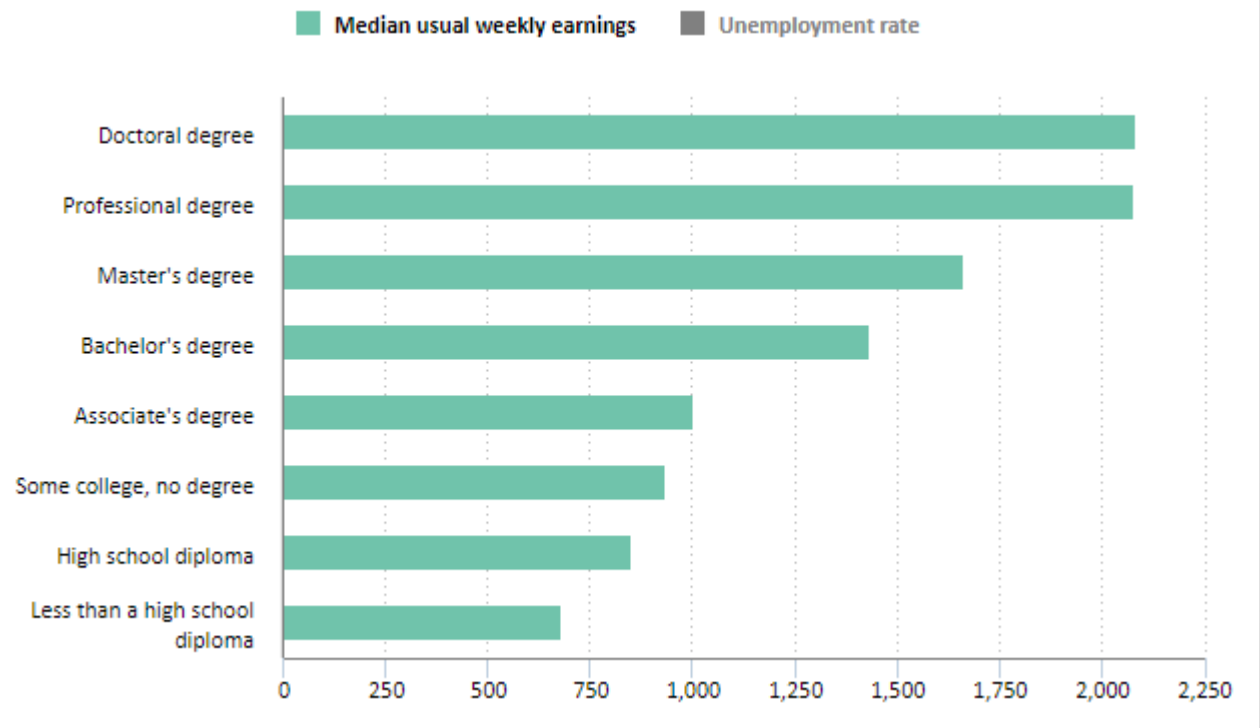
CORRELATION OR CAUSATION?

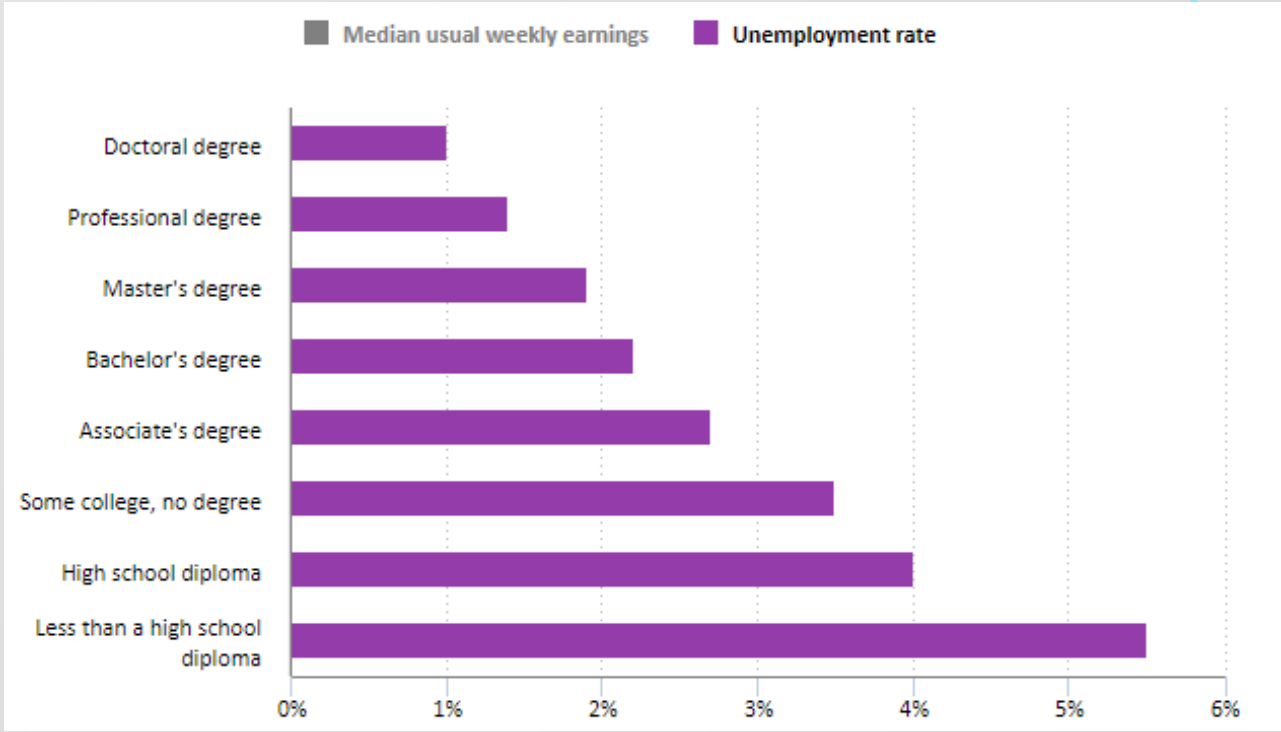
Some correlations are due to "chance" and don't reflect a meaningful causal connection.

