GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS, PART 2

Tan Kwan Chong

Chief Data Scientist, Booz Allen Hamilton

LEARNING OBJECTIVES

- Explain the difference between causation and correlation
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (p-values, confidence intervals)

INTRODUCTION

CAUSATION AND CORRELATION

CAUSATION AND CORRELATION

- If an association is observed, the first question to ask should always be… is it real?

- Think of various examples you've seen in the media related to food.

CAUSATION AND CORRELATION

FRIDAY, MAY 23, 2014 08:15 AM SGT

10 foods touted as health miracles, then vilified as health hazards

One reason Americans have trouble maintaing a healthy diet: They're suffering from "food information overload"

ALEX HENDERSON, ALTERNET

Pocket | SKIP TO COMMENTS

TOPICS: ALTERNET, CAFFEINE, COFFEE, OLIVE OIL, ORANGE JUICE, UNIVERSITY OF MINNESOTA, LIFE NEWS

Home ›

HEALTH NEWS

E-mail this page | Print this page

A few cups of coffee may lower colon cancer risk

Posted: 01 August 2007 1708 hrs

TOKYO : Drinking a few cups of coffee a day may lower the risk of advanced colon cancer, at least for women, Japanese researchers said Wednesday.

Photos 1 of 1

The study, supported by Japan's health ministry, showed women who drink more than three cups of coffee a day were 56 percent less likely to develop advanced colon cancer than those who drink no coffee at all.

"Drinking coffee sustains the secretion of bile acid and keeps down cholesterol levels, the mechanisms thought to prevent colon cancer," the report said.

But unfortunately the effect was not seen in men, the medical research team said.

Many men smoke and drink alcohol more than women, and those habits probably offset the effect of coffee, the study said.

The research team tracked down about 96,000 people in Japan aged from 40 to 69 between the early 1990s and 2002, of whom 726 men and 437 wom[…] suffered colon cancer.

Causal claims are often inconsistent and contradictory!

CancerConsultants.com
oncology resource center

Patient Home | Professional Home | Newsletters | Feedback Survey

Search: Medline | CancerConsultants.com | Both | Search

Main Menu: Home | Conference Coverage | Current Topics in Oncology | Cancer News | Disease Centers | Physician Resources | About Us

Cancer News: Rectal Cancer: Article | Printable Version

Rectal Cancer News

Coffee Does Not Decrease Risk of Colorectal Cancer

Researchers from the Harvard School of Public Health have reported that, contrary to the results of several previous studies, coffee consumption does not appear to reduce the risk of colorectal cancer. The details of this study were reported in the April 1, 2009 issue of the *International Journal of Cancer*.[1]

Habitual coffee drinking has been associated with a reduced risk of mortality and chronic diseases, including cancer. Current evidence suggests that coffee consumption is associated with a reduced risk of liver, kidney, and to a lesser extent, premenopausal breast cancer and colorectal cancer; coffee consumption has no association with prostate, pancreas, and ovarian cancers.

Some studies have indicated that coffee may have a protective effect against colon cancer; however, researchers continue to evaluate this link in an effort to establish more direct evidence. In order to examine the relationship between coffee consumption and colorectal cancer, researchers from Harvard conducted a review of 12 studies that included 646,848 participants and 5,403 cases of colorectal cancer.

They evaluated high versus low coffee consumption and found no significant effect of coffee consumption on colorectal cancer risk. The review included four studies in the United States, five in Europe, and three in Japan. The data from each country was very similar. There were no significant differences by gender or site of cancer; however, there was a slight inverse relationship between coffee consumption and colon cancer for women, which was even more pronounced among Japanese women (21% for total study, 38% for Japanese women).

The researchers observed that inverse associations between coffee consumption and colorectal cancer "were slightly stronger in studies that controlled for smoking and alcohol and in studies with shorter follow-up times."

They concluded that coffee is "unlikely to have a strong protective effect on colorectal cancer risk"; however, they also note that it does not appear to increase the risk of colorectal cancer either.

