GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS II

Angelo Klin
Katra Analytics

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

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Explain the difference between Variance and Bias
- Use descriptive statistics to understand your data

OPENING

STATISTICS FUNDAMENTALS II

OPENING

INTRODUCTION

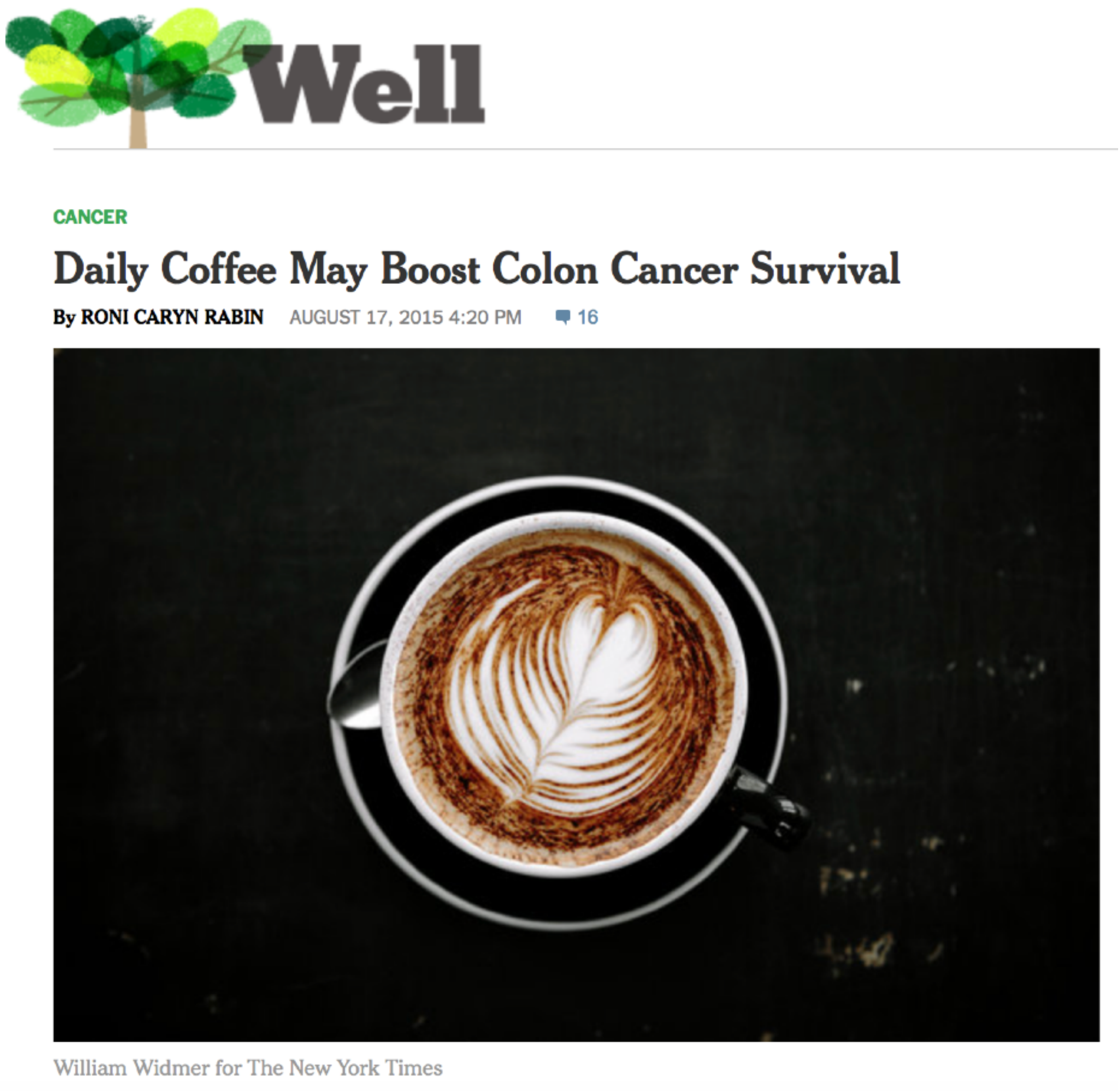
CAUSATION AND CORRELATION

CAUSATION AND CORRELATION

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

- Think of various examples you have seen in the media related to food being both good and bad