Examples: Spurious Correlations



CORRELATION OR CAUSATION?



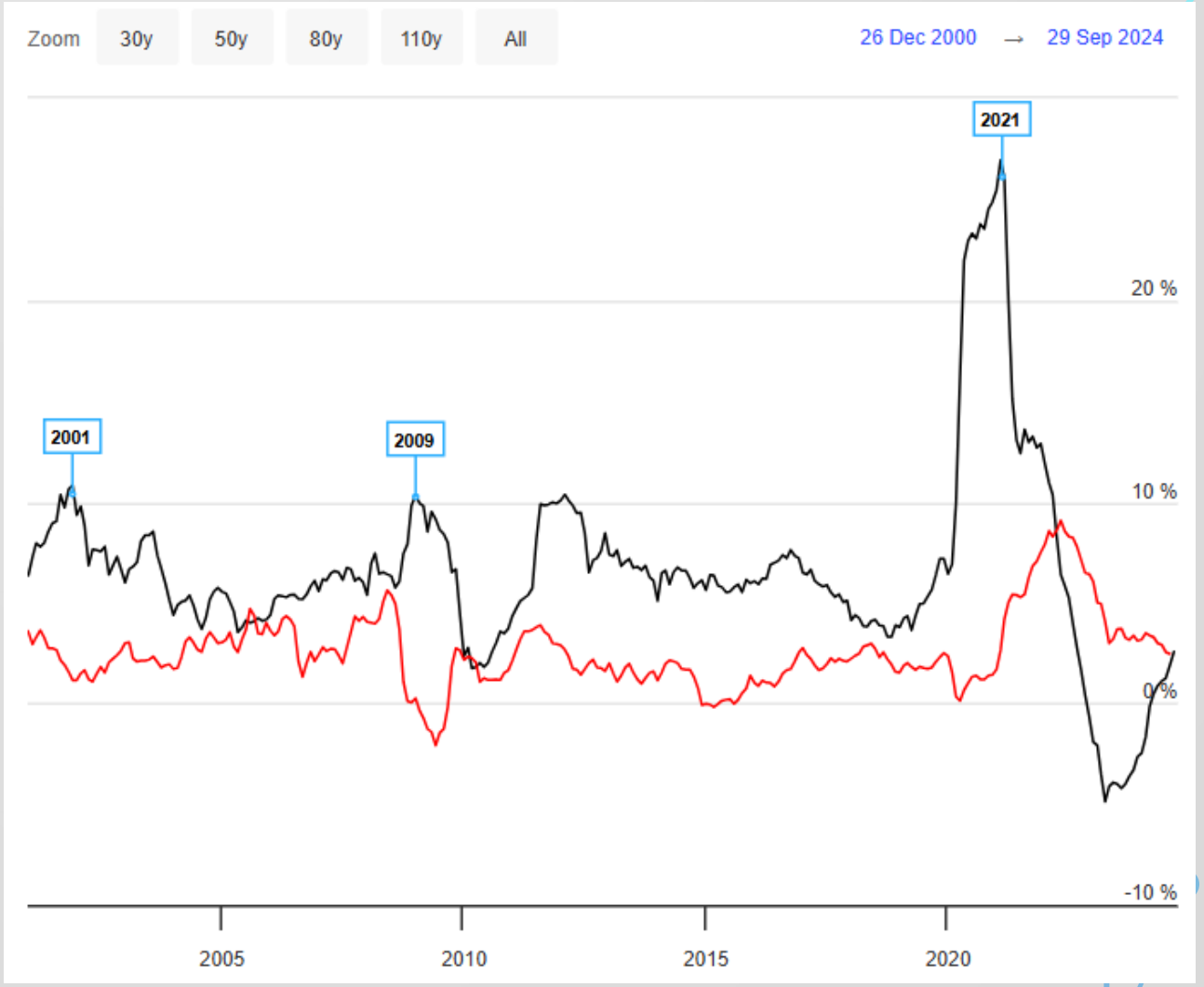


Some correlations are “confounded” or due to other factors.

Source: Education pays, 2022 : Career Outlook: U.S. Bureau of Labor Statistics (bls.gov)

CORRELATION OR CAUSATION?

"Non-correlation does not imply no causation!" Some things that appear uncorrelated may be causally related…



Legend: **M2 Growth Rate** and the **Inflation Rate**
Source: M2 Money Supply Growth vs. Inflation - 154 Year Chart | Longtermtrends

CAUSAL MODELING QUESTIONS

- Consider questions that we may want to know the answer to:
  - Is a new drug effective in improving health outcomes?
  - Does schooling improve future earnings?
  - Do masks prevent transmission of airborne diseases?
  - Did the Sarbanes-Oxley Act improve financial reporting?
  - Do investors respond to material weakness disclosures?
  - Is Sam the bartender skimming?
- If you wanted to know the answer to these questions what information (or data) would you want?

OVERVIEW OF CAUSAL MODELING: STUDY TYPES

Experimental studies: Settings where the researcher (or someone/something else) intervenes to assign conditions or treatments.

Observational studies: Settings where the researcher "observes" but does not intervene to assign conditions or treatments.

EXPERIMENTAL IDEAL

When investigating a causal effect, it is helpful to think of a hypothetical experiment. That is, we might use random assignment to investigate the effects of a treatment. Random assignment is a way of "holding all else equal." Experiments are the "gold standard" in causal inference.

EXPERIMENTAL IDEAL

Often, we can't run experiments and are limited to drawing inferences from observational settings. To hold all else equal in observational settings, we frequently rely on statistical techniques in an attempt to isolate causal connections. The technique we focus on in this course is regression, but there are several other approaches (e.g., instrumental variables). *Note that all approaches require* ***untestable*** *assumptions to draw causal inferences.*

We should always use caution in drawing causal inferences from observational data, particularly when there is good reason to believe our model/data is incomplete or misspecified.

