GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS, PART 2

Bethany Baker

DS-SF-33//March 22, 2017

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)

COURSE

PRE-WORK

PRE-WORK REVIEW

- Explain the difference between variance and bias
- Use descriptive statistics to understand your data

WHY PART 1 & 2?

- STATISTICS: “the collection, presentation, analysis, and utilization of numerical data to make *inferences* and reach *decisions* in the face of *uncertainty* or *incomplete information*
  a. DESCRIPTIVE STATISTICS (PART 1) - summarize and describe a body of data
  b. INFERENTIAL STATISTICS (PART 2) - process of reaching generalizations about a whole (the *population*) by examining a portion (the *sample*)

BIAS-VARIANCE TRADEOFF

- MOTIVATION: for prediction models, prediction error can be decomposed into two main subcomponents: error due to “bias” and error due to “variance”
- There is a tradeoff between a model’s ability to minimize bias + variance
- Understanding these concepts can help us diagnose model results and avoid over- or under-fitting

BIAS-VARIANCE TRADEOFF

- BIAS: The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict
- VARIANCE: The error due to variance is taken as the variability of a model prediction for a given data point
- EXAMPLE: Election polling

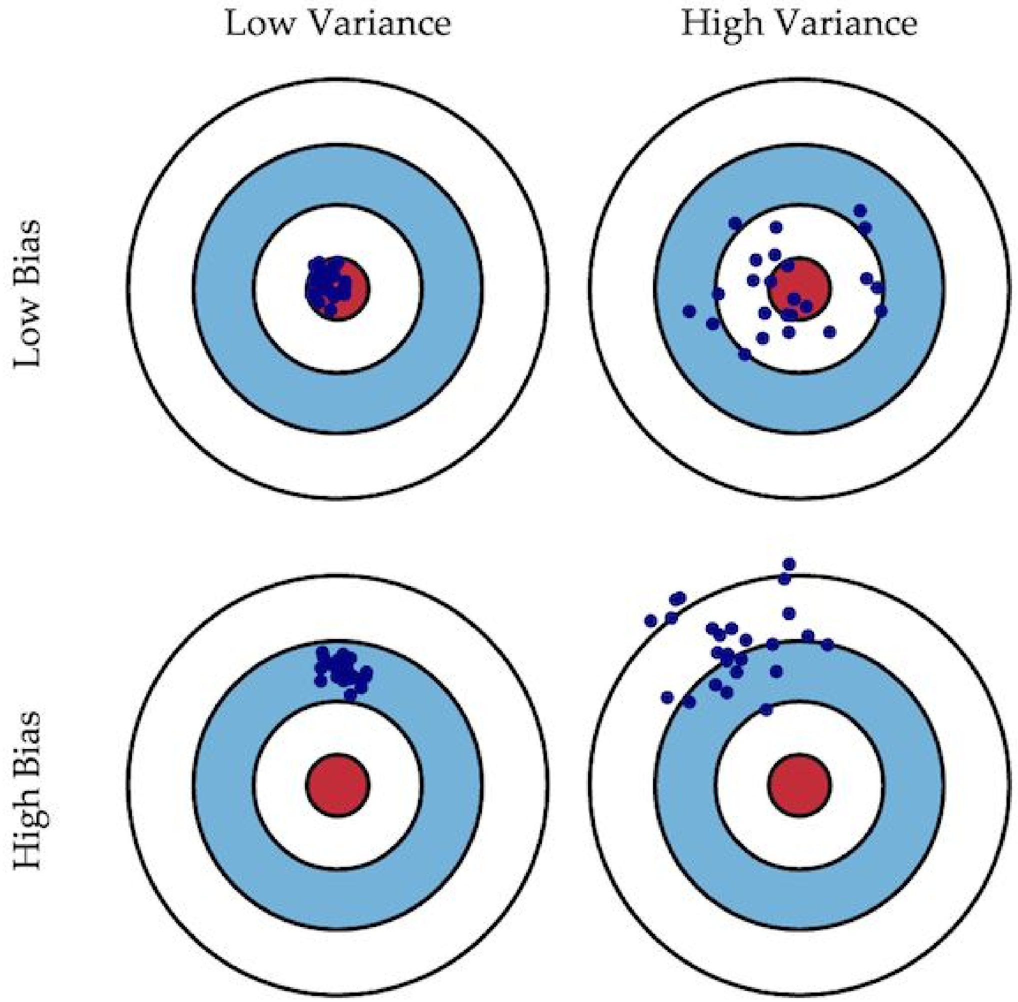


Fig. 1 Graphical illustration of bias and variance.

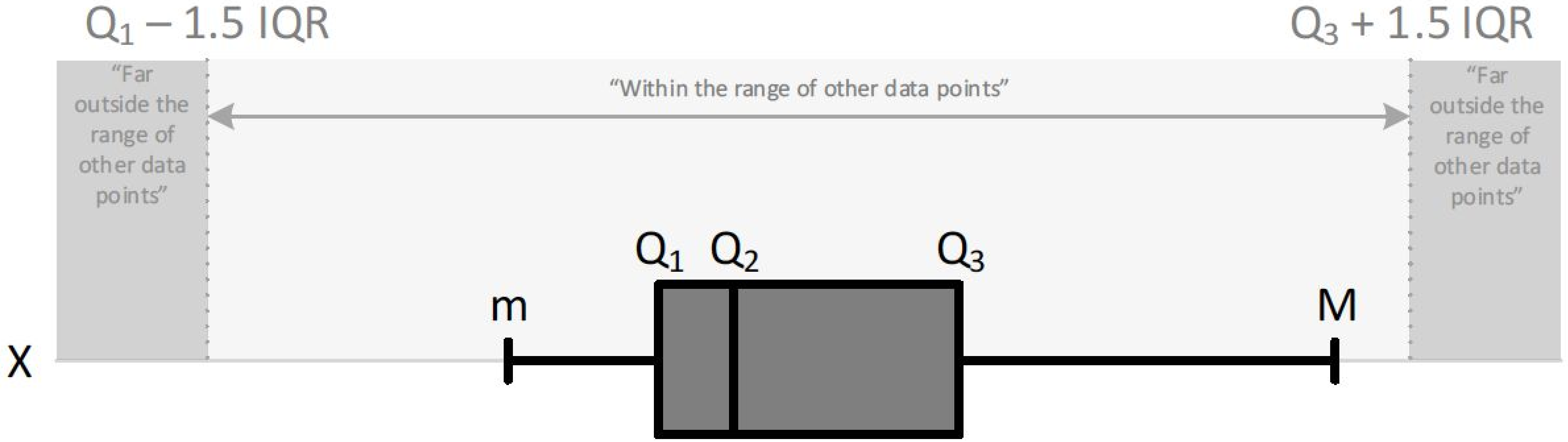
BIAS-VARIANCE TRADEOFF

- EXAMPLE: Say we want to create a model that will predict the percentage of people who will vote Republican in the next presidential election
- We can build a model by selecting 50 people at random in the phone book and polling their answer
- What are sources of bias? variance?