CAUSATION AND CORRELATION

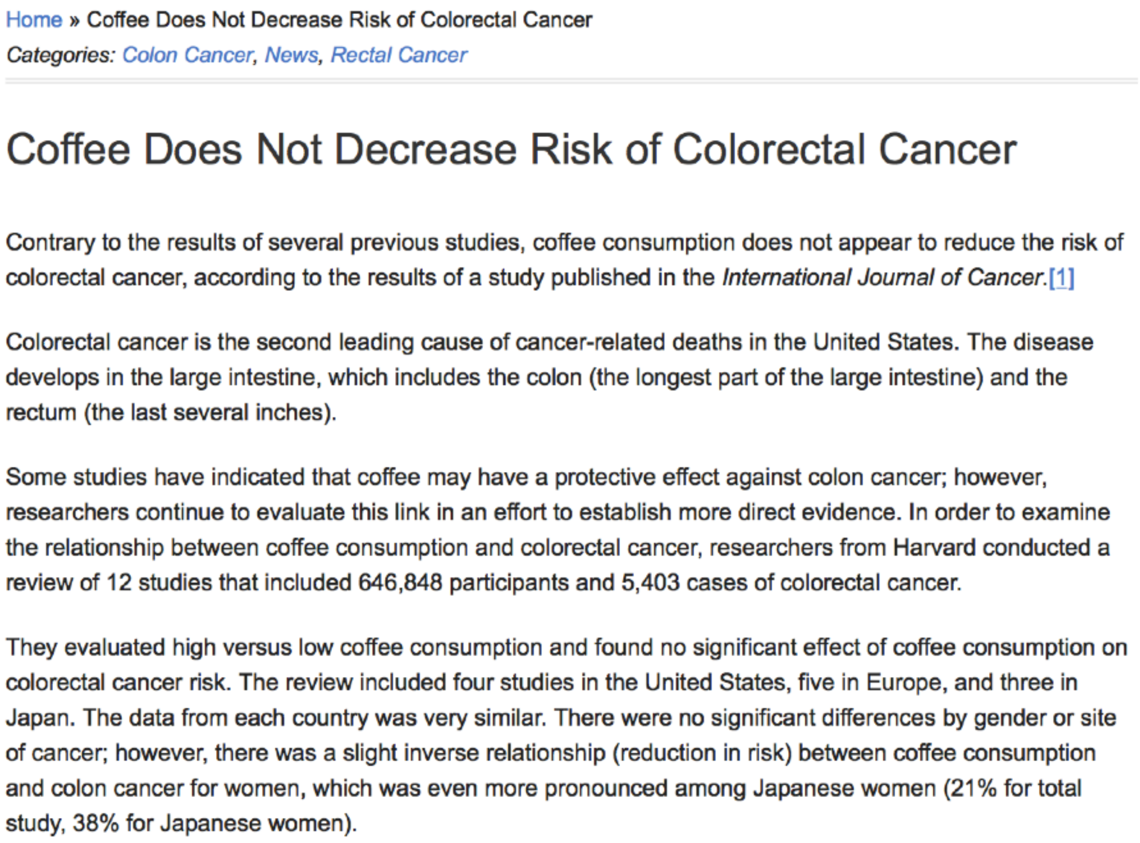


CANCER

Daily Coffee May Boost Colon Cancer Survival

By RONI CARYN RABIN AUGUST 17, 2015 4:20 PM 16

William Widmer for The New York Times



Home » Coffee Does Not Decrease Risk of Colorectal Cancer

Categories: Colon Cancer, News, Rectal Cancer

Coffee Does Not Decrease Risk of Colorectal Cancer

Contrary to the results of several previous studies, coffee consumption does not appear to reduce the risk of colorectal cancer, according to the results of a study published in the *International Journal of Cancer*.[1]

Colorectal cancer is the second leading cause of cancer-related deaths in the United States. The disease develops in the large intestine, which includes the colon (the longest part of the large intestine) and the rectum (the last several inches).

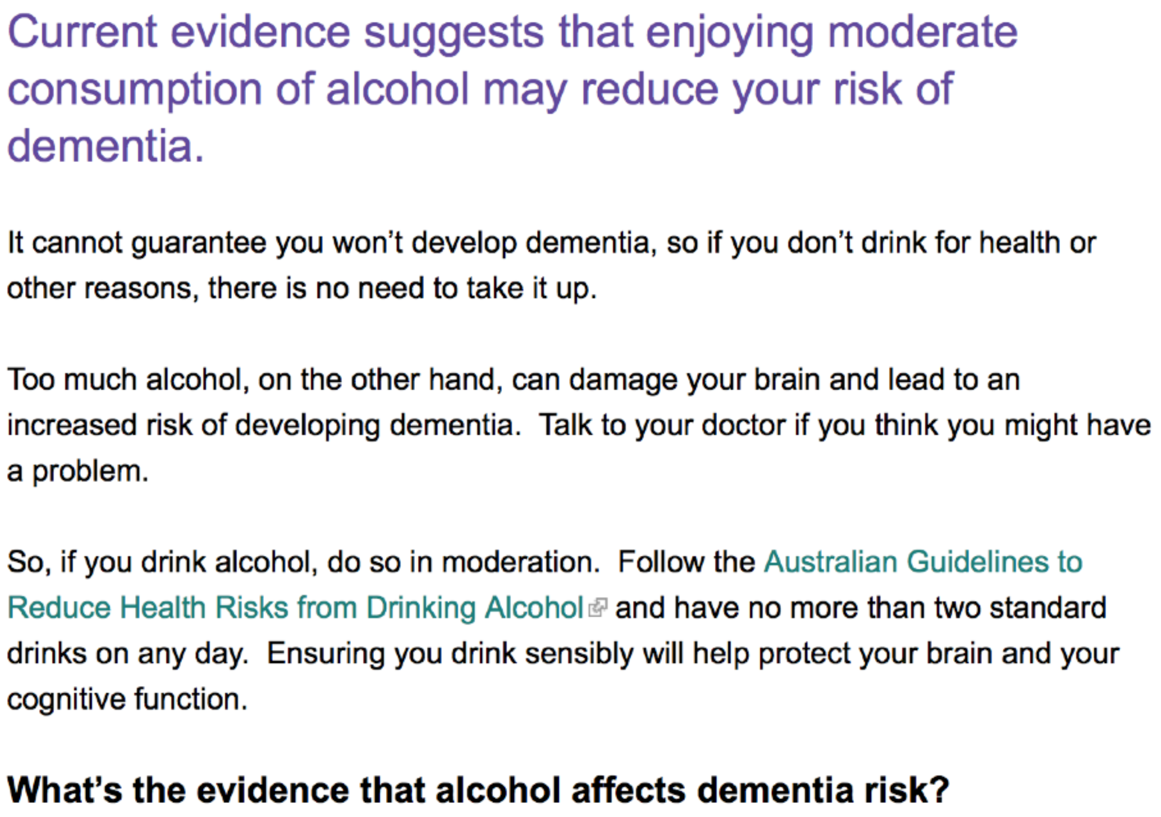
Some studies have indicated that coffee may have a protective effect against colon cancer; however, researchers continue to evaluate this link in an effort to establish more direct evidence. In order to examine the relationship between coffee consumption and colorectal cancer, researchers from Harvard 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 (reduction in risk) between coffee consumption and colon cancer for women, which was even more pronounced among Japanese women (21% for total study, 38% for Japanese women).