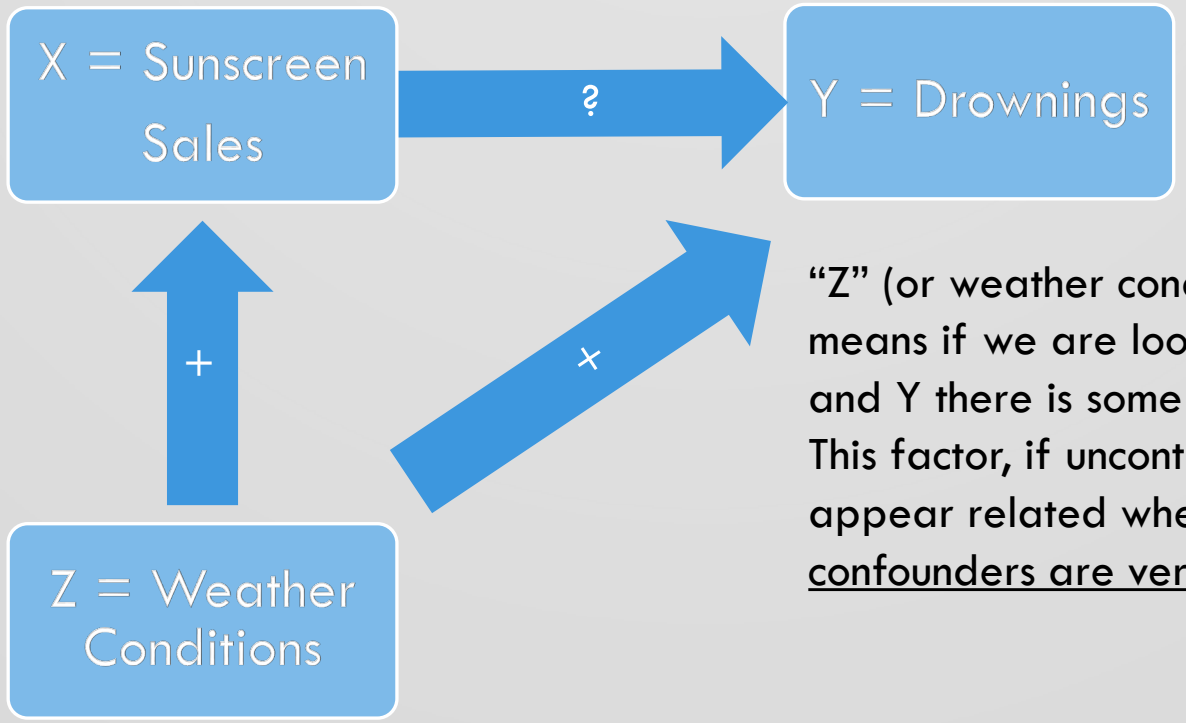
CAUSAL MODEL – SUNSCREEN AND DROWNINGS

- Investigation of the relation between sunscreen sales and drownings:
  - $drownings = \beta_0 + \beta_1 sunscreen\ sales + \varepsilon$
- If you want determine if sunscreen sales <u>cause</u> drownings you need to "hold all else equal" (also referred to as "ceteris paribus").
- By adding additional variables to a regression model, we can adjust for the linear effects of those variables. To the extent we have the "correct" model, we can approximate the effect of sunscreen sales "holding all else equal."

$$drownings = \beta_0 + \beta_1 sunscreen\ sales + \beta_k controls + \varepsilon$$