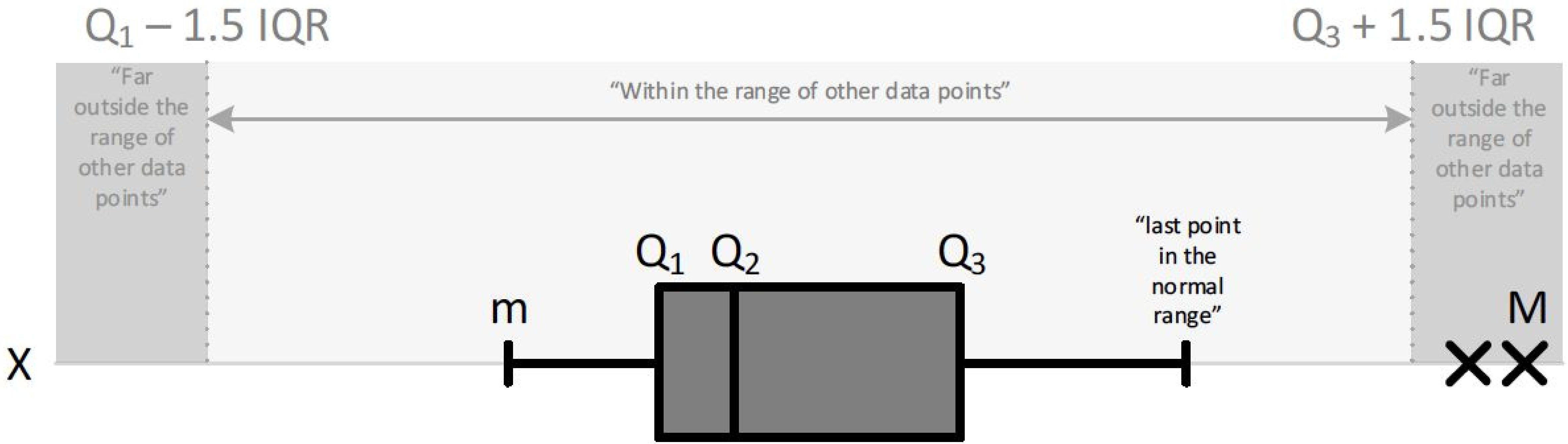
Voting Republican	Voting Democrat	Non-Respondent	Total
13	16	21	50
44.8%	55.2%		

http://scott.fortmann-roe.com/docs/BiasVariance.html

BOXPLOTS - NO OUTLIERS


$Q_1 - 1.5$ IQR
$Q_3 + 1.5$ IQR
"Far outside the range of other data points"
"Within the range of other data points"
"Far outside the range of other data points"
Q_1
Q_2
Q_3
m
M
X

BOXPLOTS - WITH OUTLIERS


$Q_1 – 1.5$ IQR
$Q_3 + 1.5$ IQR
“Far outside the range of other data points”
“Within the range of other data points”
“Far outside the range of other data points”
“last point in the normal range”
Q_1
Q_2
Q_3
m
M
X

OPENING

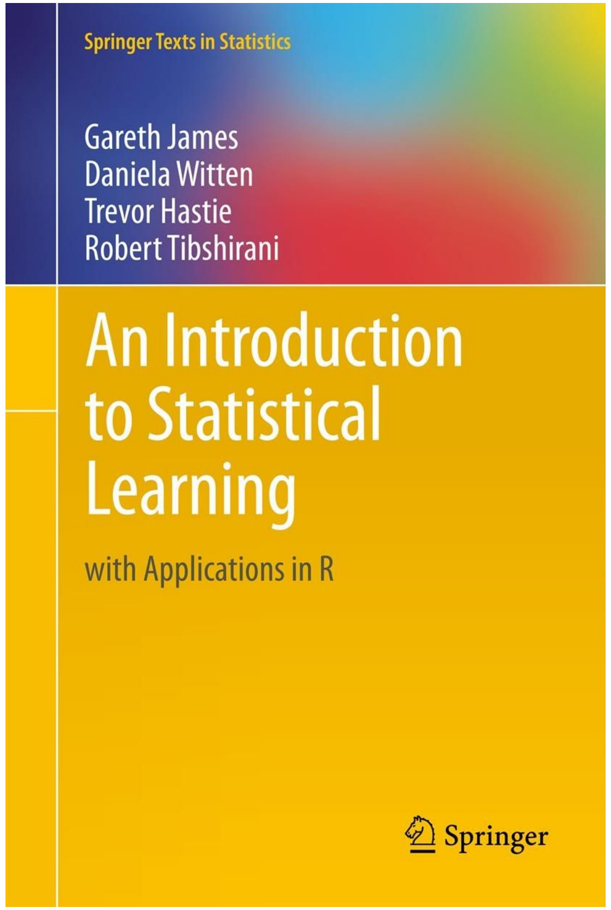
STATISTICS FUNDAMENTALS, PART 2

LAST SESSION

- Any questions from last class?
- Exit tickets

DATA SOURCE

- Today, we’ll use advertising data from an example in *An Introduction to Statistical Learning*.



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

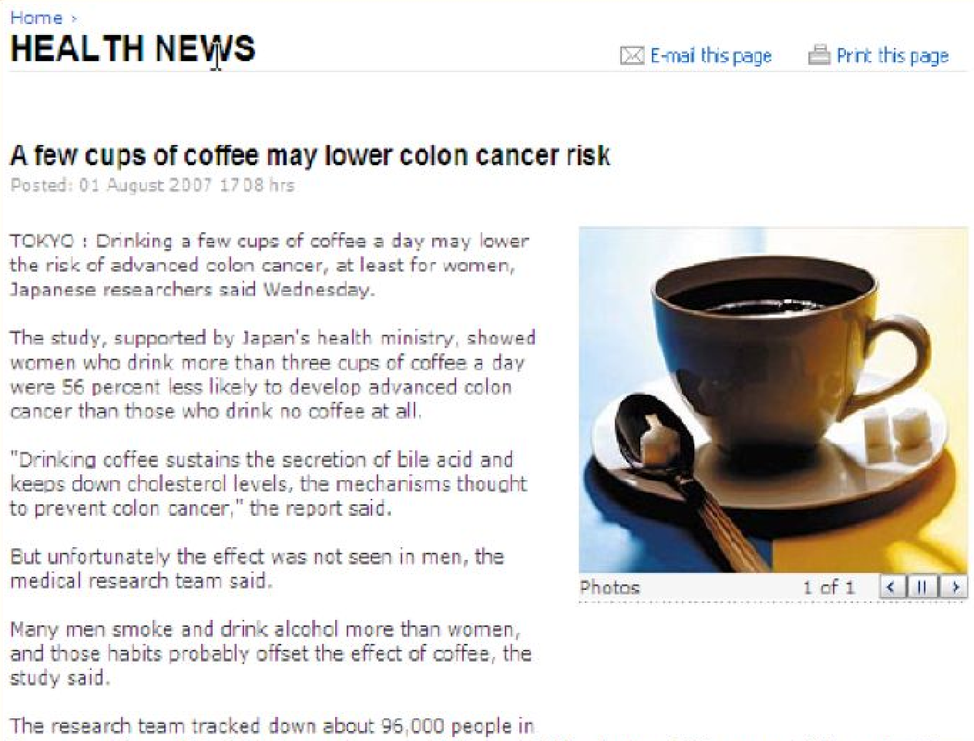
INTRODUCTION

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.