CAUSATION AND CORRELATION

- Why is this?
  - Sensational headlines?
  - Neglect of a robust data analysis?

CAUSATION AND CORRELATION



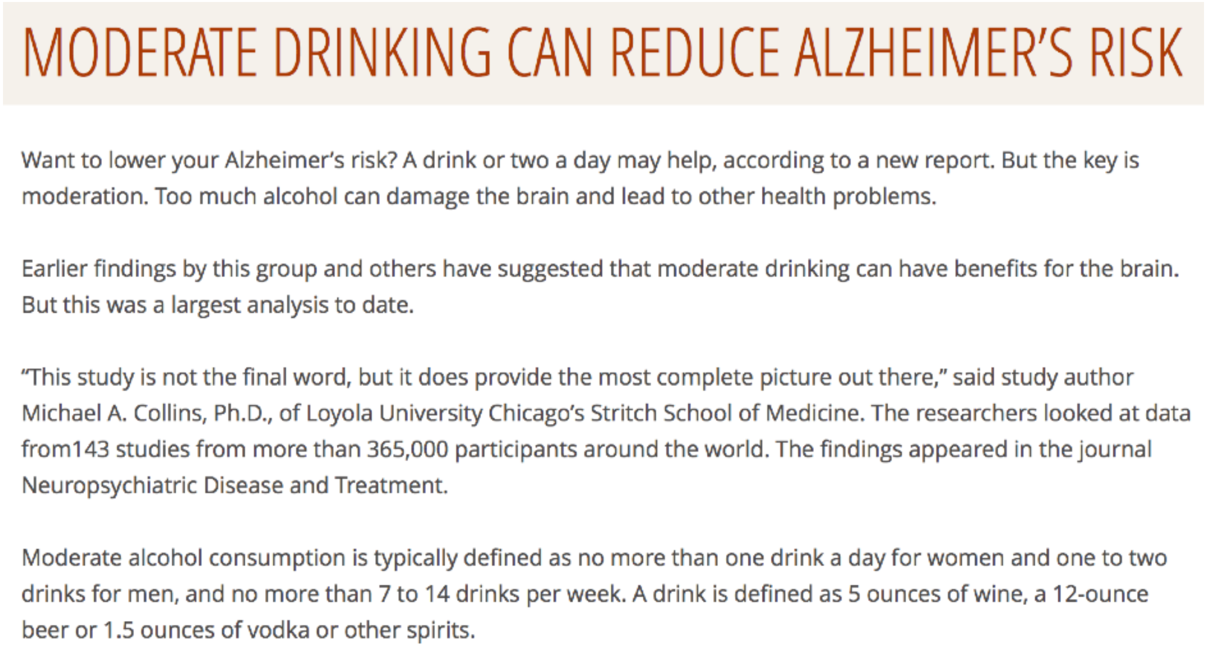
Current evidence suggests that enjoying moderate consumption of alcohol may reduce your risk of dementia.

It cannot guarantee you won't develop dementia, so if you don't drink for health or other reasons, there is no need to take it up.

Too much alcohol, on the other hand, can damage your brain and lead to an increased risk of developing dementia. Talk to your doctor if you think you might have a problem.

So, if you drink alcohol, do so in moderation. Follow the Australian Guidelines to Reduce Health Risks from Drinking Alcohol and have no more than two standard drinks on any day. Ensuring you drink sensibly will help protect your brain and your cognitive function.