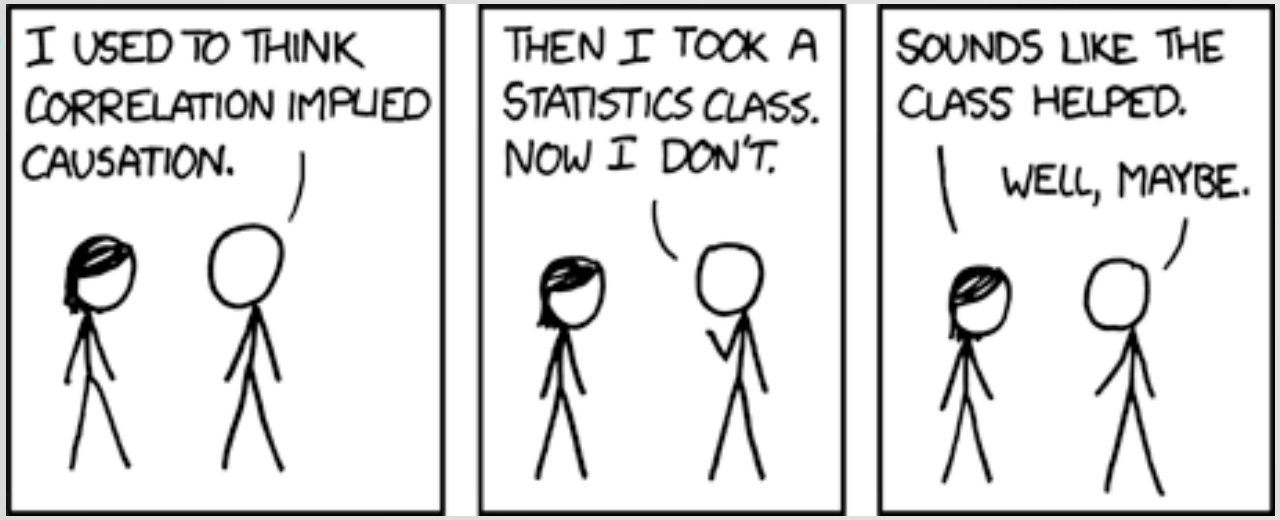
THE CAUSAL MODEL – SUNSCREEN AND DROWNINGS

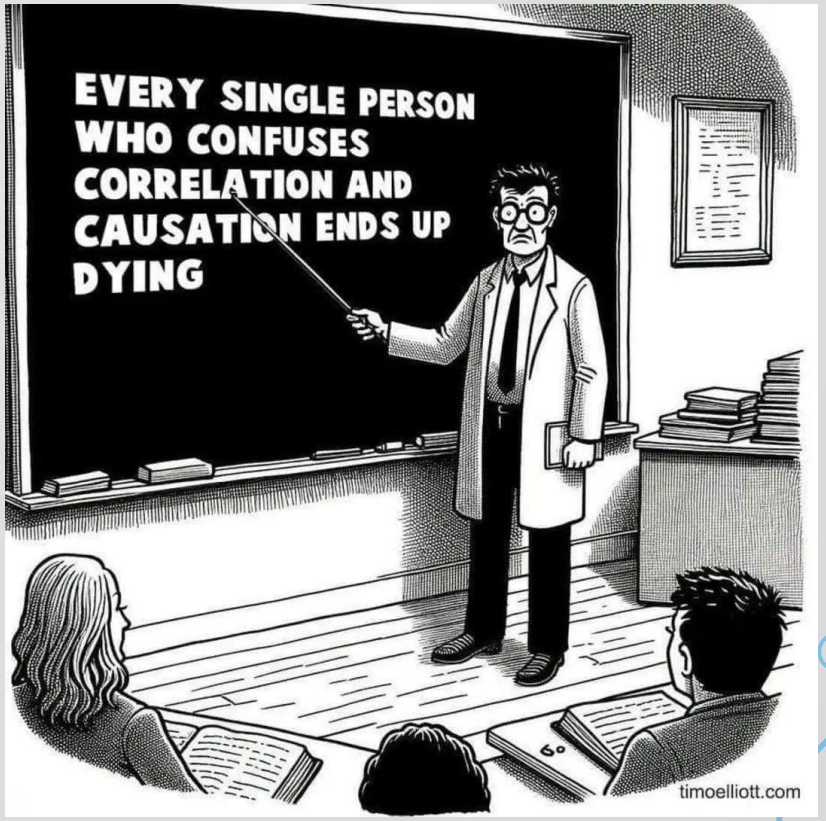
- $drownings = \beta_0 + \beta_1 sunscreen\_sales + \varepsilon$



"Z" (or weather conditions), is a "confounder." That means if we are looking at the relations between X and Y there is some other factor that determines both. This factor, if uncontrolled, will cause X and Y to appear related when they are not. Uncontrolled confounders are very bad for causal inference.

DISCUSSION EXAMPLES





Nate Silver ✓
@NateSilver538

To take a relatively non-spicy example, there's a correlation between how harshly the media covered Enron and how much legal trouble they got in. But the media very much wasn't the cause of Enron's problems. There's an objective reality we're describing.

10:33 PM · Jul 5, 2024 · **159.6K** Views

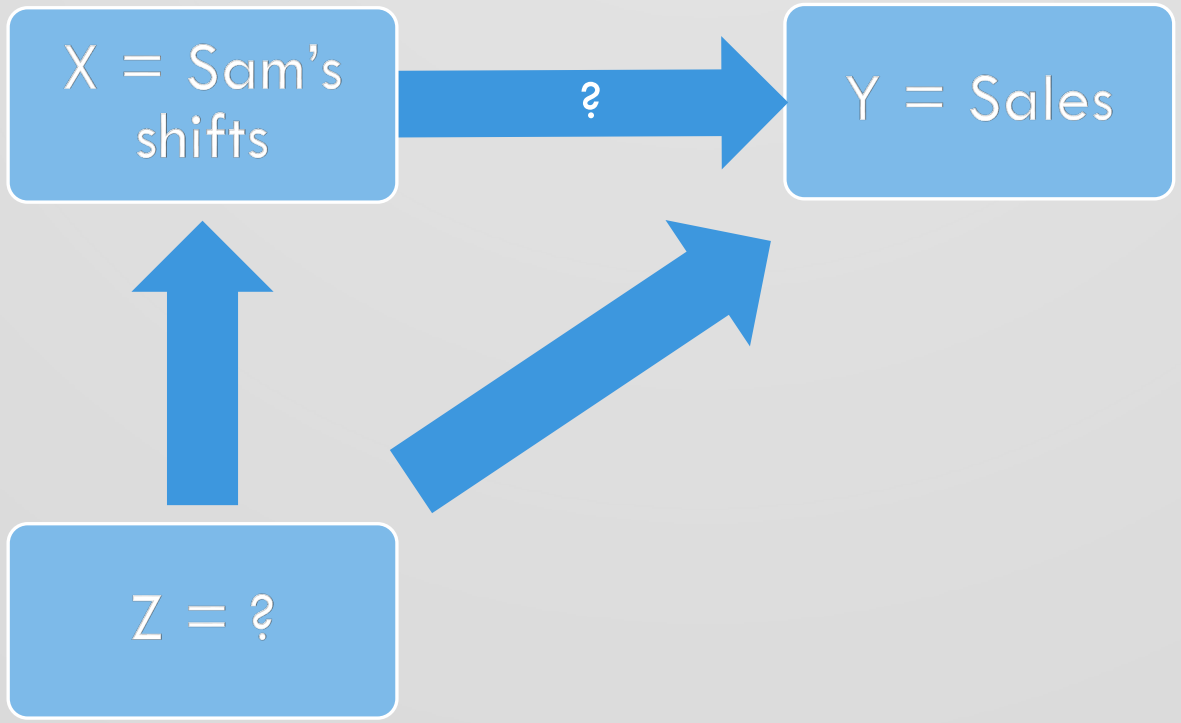
Sources: [Correlation vs Causation Cartoon – Innovation Evangelism]
[xkcd: Correlation]

EXERCISE – SAM THE BARTENDER

You are working as a forensic data scientist at an accounting firm. One of your clients, the owner of a local restaurant, is concerned that Sam, a bartender, has been "skimming" from drink sales to customers. The basis for the owner's concern is that the accounting system clearly shows that sales and cash collections are consistently lower when Sam is working at the bar. The owner is considering firing Sam but asks you to look into it first.

You are not convinced by the validity of the owner's conclusions related to Sam's culpability. Come up with a plausible alternative explanation that might explain the seemingly suspicious pattern. Discuss what data and analyses you might use to confirm or disconfirm this explanation.