SPURIOUS CORRELATIONS

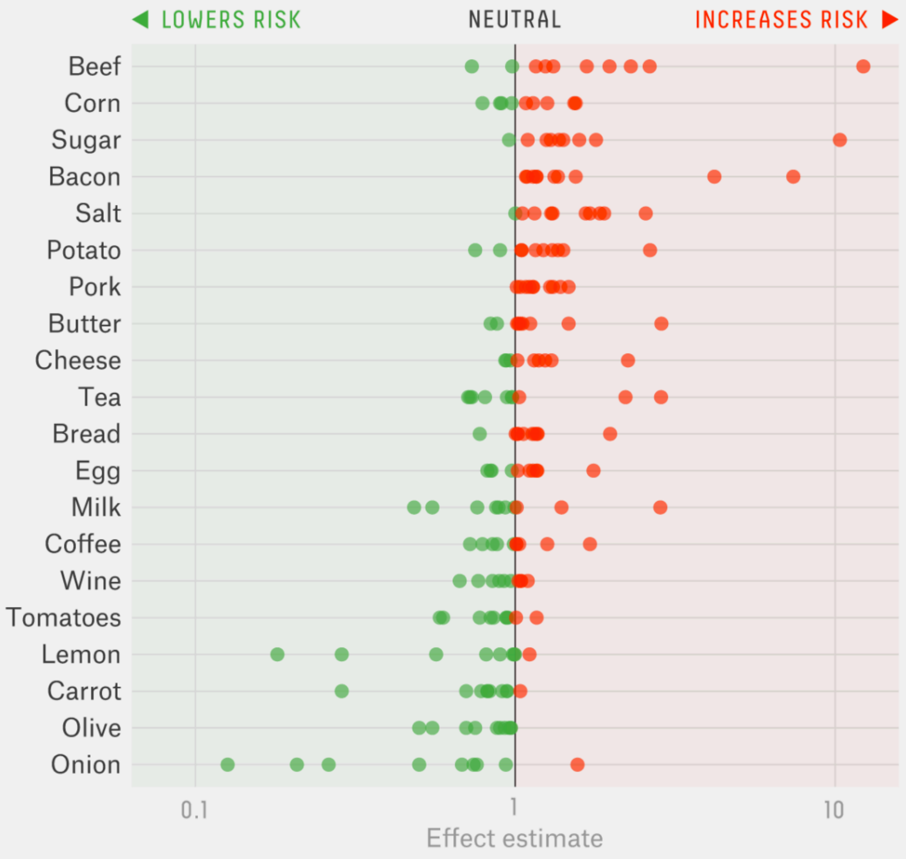
Our shocking new study finds that ...

EATING OR DRINKING	IS LINKED TO	P-VALUE
Raw tomatoes	Judaism	<0.0001
Egg rolls	Dog ownership	<0.0001
Energy drinks	Smoking	<0.0001
Potato chips	Higher score on SAT math vs. verbal	0.0001
Soda	Weird rash in the past year	0.0002
Shellfish	Right-handedness	0.0002
Lemonade	Belief that “Crash” deserved to win best picture	0.0004
Fried/breaded fish	Democratic Party affiliation	0.0007
Beer	Frequent smoking	0.0013
Coffee	Cat ownership	0.0016
Table salt	Positive relationship with Internet service provider	0.0014
Steak with fat trimmed	Lack of belief in a god	0.0030
Iced tea	Belief that “Crash” didn’t deserve to win best picture	0.0043
Bananas	Higher score on SAT verbal vs. math	0.0073
Cabbage	Innie bellybutton	0.0097

SOURCE: FFQ & FIVETHIRTYEIGHT SUPPLEMENT

Foods that may or may not give you cancer

Risk estimates for 20 foods (each studied at least 10 times) from a 2012 meta-analysis



One outlier study not shown (corn, risk estimate of 19.43).

FIVETHIRTYEIGHT SOURCE: AMERICAN JOURNAL OF CLINICAL NUTRITION

http://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

CAUSATION AND CORRELATION

- Why is this?
- Sensational headlines?
- There is neglect of a robust data analysis.

CAUSATION AND CORRELATION

- There is also often a lack of understanding of the difference between *causation* and *correlation*.

- Understanding this difference is critical in the data science workflow, especially when **Identifying** and **Acquiring** data.