HEALTH NEWS

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.

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.

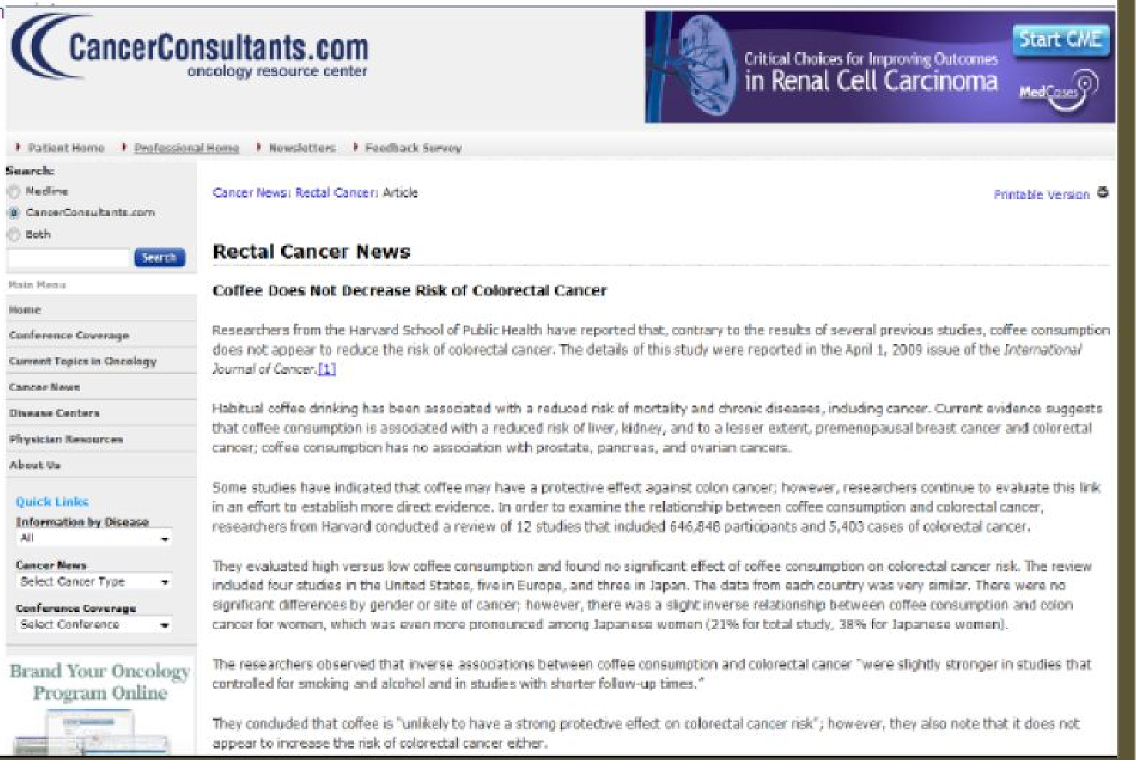
But unfortunately the effect was not seen in men, the medical research team said.

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!



Cancer News: Rectal Cancer: Article

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]

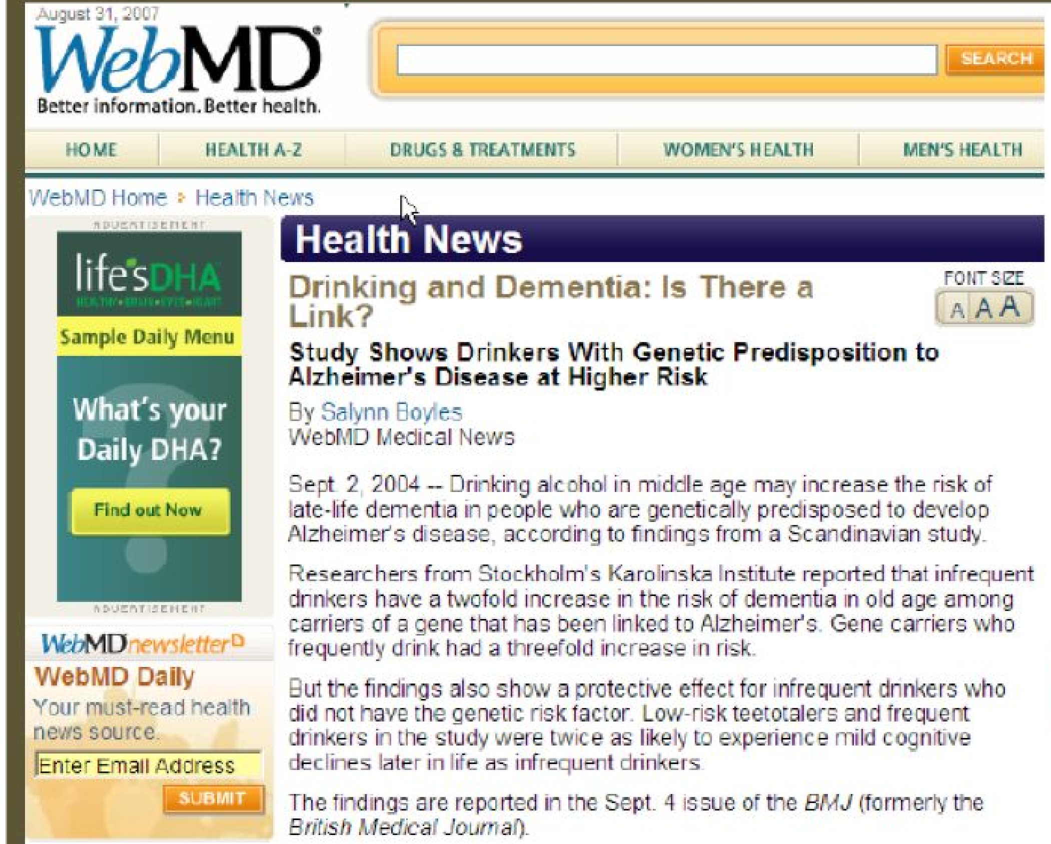
Habitual coffee drinking has been associated with a reduced risk of mortality and chronic diseases, including cancer. Current evidence suggests that coffee consumption is associated with a reduced risk of liver, kidney, and to a lesser extent, premenopausal breast cancer and colorectal cancer; coffee 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.



Health News

Drinking and Dementia: Is There a Link?