What's the evidence that alcohol affects dementia risk?



MODERATE DRINKING CAN REDUCE ALZHEIMER'S RISK

Want to lower your Alzheimer's risk? A drink or two a day may help, according to a new report. But the key is moderation. Too much alcohol can damage the brain and lead to other health problems.

Earlier findings by this group and others have suggested that moderate drinking can have benefits for the brain. But this was a largest analysis to date.

"This study is not the final word, but it does provide the most complete picture out there," said study author Michael A. Collins, Ph.D., of Loyola University Chicago's Stritch School of Medicine. The researchers looked at data from143 studies from more than 365,000 participants around the world. The findings appeared in the journal Neuropsychiatric Disease and Treatment.

Moderate alcohol consumption is typically defined as no more than one drink a day for women and one to two drinks for men, and no more than 7 to 14 drinks per week. A drink is defined as 5 ounces of wine, a 12-ounce beer or 1.5 ounces of vodka or other spirits.

CAUSATION AND CORRELATION

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

- 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
- We do not want to overstate what our model measures
- Be careful not to say "caused" when you really mean "measured" or "associated"

LECTURE

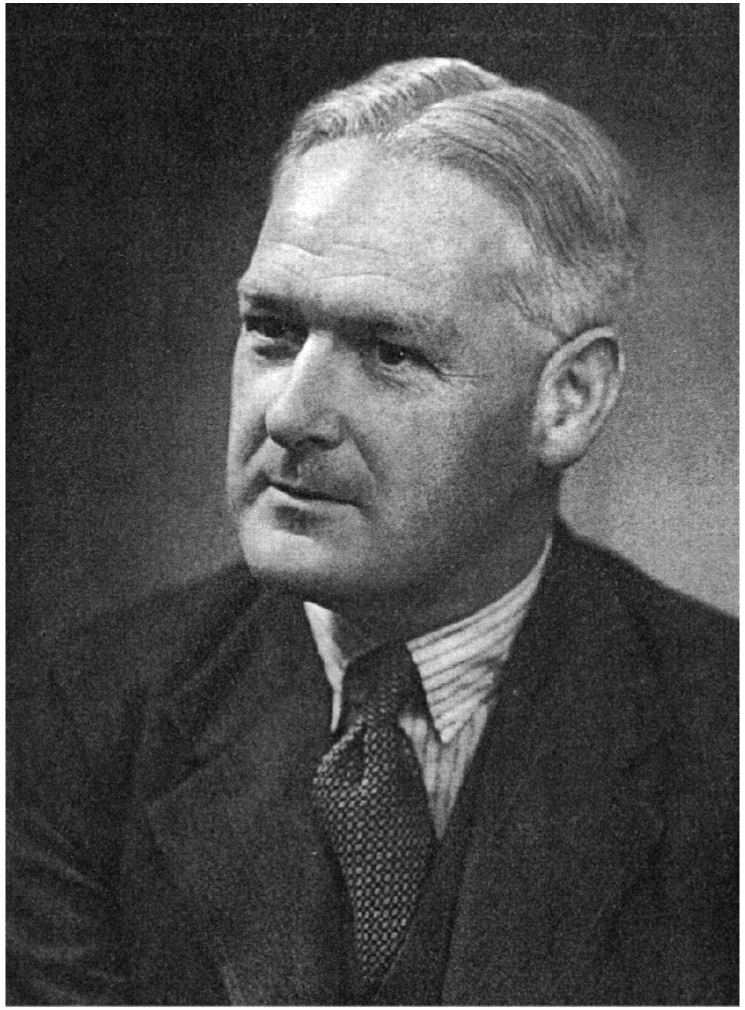
CAUSATION VERSUS CORRELATION

CAUSAL CRITERIA

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

CAUSAL CRITERIA

- Bradford Hill developed a list of “tests” that an analysis must pass in order to indicate a causal relationship
  1. Strength of association
  2. Consistency
  3. Specificity
  4. Temporality
  5. Biological gradient
  6. Plausibility
  7. Coherence
  8. Experiment
  9. Analogy



CAUSAL CRITERIA

- This is not an exhaustive checklist, but it is 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

- Most commonly we find an association between two variables. This means that we observe a **correlation** between them

- However, we may not fully understand the causal direction or how other factors may be influencing the association we are observing

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (5 MINUTES)

1. What is the difference between causation and association?

INTRODUCTION

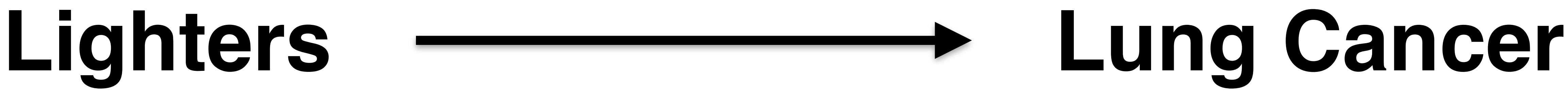
CONFOUNDING

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 lighters are 2.4 times more likely to contract lung cancer as people who do not carry lighters