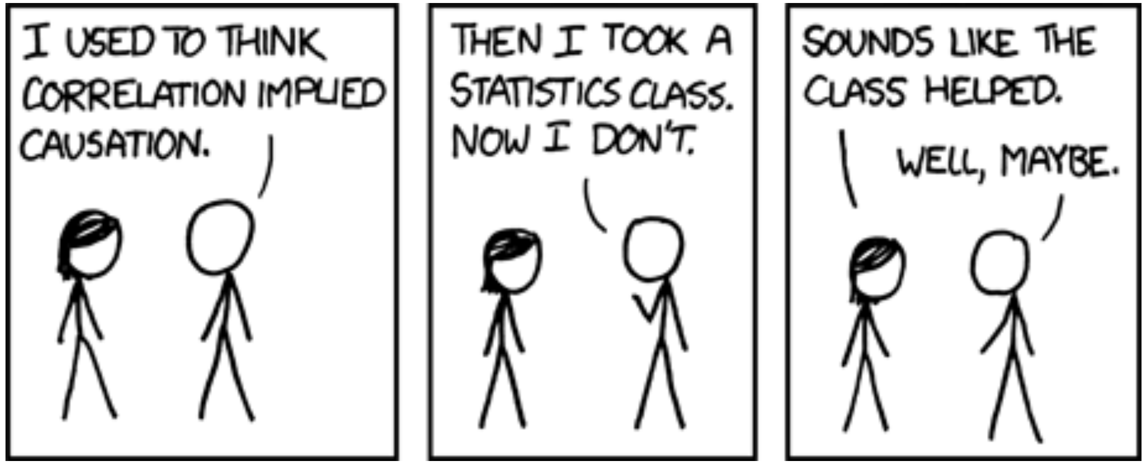
- We need to fully articulate our question and use the right data to answer it, including any *confounders*.

CAUSATION AND CORRELATION

- Additionally, this comes up when we **Present** our results to stakeholders.

- We don't want to overstate what our model measures.

- Be careful not to say "caused" when you really mean "measured" or "associated".



LECTURE

CAUSATION VS CORRELATION

CAUSAL CRITERIA

- Causal criteria is one approach to assessing causal relationships.
- However, it’s ***very hard to define*** universal causal criteria.
- One attempt that is commonly used in the medical field is based on work by Bradford Hill.

CAUSAL CRITERIA

- He developed a list of "tests" that an analysis must pass in order to indicate a causal relationship. A relationship is more likely to be causal if:
  - **Strength**: The correlation coefficient is large and statistically significant
  - **Consistency**: It can be replicated
  - **Specificity**: There is no other likely explanation
  - **Temporality**: The effect always occurs after the cause
  - **Gradient**: A greater exposure to the suspected cause leads to a greater effect
  - **Plausibility**: There is a plausible mechanism between the cause and the effect
  - **Coherence**: It is compatible with related facts and theories
  - **Experiment**: It can be verified experimentally
  - **Analogy**: There are proven relationships between similar causes and effects

CAUSAL CRITERIA

- This is not an exhaustive checklist, but it’s useful for understanding that your predictor/exposure **must have occurred before your outcome**.

- For example, in order for smoking to cause cancer, one must have started smoking prior to getting cancer.

CAUSAL CRITERIA

- Most commonly, we find an *association* between two variables. This means there is an observed **correlation** between the variables.
- We may not fully understand the causal direction (e.g. does smoking cause cancer or does cancer cause smoking?).
- We also might not understand *other* factors influencing the association.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is the difference between causation and association?