Study Shows Drinkers With Genetic Predisposition to Alzheimer's Disease at Higher Risk

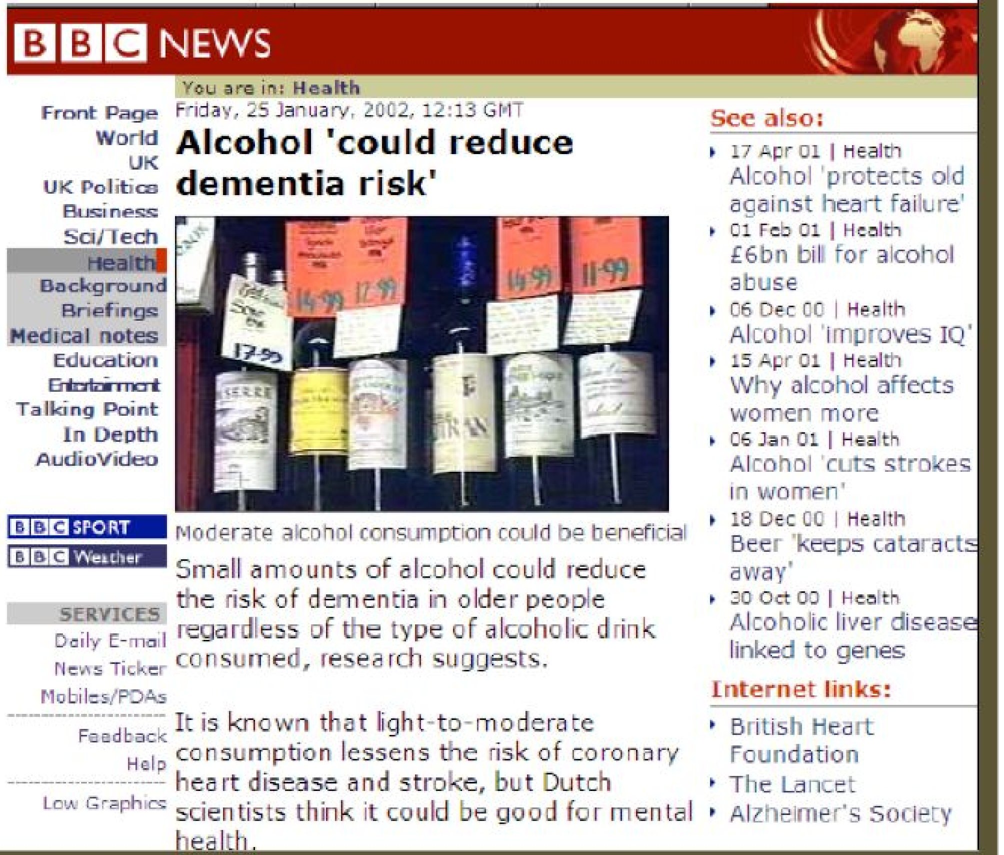
By Salynn Boyles
WebMD Medical News

Sept. 2, 2004 -- Drinking alcohol in middle age may increase the risk of late-life dementia in people who are genetically predisposed to develop Alzheimer's disease, according to findings from a Scandinavian study.

Researchers from Stockholm's Karolinska Institute reported that infrequent drinkers have a twofold increase in the risk of dementia in old age among carriers of a gene that has been linked to Alzheimer's. Gene carriers who frequently drink had a threefold increase in risk.

But the findings also show a protective effect for infrequent drinkers who did not have the genetic risk factor. Low-risk teetotalers and frequent drinkers in the study were twice as likely to experience mild cognitive declines later in life as infrequent drinkers.

The findings are reported in the Sept. 4 issue of the *BMJ* (formerly the *British Medical Journal*).



Friday, 25 January, 2002, 12:13 GMT

Alcohol 'could reduce dementia risk'

Moderate alcohol consumption could be beneficial

Small amounts of alcohol could reduce the risk of dementia in older people regardless of the type of alcoholic drink consumed, research suggests.

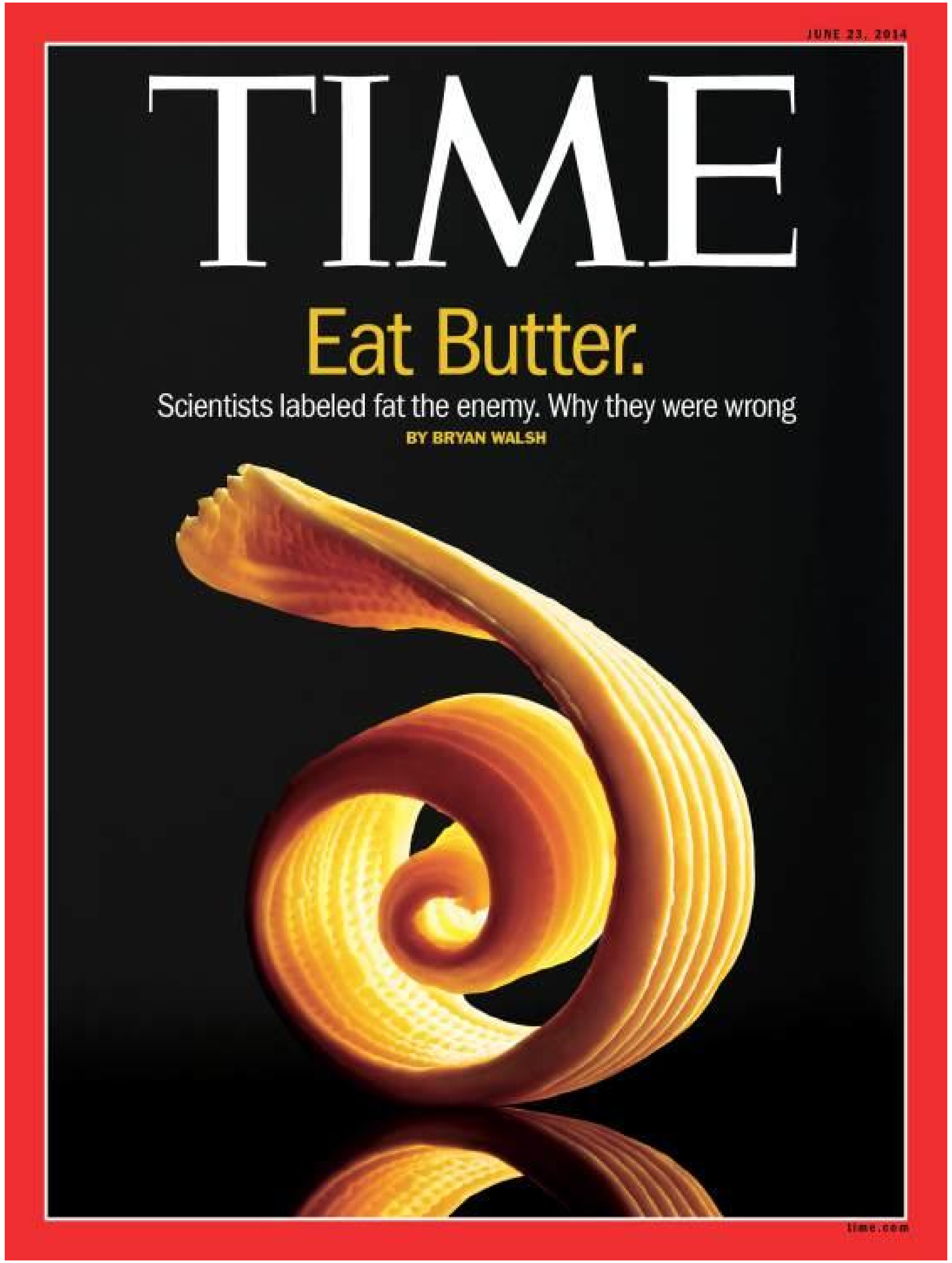
It is known that light-to-moderate consumption lessens the risk of coronary heart disease and stroke, but Dutch scientists think it could be good for mental health.