THE CAUSAL MODEL – SAM THE BARTENDER



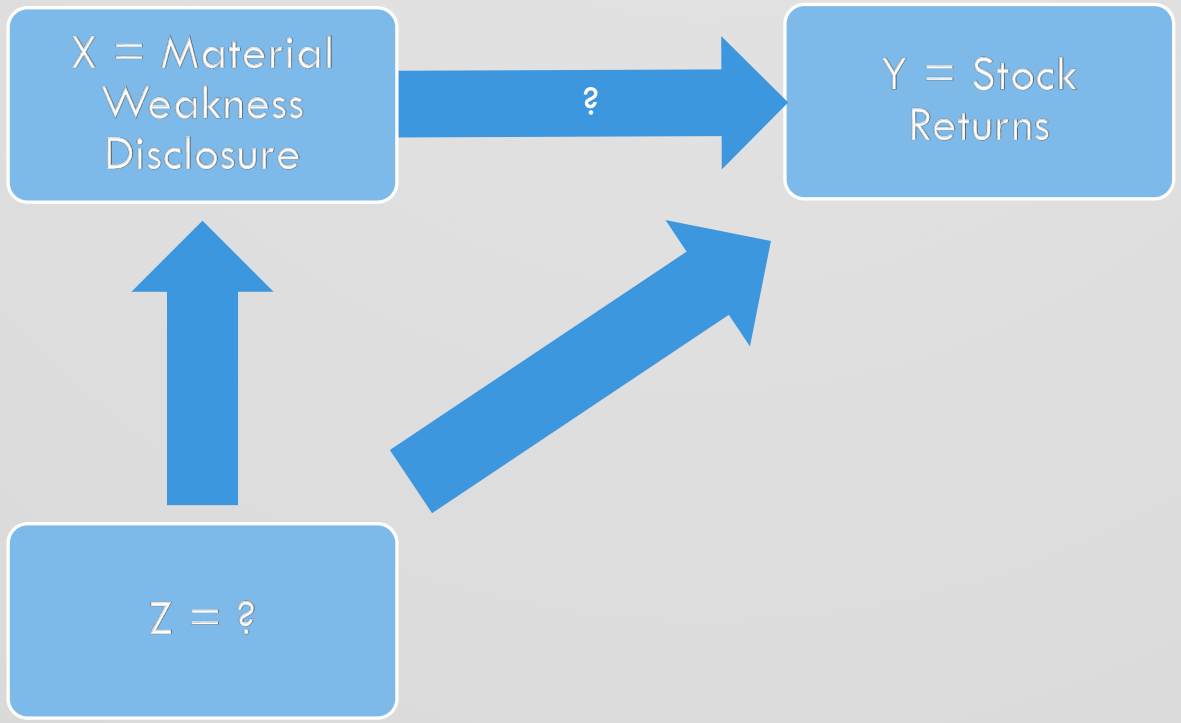
THE CAUSAL MODEL – MATERIAL WEAKNESS DISCLOSURES

Suppose you want to know if material weakness disclosures are informative to investors. You put together an analysis that investigates the market response around the disclosure of an audit report that notes a material weakness.

Consider the following example:

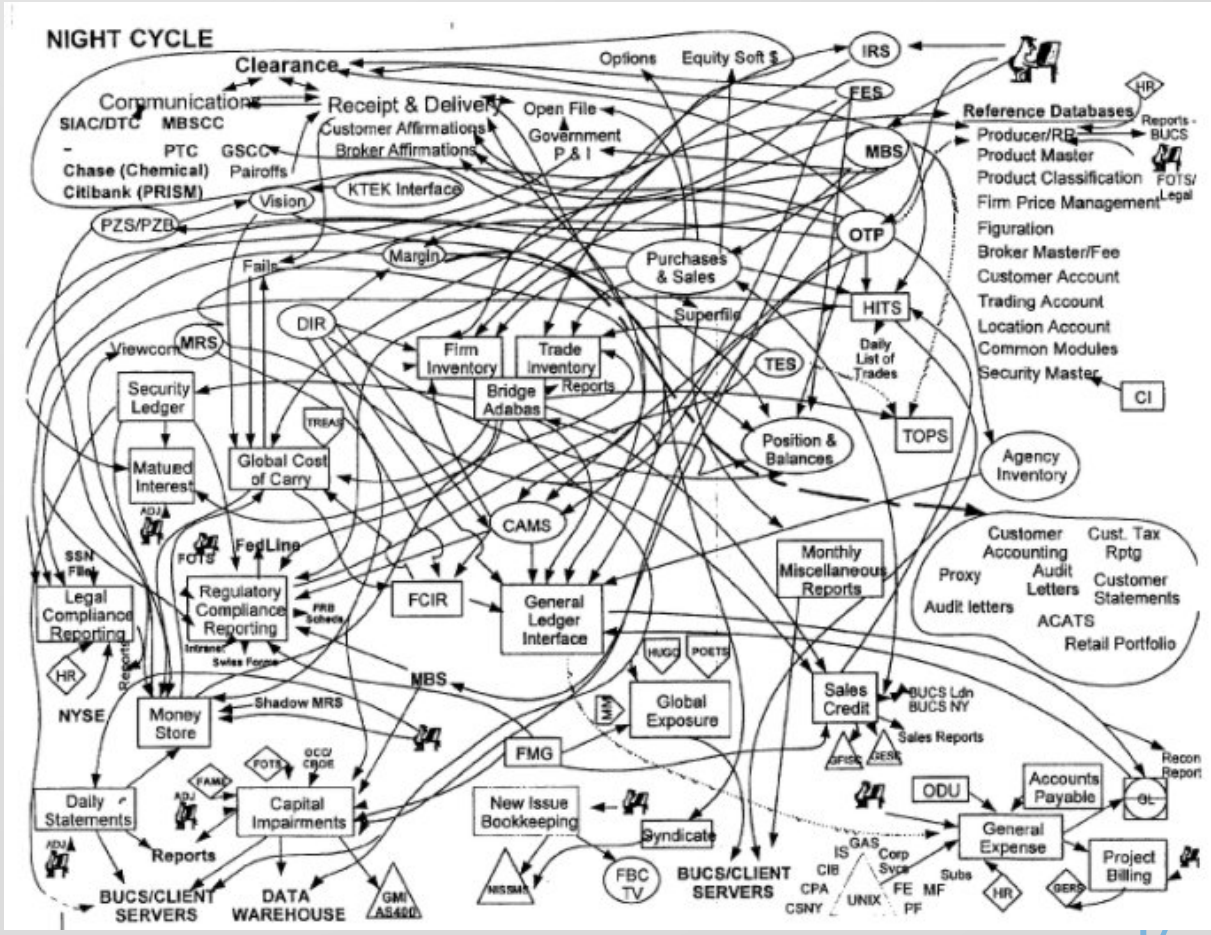
Costco 2019 10-K

THE CAUSAL MODEL –MATERIAL WEAKNESS DISCLOSURES



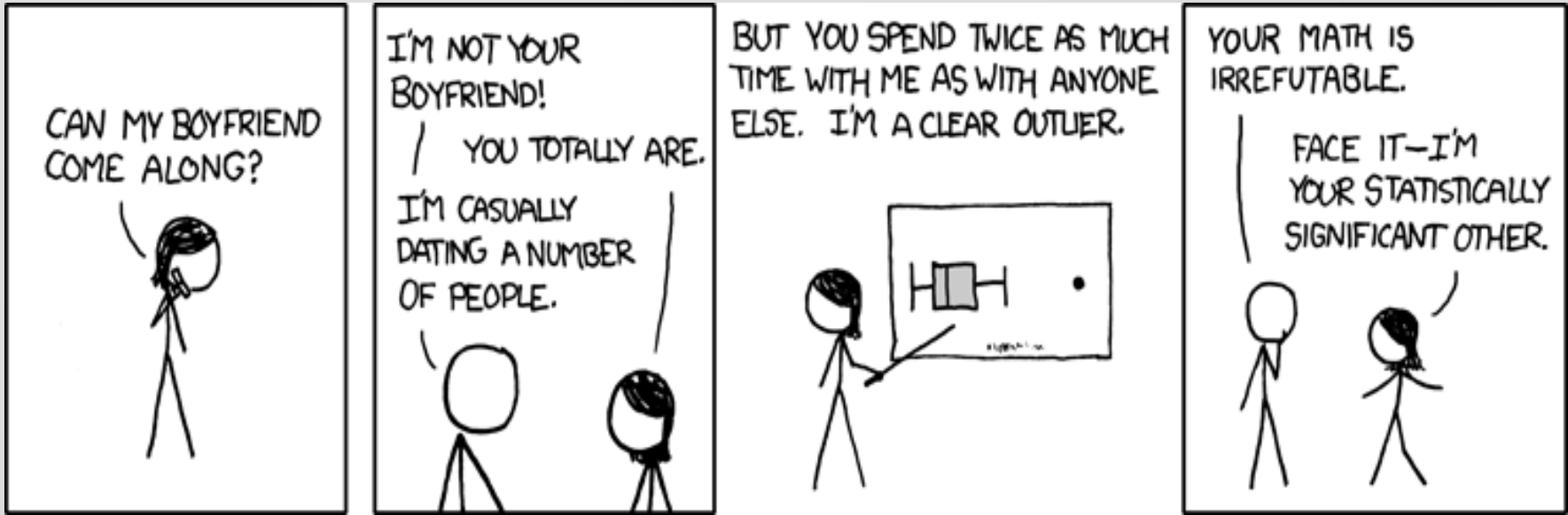
OUR WORLD IS COMPLEX AND SO IS OUR DATA

- The real world is complex and requires complex models.
- Nonetheless, the better we can design our models the more meaningful our inferences. *This is an area of continuous improvement.*



OVERVIEW OF STATISTICAL SIGNIFICANCE

- What does it mean when a scientist or media article mentions "statistical significance"?
  - Important?
  - Causal?
  - Valid?

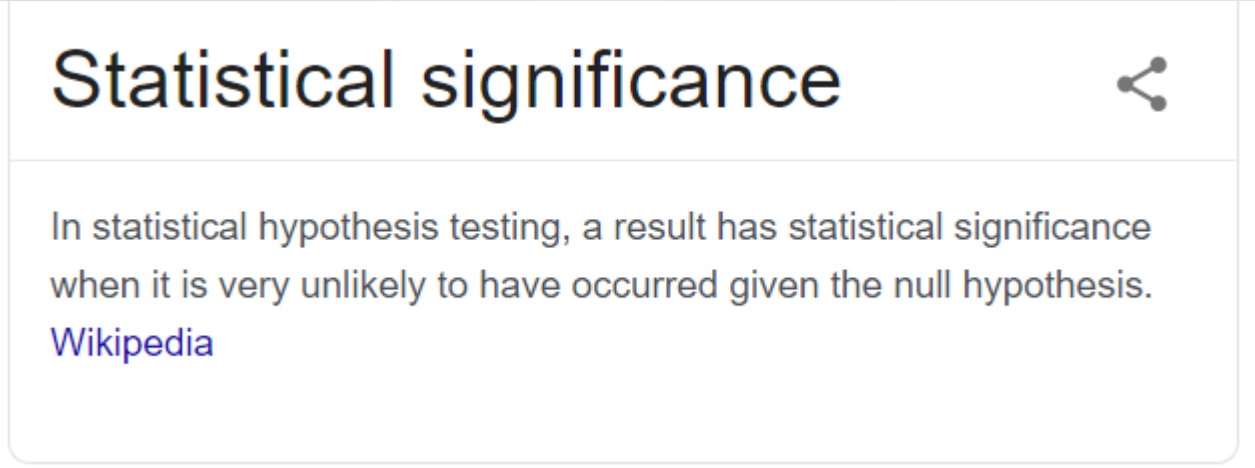


OVERVIEW OF STATISTICAL SIGNIFICANCE

- Google “significance”:

OVERVIEW OF STATISTICAL SIGNIFICANCE

- Google “statistical significance”… Wikipedia:



Much better… but what does this mean?

OVERVIEW OF STATISTICAL SIGNIFICANCE

- When estimating regression models we have obtained “standard errors”, “t-statistics”, and “p-values”
- Statistical significance is determined by p-value cutoffs, or alpha (α). Typical cutoffs are 0.10, 0.05, 0.01. See the default “star assignments” from stargazer output:

Note: *p<0.1; **p<0.05; ***p<0.01

OVERVIEW OF STATISTICAL SIGNIFICANCE

- Interpreting p-values: the probability that we would observe an effect as great or greater than the estimated effect *if the true effect were just noise or randomness* (or if the null hypothesis is true).
- This is NOT the same as:
  - The probability that the effect is true.
  - The probability that the effect is due to chance.
  - The probability that the effect is economically meaningful.