DELIVERABLE

Answers to the above questions

INTRODUCTION

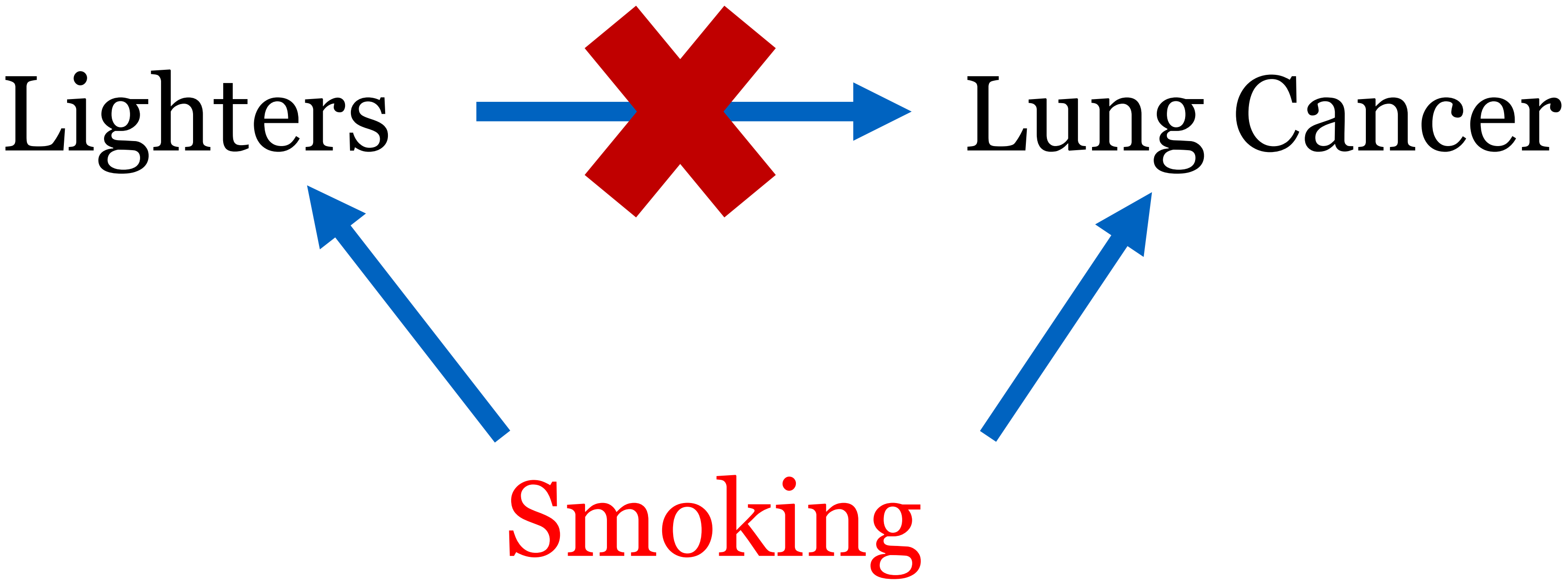
CONFOUNDING AND DAGS

CONFOUNDING

- Often times, associations may be influenced by another *confounding* factor.
- Let's say we did an analysis to understand what causes lung cancer.
- We find that people who carry cigarette lights are 2.4 times more likely to contract lung cancer as people who don't carry lighters.
- Does this mean that the lighters are causing cancer?