CONFOUNDING

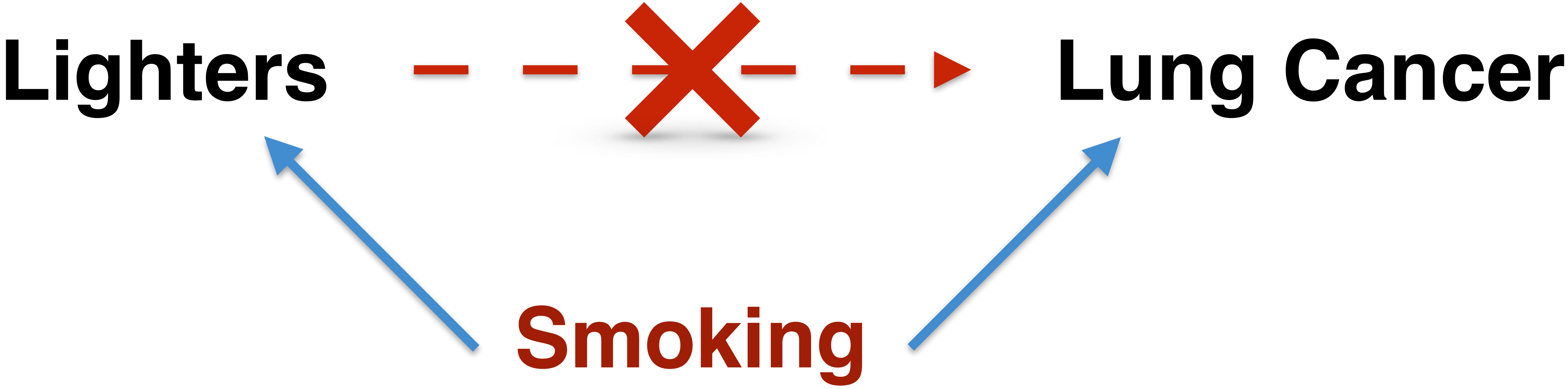
- Does this mean that the lighters are causing cancer?