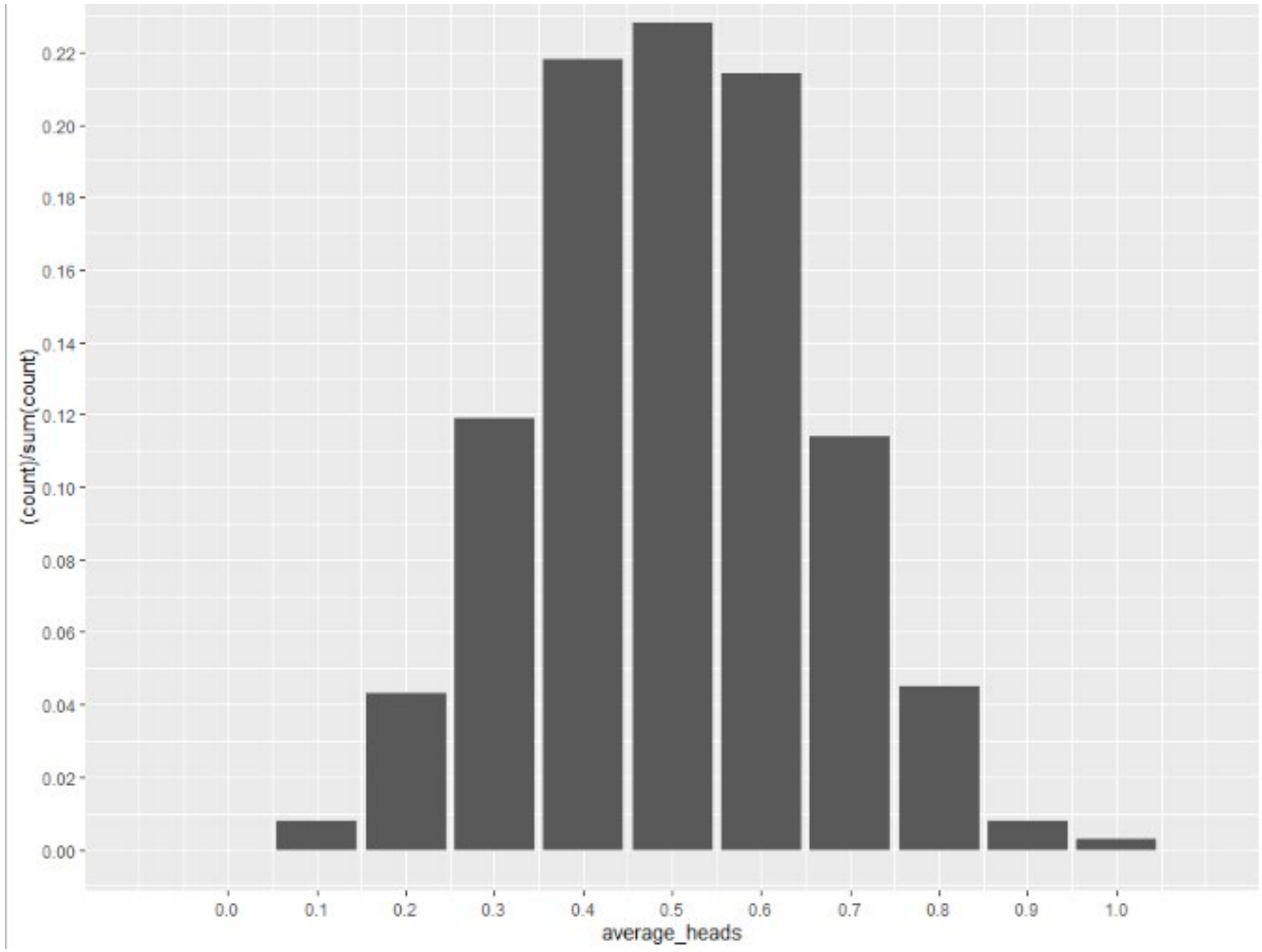
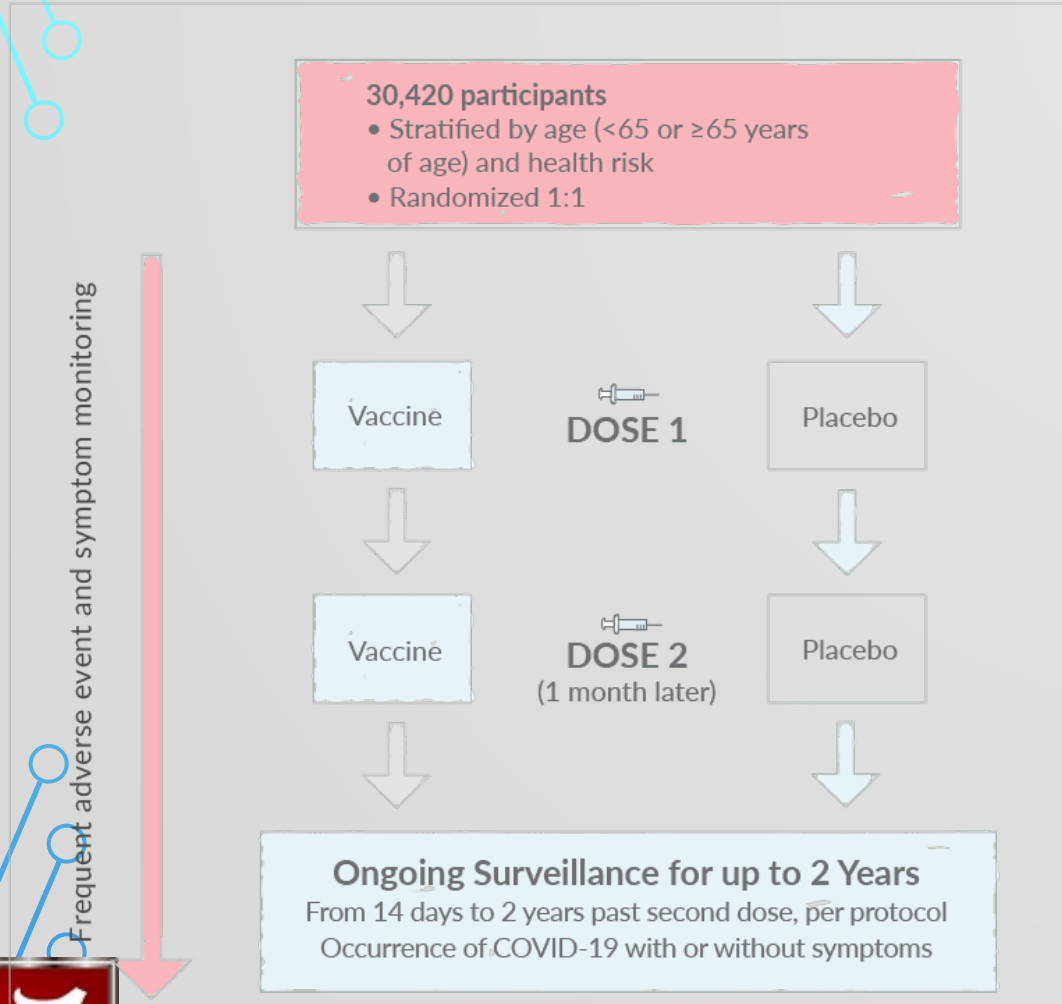
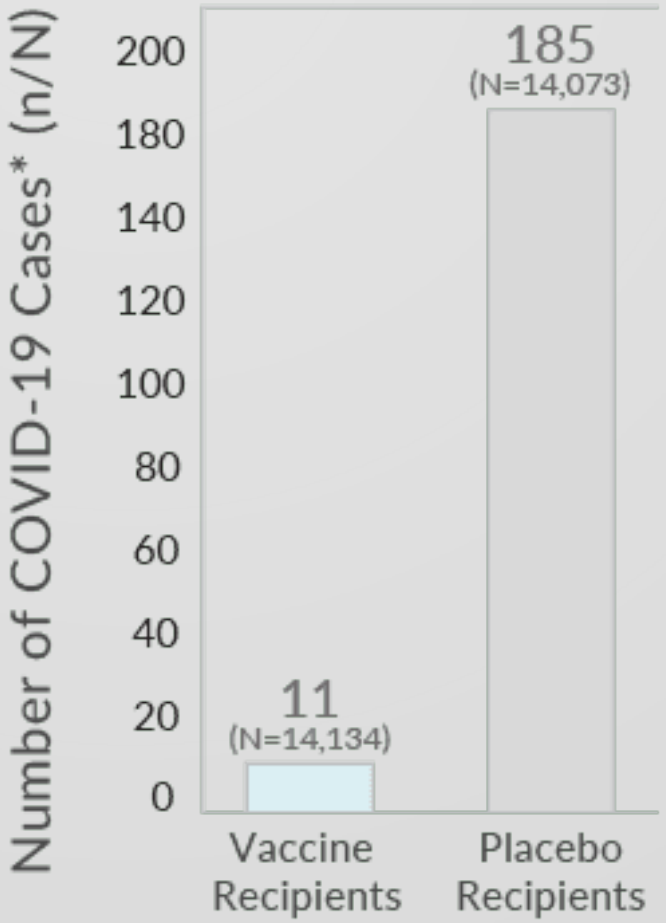