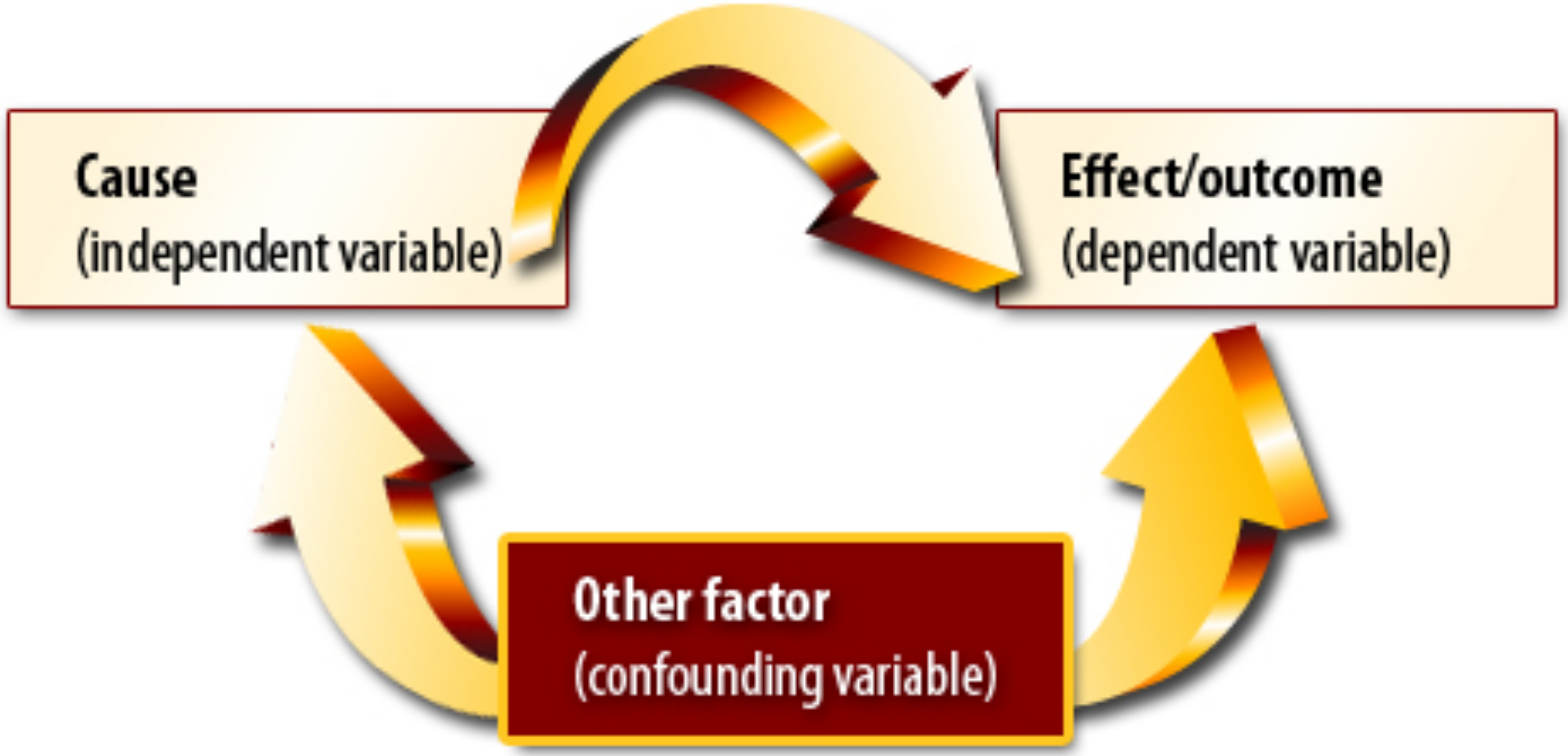
CONFOUNDING

- No!



CONFOUNDING

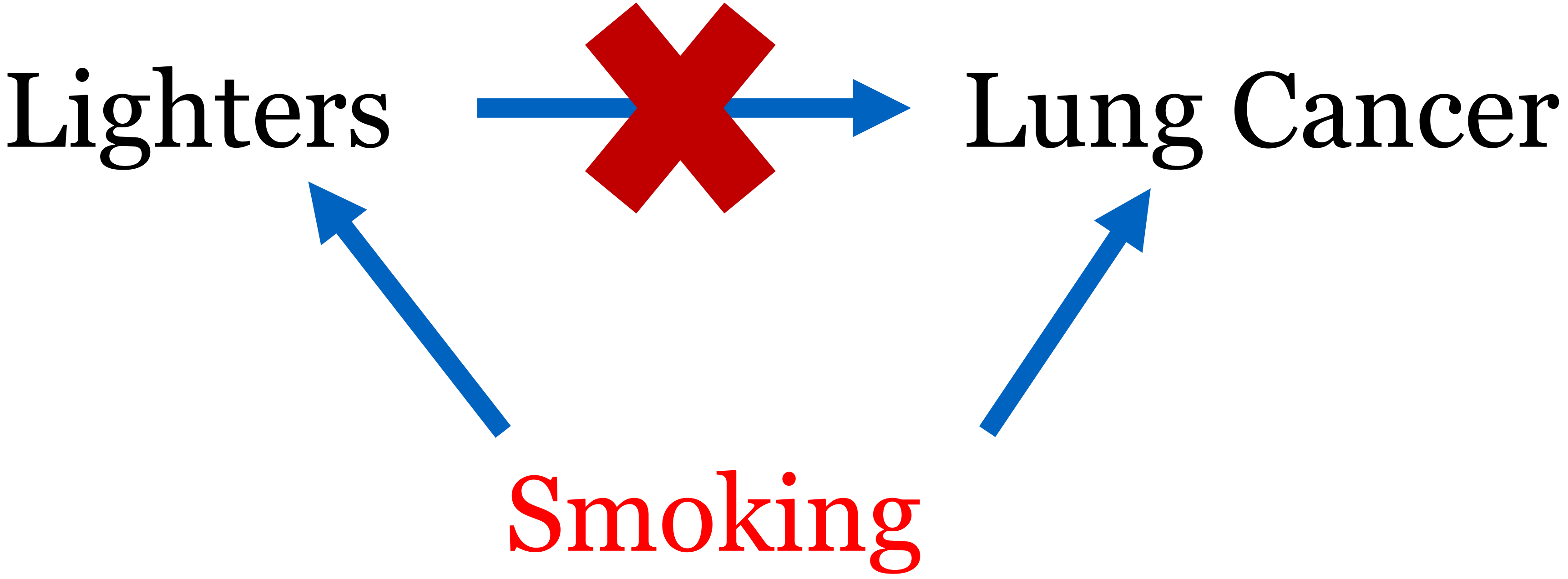
- Confounding variables often hide the true association between causes and outcomes.



DIRECTED ACYCLIC GRAPH

- A *Directed Acyclic Graph* (DAG) can help determine which variables are most important for your model. It helps visually demonstrate the logic of your models.

- A DAG always includes at least one exposure/predictor and one outcome.