Lighters → Lung Cancer

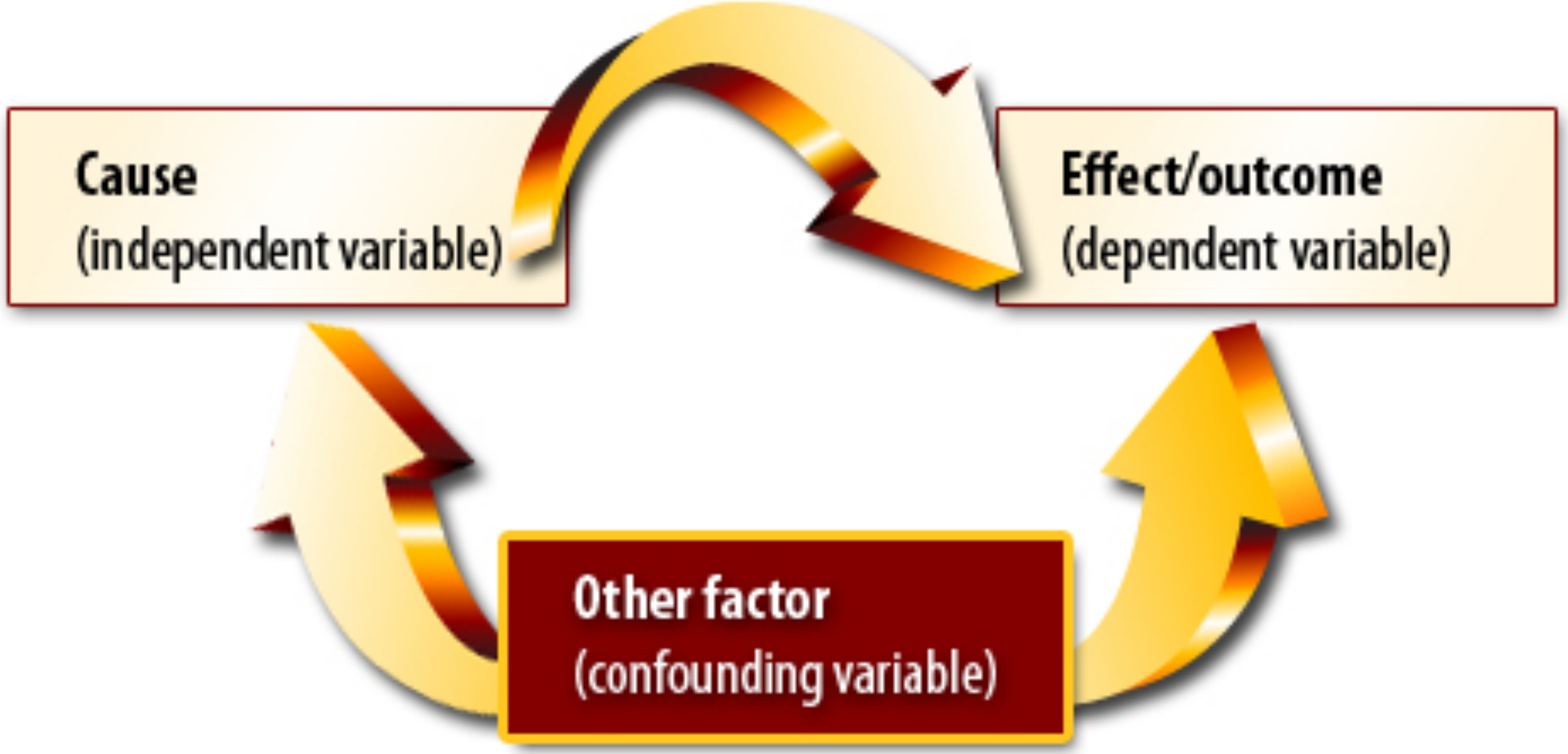
CONFOUNDING

- Does this mean that the lighters are causing cancer?
- No!



CONFOUNDING

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



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (5 MINUTES)

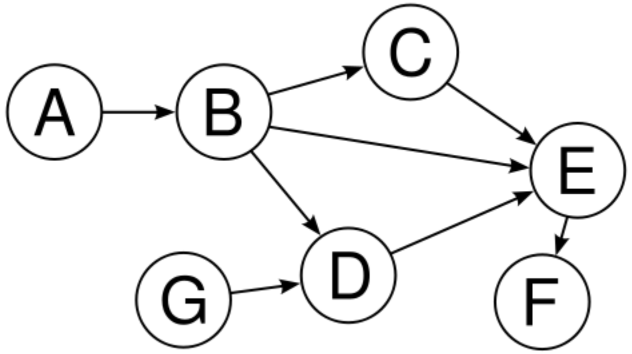
1. What factors are missing from this model?
2. How might we measure for these?

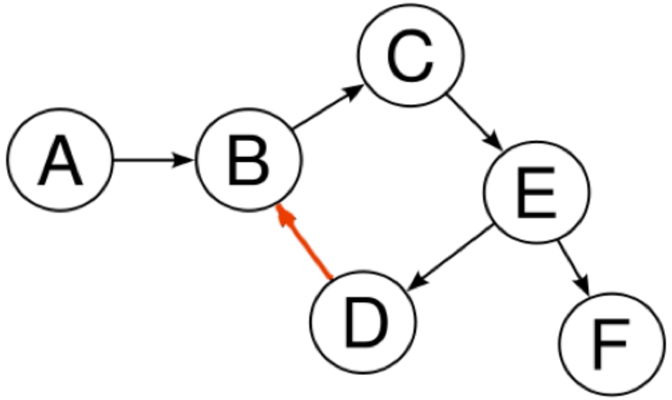
THINK, PAIR, SHARE

DIRECTED ACYCLIC GRAPHS

DIRECTED ACYCLIC GRAPH

- **Graph** is a structure consisting of **nodes**, aka **vertices**, that are connected to each other with **edges**
- **Directed** means that edges have a direction
  - A -> B is NOT the same as B -> A
- **Acyclic** or “non-circular” meaning after passing by a node it will never be visited again by following the edges
- DAGs are associated with problem solving by representing connectivity, causality and dependency



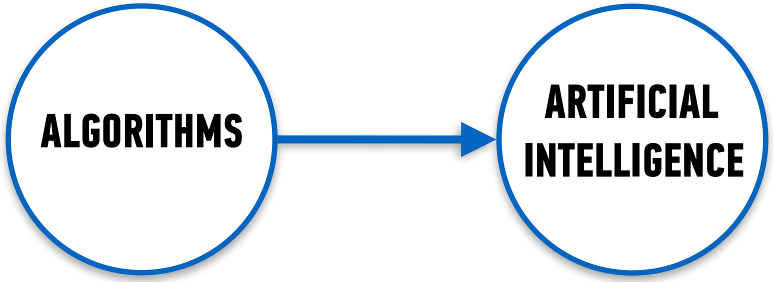


DIRECTED ACYCLIC GRAPH - SOME USES

- Shortest path
- Scheduling
- Data processing networks
- Causal structures
- Genealogy and version history
- Citation graphs
- Data compression

DIRECTED ACYCLIC GRAPH - EXAMPLE: PRE-REQUISITE

- On a course a student faces a task of choosing subjects that follows pre-requisites. One cannot take a class on Artificial Intelligence[B] without a pre-requisite course on Algorithms[A].
- Hence B depends on A or A has an edge directed to B. So in order to reach Node B you have to visit Node A.
- After adding all the subjects with their pre-requisites into a graph, it will turn out to be a Directed Acyclic Graph
- The university only allows students to register for courses that they have taken all pre-requisite courses before

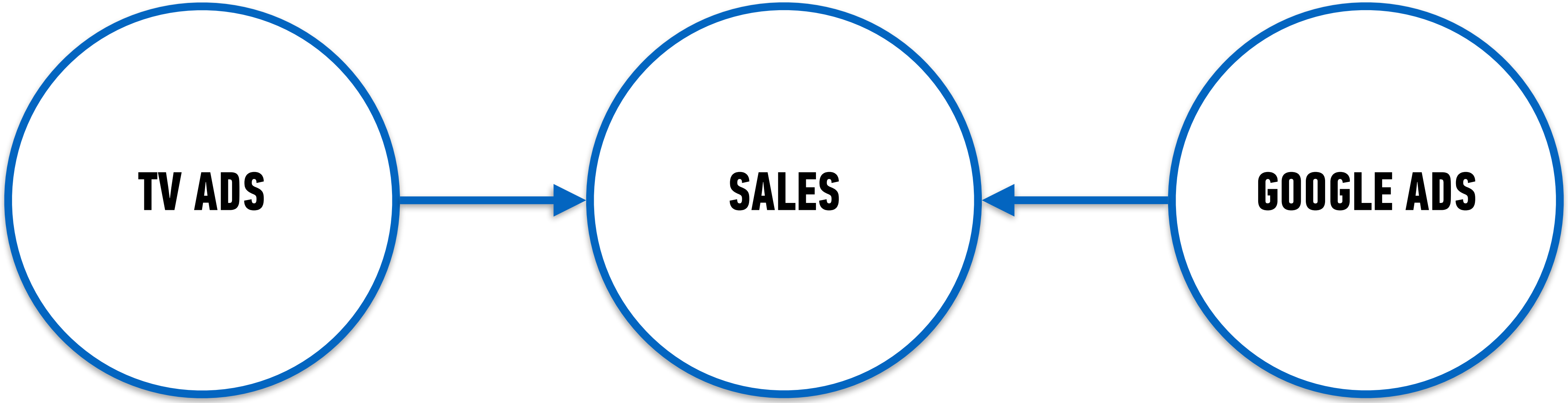


DIRECTED ACYCLIC GRAPH

- The exposure / predictor is TV advertising, 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 advertising?

DIRECTED ACYCLIC GRAPH

- The DAG for this might look like the following



ACTIVITY: DIRECTED ACYCLIC GRAPHS

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (5 MINUTES)

1. Let's say we want to evaluate which type of advertising is associated with higher sales
   a. Break in small groups
   b. Draw a basic DAG on your table or on the board. This DAG should show the relationship between ads and higher sales
   c. Discuss your DAGs in small groups and be ready to share one or two examples with the class

SEASONALITY

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

SEASONALITY

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

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

SEASONALITY

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

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

- This is an example of bias and confounding
  - It is not that TV ads are better than Google ads
  - It is 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 Advertising and Sales while taking into account seasonality (recurring regular patterns over time)

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

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS (10 MINUTES)

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

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 the 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

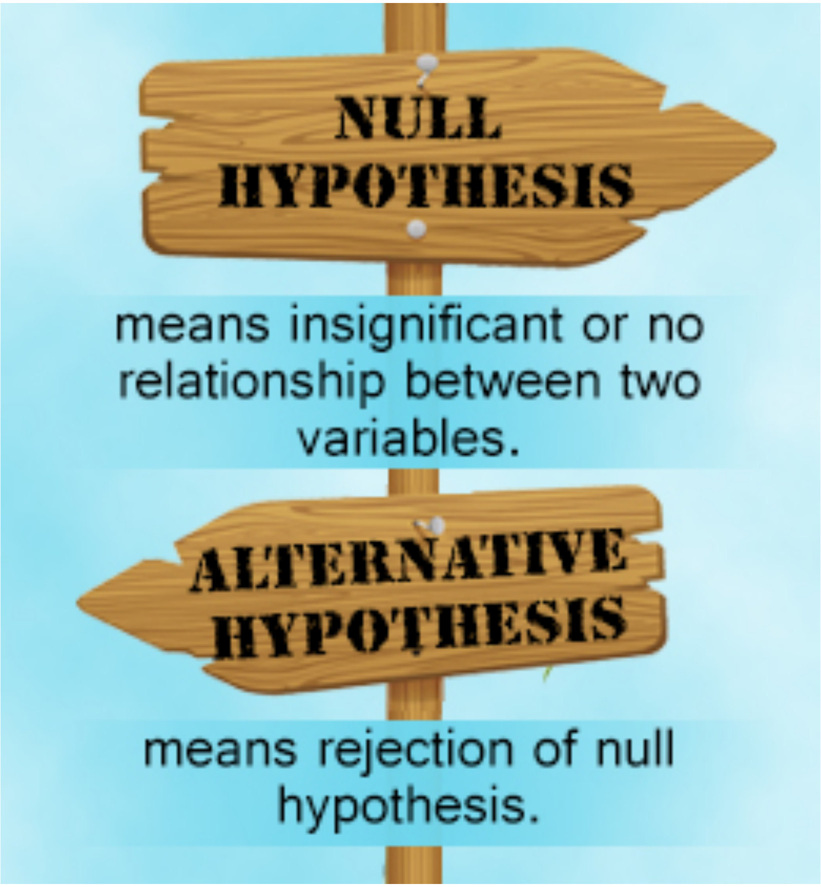
HYPOTHESIS TESTING

- 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
  - Is the **Alternative Hypothesis**