See also:

- 17 Apr 01 | Health
  Alcohol 'protects old against heart failure'
- 01 Feb 01 | Health
  £6bn bill for alcohol abuse
- 06 Dec 00 | Health
  Alcohol 'improves IQ'
- 15 Apr 01 | Health
  Why alcohol affects women more
- 06 Jan 01 | Health
  Alcohol 'cuts strokes in women'
- 18 Dec 00 | Health
  Beer 'keeps cataracts away'
- 30 Oct 00 | Health
  Alcoholic liver disease linked to genes

Internet links:

- British Heart Foundation
- The Lancet
- Alzheimer's Society



JUNE 23, 2014

TIME

Eat Butter.

Scientists labeled fat the enemy. Why they were wrong

BY BRYAN WALSH

time.com

CAUSATION AND CORRELATION

- Why is this?
- Sensational headlines?

CAUSATION AND CORRELATION

- 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

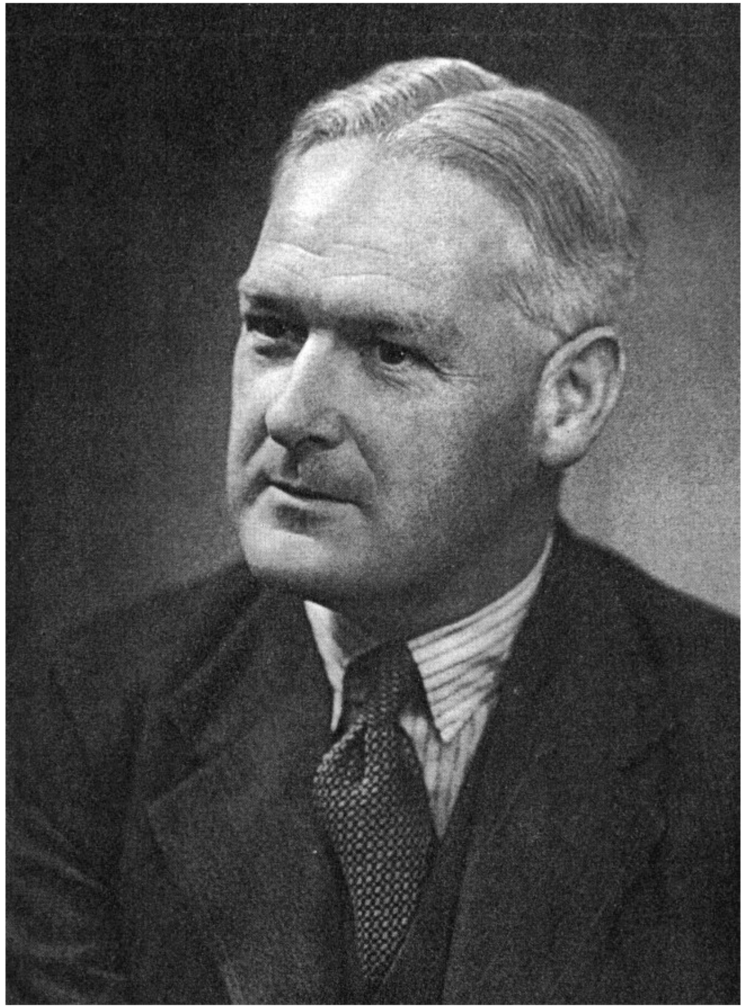
CAUSAL CRITERIA

- 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. Strength of association
  b. Consistency
  c. Specificity
  d. Temporality
  e. Biological gradient
  f. Plausibility
  g. Coherence
  h. Experiment
  i. Analogy



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.

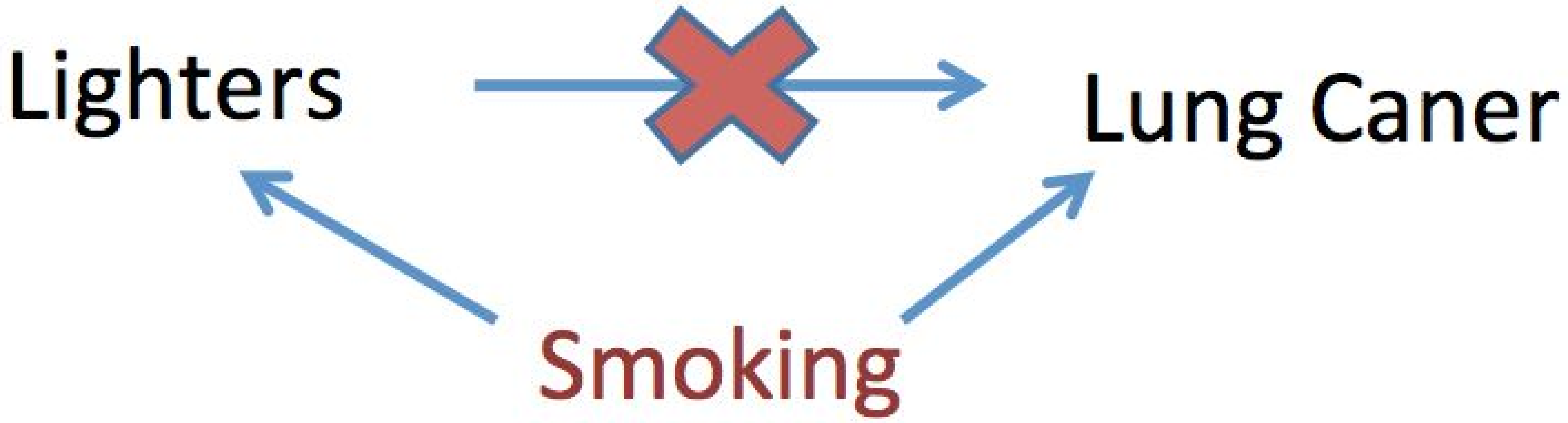
- Let’s say we did an analysis to understand what causes lung cancer.

- 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?