ILLUSTRATION WITH COIN FLIPS

MODERNA COVID-19 VACCINE





Vaccine Efficacy†
94.1%
(95% CI, 89.3%, 96.8%)
p<0.001‡

Source: https://eua.modernatx.com/covid19vaccine-eua/providers/clinical-trial-data-primary-series
Note: No longer available, can be found on Wayback Machine Sept. 28, 2022

MPG REGRESSION – STATISTICAL SIGNIFICANCE

MPG Regression Results

	Dependent variable:
	mpg
wt	-3.167***
	(0.741)
hp	-0.018
	(0.012)
cyl	-0.942*
	(0.551)
Constant	38.752***
	(1.787)
Observations	32
R^2	0.843
Adjusted R^2	0.826
Note:	*p<0.1; **p<0.05; ***p<0.01

Coefficient or slope = -0.942

Standard error – estimate of the coefficient's precision = 0.551

t-statistic = coefficient / standard error = -1.7

t-statistics are used to determine the p-value. The higher the t-statistic the lower the p-value and the greater the "statistical significance."

Rule of thumb: t-stats > 2 are significant at the 0.05 level.

In this case the p-value is 0.098.

CPU REGRESSION – ECONOMIC SIGNIFICANCE

CPU Regression Results

	Dependent variable:		
	log(price_usd)		
	(1)	(2)	(3)
log(cpu_mark)	0.712***		0.695***
	(0.017)		(0.017)
brandIntel		0.655***	0.229***
		(0.064)	(0.049)
Constant	-0.805***	4.642***	-0.829***
	(0.140)	(0.055)	(0.139)
Observations	2,026	2,026	2,026
R^2	0.475	0.049	0.481
Adjusted R^2	0.475	0.048	0.480
Note:	*p<0.1; **p<0.05; ***p<0.01		

- Can you interpret the coefficients here? Note that benchmark and price are in natural log form.
- What about statistical significance?
- R-squared?

WORDS OF CAUTION WITH P-VALUES

- P-values do NOT tell us that the effect is “true” or “important.”
- The practical use of p-values is susceptible to “false-positives”, especially when the researcher sorts through data and models to identify p-values below a threshold (sometimes called p-hacking or data dredging):
  - Example (p-hacking): 882: Significant - explain xkcd
  - Example 2 (data dredging): Thomas Steinke on X

OVERALL TAKEAWAYS

- Regressions can be used to predict outcomes or to determine how variables are related to one another.
- Uncovering causal effects requires assumptions and a firm understanding of the setting and data.
- Statistical significance is not the same as economic significance!

OTHER RESOURCES AND ACKNOWLEDGEMENTS

- Credit to the following resources for solidifying my understanding of these issues and improving my ability to communicate causal issues in understandable terms:
  - Mastering Metrics: The Path from Cause to Effect
  - Statistical Modeling, Causal Inference, and Social Science (columbia.edu)
- I also encourage Causal Inference The Mixtape (scunning.com). And includes examples in Python/R!