DIRECTED ACYCLIC GRAPH

- Suppose we have the following output from a model:

Dep. Variable:	Sales	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	312.1
Date:	Thu, 03 Sep 2015	**Prob (F-statistic):**	1.47e-42
Time:	18:58:58	**Log-Likelihood:**	-519.05
No. Observations:	200	**AIC:**	1042.
Df Residuals:	198	**BIC:**	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	7.0326	0.458	15.360	0.000	6.130 7.935
TV	0.0475	0.003	17.668	0.000	0.042 0.053

Omnibus:	0.531	**Durbin-Watson:**	1.935
Prob(Omnibus):	0.767	**Jarque-Bera (JB):**	0.669
Skew:	-0.089	**Prob(JB):**	0.716
Kurtosis:	2.779	**Cond. No.**	338.

DIRECTED ACYCLIC GRAPH

- The exposure/predictor is TV ads, associated with the outcome: sales.
- We can measure the strength to demonstrate a strong association.
- What other factors may increase sales?
- What other types of ads?

DIRECTED ACYCLIC GRAPH

- The DAG for this might look like the following:

THINK, PAIR, SHARE

DAGS

SEASONALITY

- Suppose TV ads were run in November/December (peak buying season) while Google ads were run during February/March (low buying season).

- If we compare the two, we're likely to reach the wrong conclusion! Seasonal trends are affecting our associations.

- This is an example of *bias* and *confounding*. It isn't that TV ads are better than Google ads; it's that November/December is a better buying season than February/March, an inherent bias.

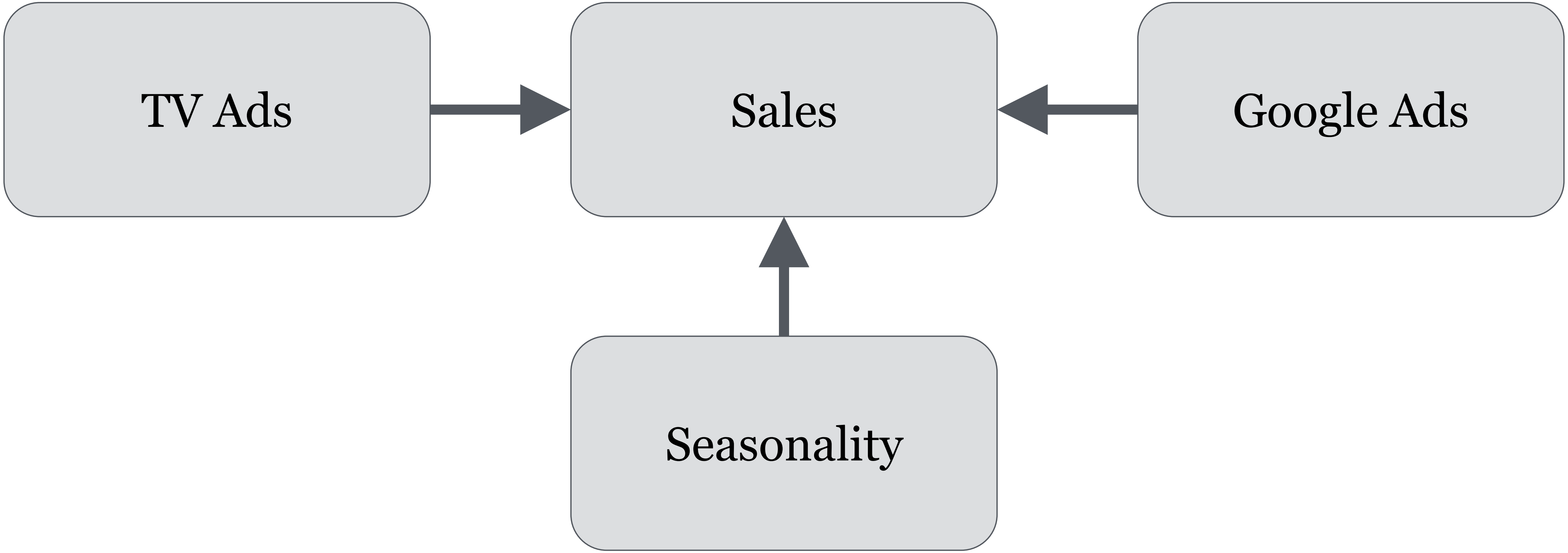
SEASONALITY

- Let’s take a look at the association between TV Ads and Sales while taking into account *seasonality* (recurring regular patterns over time).

- What are some examples of seasonality with relation to sales?

SEASONALITY

- A DAG incorporating seasonality might look like this.