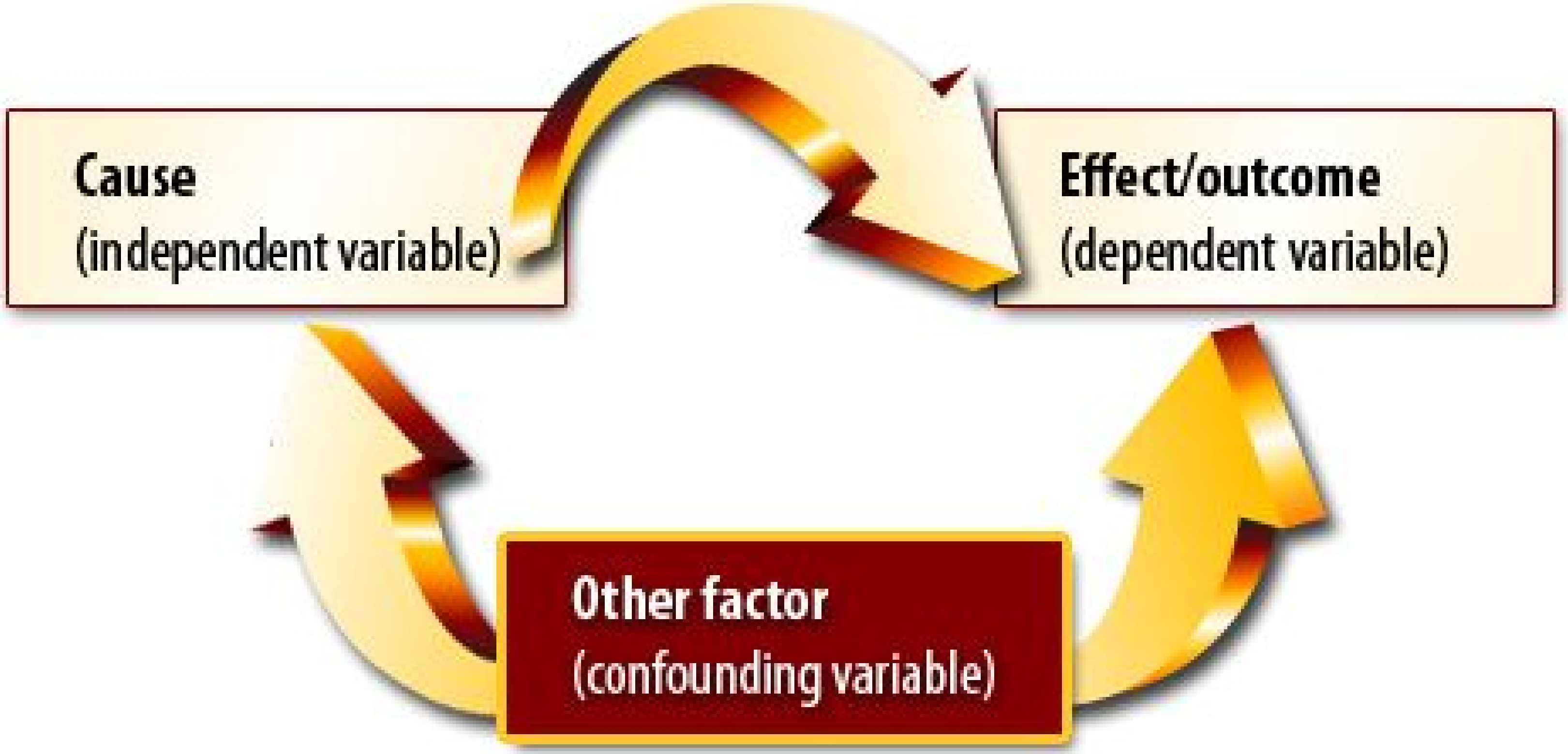
CONFOUNDING

- No!



CONFOUNDING

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



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

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

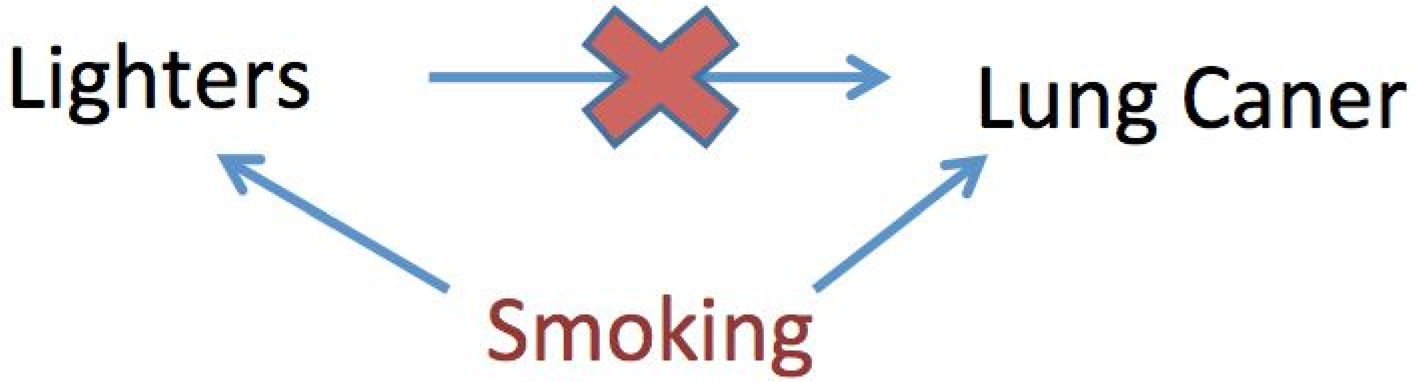
DELIVERABLE

Answers to the above questions

DIRECTED ACYCLIC GRAPH

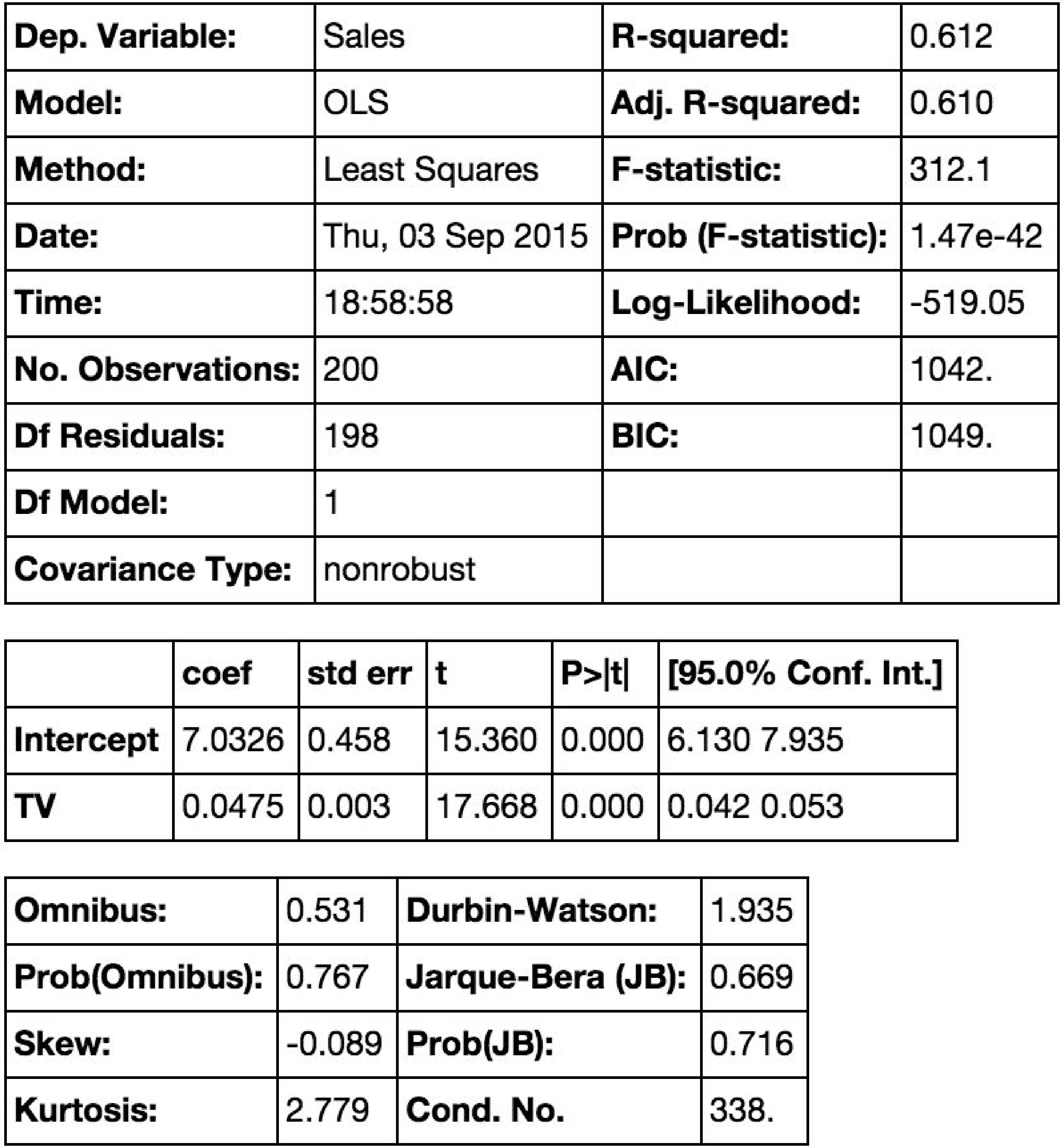
- A *Directed Acyclic Graph* (DAG) can help determine which variables are most important for your model. It helps visually demonstrate the logic of your models.

- 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

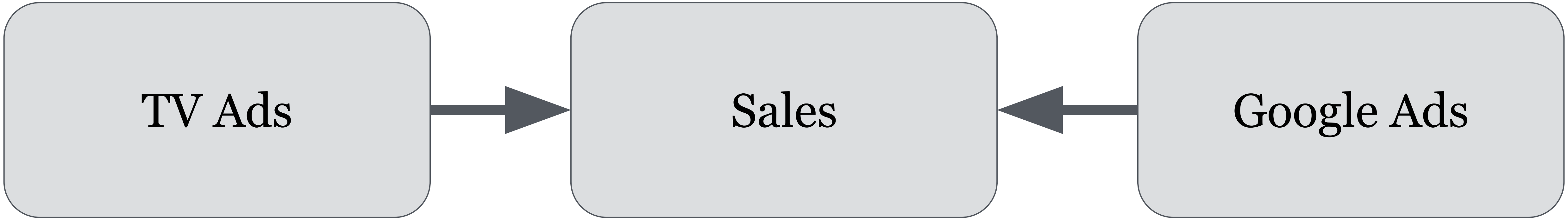
Omnibus:	0.531	**Durbin-Watson:**	1.935
Prob(Omnibus):	0.767	**Jarque-Bera (JB):**	0.669
Skew:	-0.089	**Prob(JB):**	0.716
Kurtosis:	2.779	**Cond. No.**	338.

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

ACTIVITY: DAGS

EXERCISE

DIRECTIONS

Let’s say we want to evaluate which type of ad is associated with higher sales.

1. Break small groups.
2. Draw a basic DAG on your table or on the board. This DAG should show the relationship between ads and higher sales.
3. Discuss your DAGs in small groups and be ready to share one or two examples with the class.

DELIVERABLE

Insert Deliverable

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.

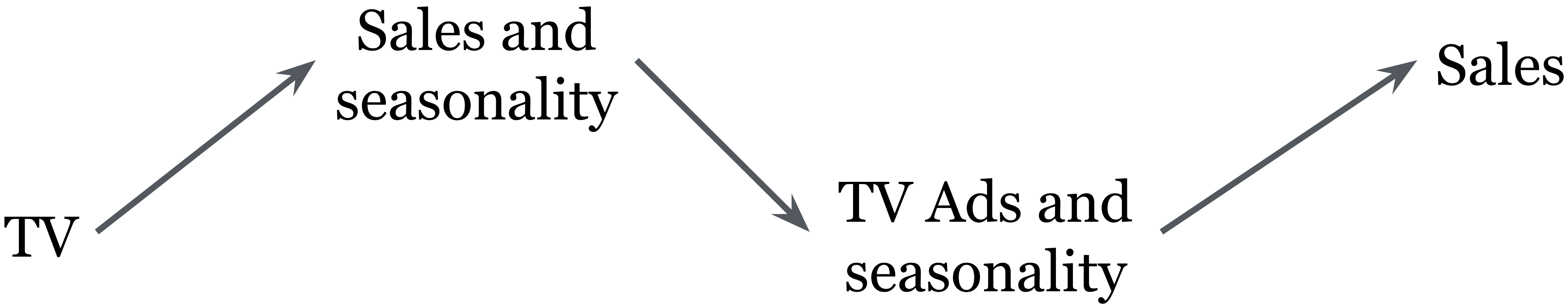
SEASONALITY

- 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)?