NULL HYPOTHESIS
means insignificant or no relationship between two variables.
ALTERNATIVE HYPOTHESIS
means rejection of null 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 in favour of the Alternative Hypothesis
  - **Failing to Reject** the Null Hypothesis

- **Note**: Failing to reject the Null Hypothesis is **NOT** the same as accepting the Alternative Hypothesis
  - While the Alternative Hypothesis **might** be true, we do not have enough data to support that claim specifically
  - Keep this in mind so you do not overstate your findings

DEMONSTRATION

HYPOTHESIS TESTING CASE STUDY

HYPOTHESIS TESTING CASE STUDY

- We are going to walk through Part 1 of the guided demonstration notebook in the class repository for lesson 4
    - ~/lessons/lesson-04/code/starter/demo-starter-4.ipynb
- There are several questions to answer
    - We will answer those questions in small groups and then discuss with the class

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS (5 MINUTES)

1. What is the Null Hypothesis?
2. Why is this important to use?

INTRODUCTION

VALIDATE YOUR FINDINGS

VALIDATING YOUR FINDINGS

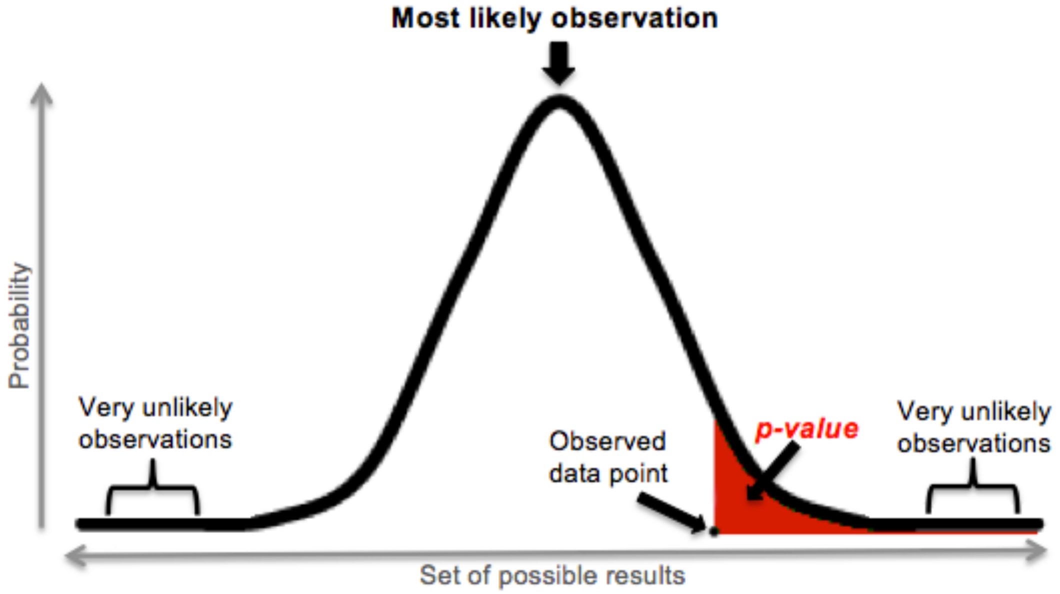
- We know how to carry out a hypothesis test
  - But how do we tell if the association we found is **Statistically Significant**?

- **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

VALIDATING 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

VALIDATING YOUR FINDINGS

- 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 (in favour of the Alternative Hypothesis)	$p < 0.05$
• Not Statistically Significant • Due to chance	The Null Hypothesis could not be rejected	$p > 0.05$

VALIDATING 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

DEMONSTRATION

CONFIDENCE INTERVALS AND P-VALUES CASE STUDY

CONFIDENCE INTERVALS AND P-VALUES CASE STUDY

- We are now going to walk through Part 2 of the guided demonstration notebook in the class repository for lesson 4
  - ~/lessons/lesson-04/code/starter/demo-starter-4.ipynb

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

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS (5 MINUTES)

1. What does a 95% confidence interval indicate?

INDEPENDENT PRACTICE

INTERPRETING RESULTS

ACTIVITY: INTERPRETING RESULTS

EXERCISE

DIRECTIONS (35 MINUTES)

1. Using the laboratory code, you will look through a variety of analyses and interpret the findings
    a. ~/lessons/lesson-04/code/starter/lab-starter-4.ipynb
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

CONCLUSION

LAB REVIEW

LAB REVIEW

- Let’s review the answers to the questions in the labs

- Any other questions?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project
  - Unit Project 2

Q & A

STATISTICS FUNDAMENTALS II

EXIT TICKETS

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

Exit Ticket Link

What's the lesson number?	04
What was the topic of the lesson?	Statistics Fundamentals II

CREDITS AND REFERENCES

STATISTICS FUNDAMENTALS II

- How juries are fooled by statistics
  - Peter Donnelly
    Mathematician, Statistician
    - Australian-born, Oxford-based mathematician, best known for his work in molecular evolution
      - [Website]
      - [TED Talk]