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is bias?
2. What is confounding?
3. What could we do differently in this example to avoid these elements?

DELIVERABLE

Answers to the above questions

A FEW KEY TAKEAWAYS

- It is important to have deep subject area knowledge to be aware of biases in your field. This knowledge supplements statistical techniques.

- A DAG can be a useful tool for thinking through the logic of your model.

- There is a difference between causation and correlation. Statistics usually show *correlation*, not *causation* (remember our smoking example).

- Good data is important. Your analysis is only as good as your understanding of the problem and the data you have to work with.

INTRODUCTION

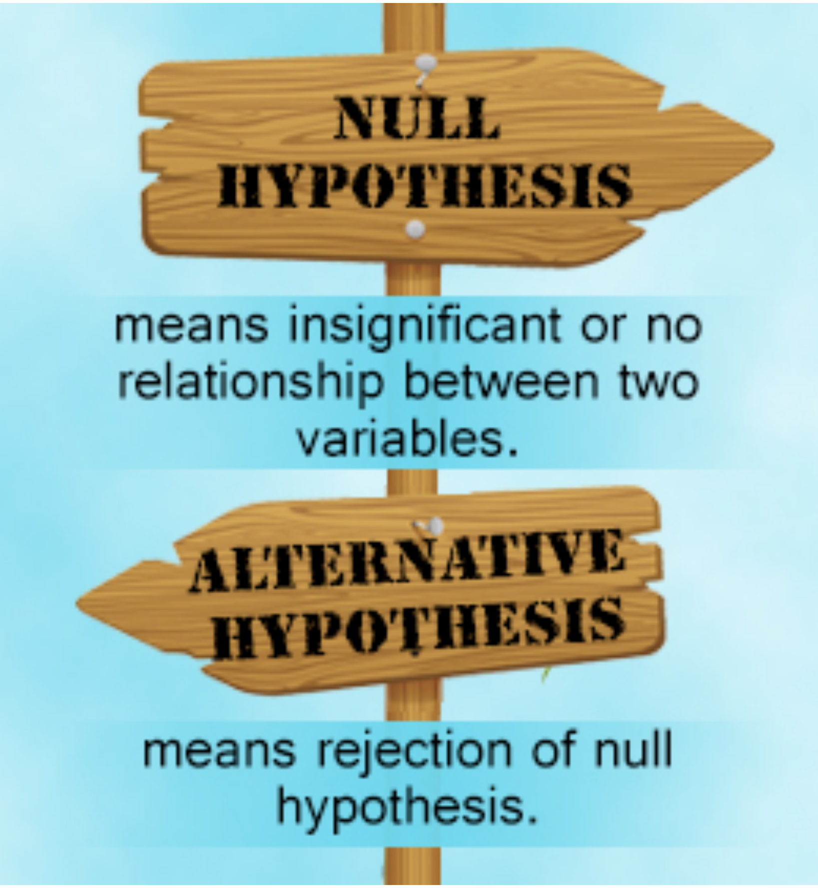
HYPOTHESIS TESTING

HYPOTHESIS TESTING

- How can we tell the difference between two groups of observations (e.g. smokers vs. non-smokers)?
- Imagine we are testing the health of smokers vs. non-smokers. At a cursory glance, our results may show that smokers are marginally healthier than non-smokers.
- Are they healthier due to random chance or is there a statistically significant difference? Maybe we happened to assemble a strange group of smoking triathletes and a group of non-smoking couch potatoes.
- This is where hypothesis testing can help.

HYPOTHESIS TESTING STEPS

- First, you need a hypothesis to test, referred to as the *null hypothesis*. The opposite of this would be the *alternative hypothesis*.



HYPOTHESIS TESTING STEPS

- For example, if we want to test the relationship between gender and sales, we may have the following hypotheses.
- Null hypothesis: There is no relationship between Gender and Sales.
- Alternative hypothesis: There is a relationship between Gender and Sales.

HYPOTHESIS TESTING STEPS

- Once you have your hypotheses, you can check whether the data supports rejecting the null hypothesis or failing to reject the hypothesis.

- **Note**: Failing to reject the null is **NOT** the same as accepting the alternate. While the alternative hypothesis **might** be true, we don't have enough data to support that claim specifically.

- Keep this in mind so you don't overstate your findings.

DEMO

HYPOTHESIS TESTING CASE STUDY

HYPOTHESIS TESTING CASE STUDY

- We’re going to walk through Part 1 of the guided-demo-starter-code notebook in the class repo for lesson 4.
- There are several questions to answer. We’ll answer those questions in small groups and then discuss with the class.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is the null hypothesis?
2. Why is this important to use?