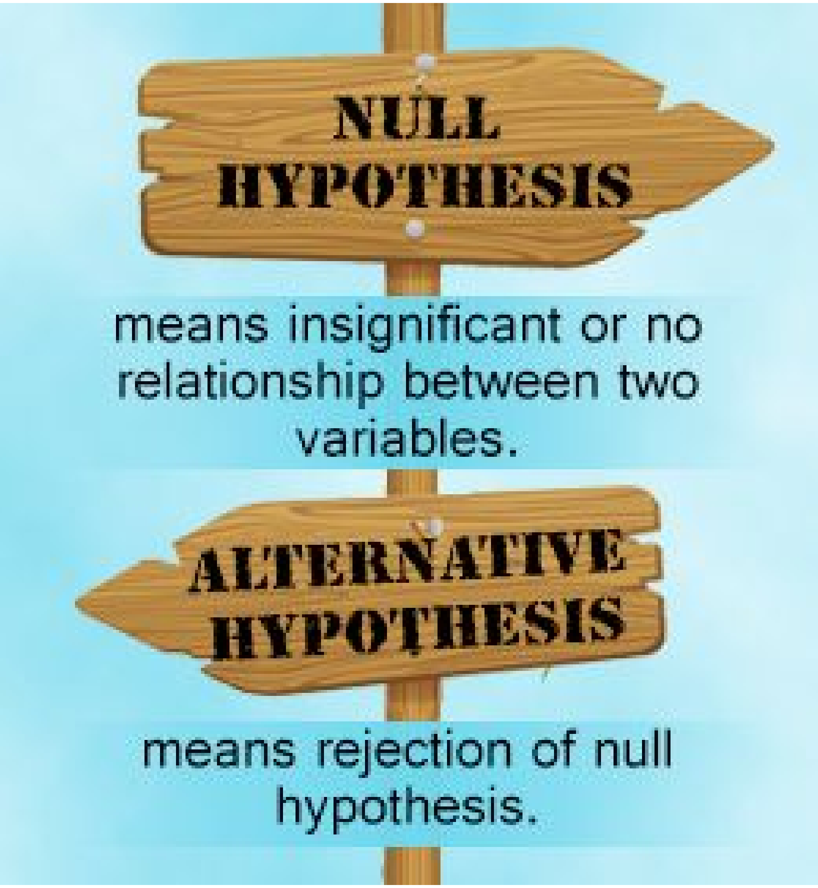
- Imagine we are testing the health of smokers vs. non-smokers. At a cursory glance, our results may show that smokers are marginally healthier than non-smokers.

- 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*.


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 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

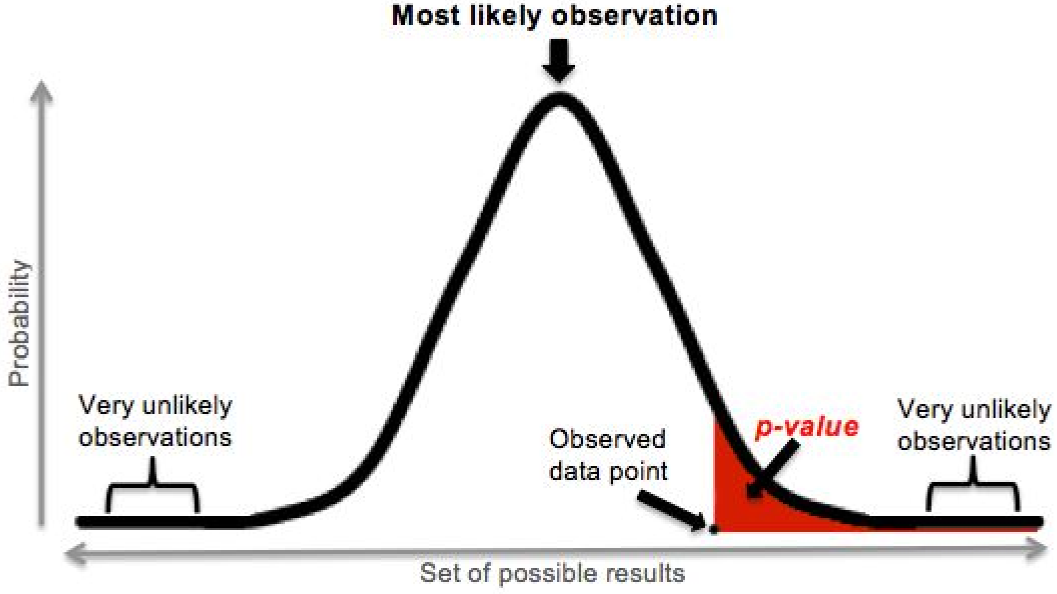
- We know how to carry out a hypothesis test, but how do we tell if the association we found is *statistically significant*?

- *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

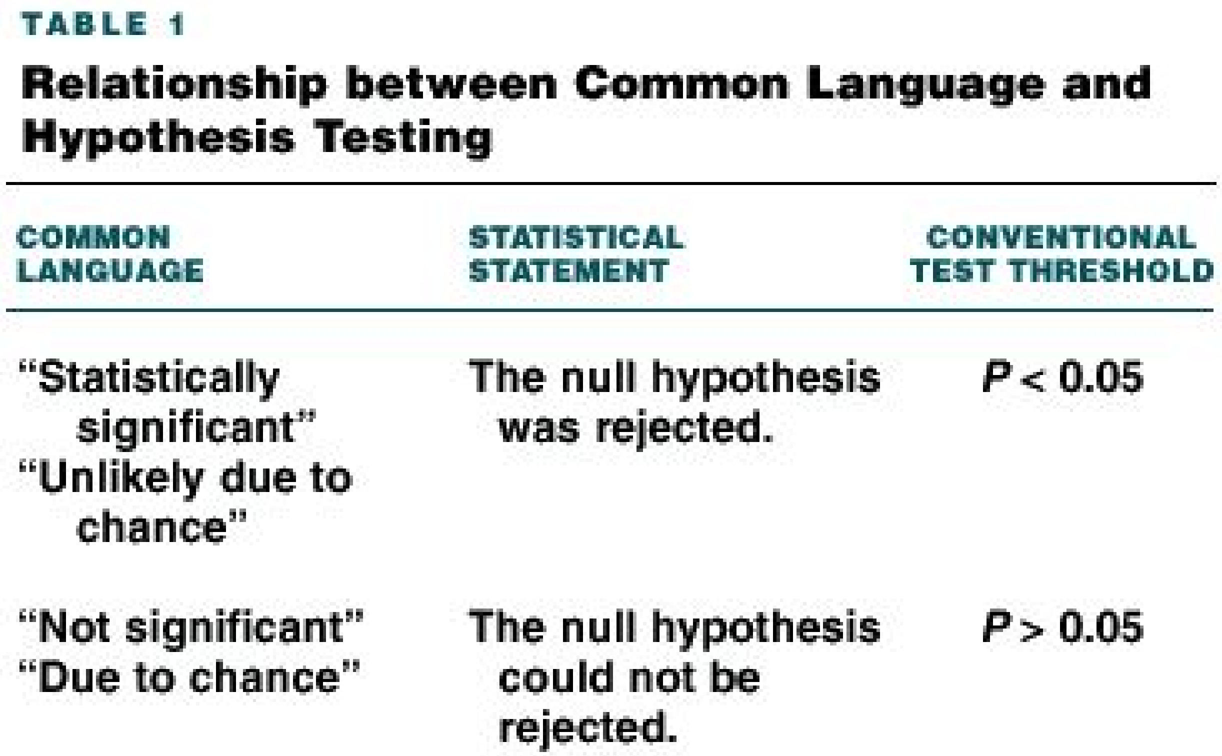


TABLE 1

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

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