DELIVERABLE

Answers to the above questions

INTRODUCTION

VALIDATE YOUR FINDINGS

VALIDATE YOUR FINDINGS

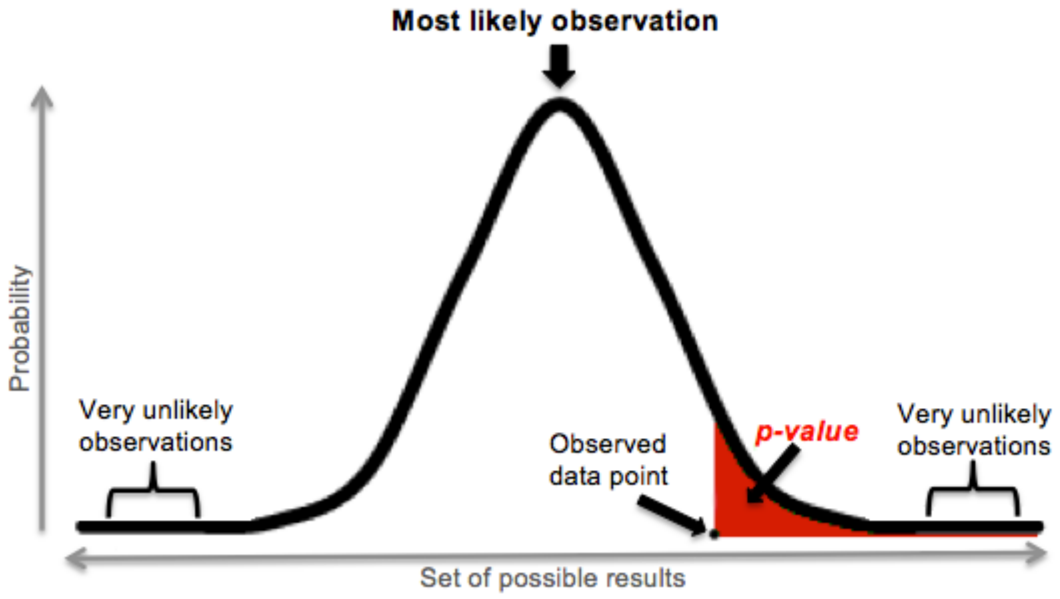
- We know how to carry out a hypothesis test, but how do we tell if the association we found is *statistically significant*?

- *Statistical significance* is the likelihood that a result or relationship is caused by something other than random chance.

- Statistical hypothesis testing is traditionally employed to determine if a result is statistically significant or not.

VALIDATE YOUR FINDINGS

- Typically, a cut point of 5% is used. This means that we say something is statistically significant if there is a less than a 5% chance that our finding was due to random chance alone.



A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

VALIDATE YOUR FINDINGS

TABLE 1

Relationship between Common Language and Hypothesis Testing

COMMON LANGUAGE	STATISTICAL STATEMENT	CONVENTIONAL TEST THRESHOLD
"Statistically significant" "Unlikely due to chance"	The null hypothesis was rejected.	$P < 0.05$
"Not significant" "Due to chance"	The null hypothesis could not be rejected.	$P > 0.05$

VALIDATE YOUR FINDINGS

- When we present results, we say we found something significant using this criteria.

- We will use an example to dive further into this and understand p-values and confidence intervals.

DEMO

P-VALUES AND CONFIDENCE INTERVALS CASE STUDY

P-VALUES AND CONFIDENCE INTERVALS CASE STUDY

- We’re now going to walk through Part 2 of the guided-demo-starter-code notebook in the class repo for lesson 4.

- There are several questions to answer. We’ll answer those questions in small groups and then discuss with the class.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What does a 95% confidence interval indicate?

DELIVERABLE

Answers to the above questions

DEMO

A/B TESTING

INDEPENDENT PRACTICE

INTERPRETING RESULTS

ACTIVITY: INTERPRETING RESULTS

EXERCISE

DIRECTIONS (35 minutes)

1. Using the lab-start-code-4, you will look through a variety of analyses and interpret the findings.
2. You will be presented with a series of outputs and tables from a published analysis.
3. Read the outputs and determine if the findings are statistically significant or not.

DELIVERABLE

Answers to the questions in the notebook

CONCLUSION

LAB REVIEW

LAB REVIEW

- Let’s review the answers to the questions in the labs.
- Any other questions?

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project: Unit Project 1

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET