GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS, PART 2

Chirayu Sariya

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

- Use descriptive statistics to understand your data
- Explain multicollinearity

OPENING

STATISTICS FUNDAMENTALS, PART 2

LAST SESSION

- Any questions from last class?

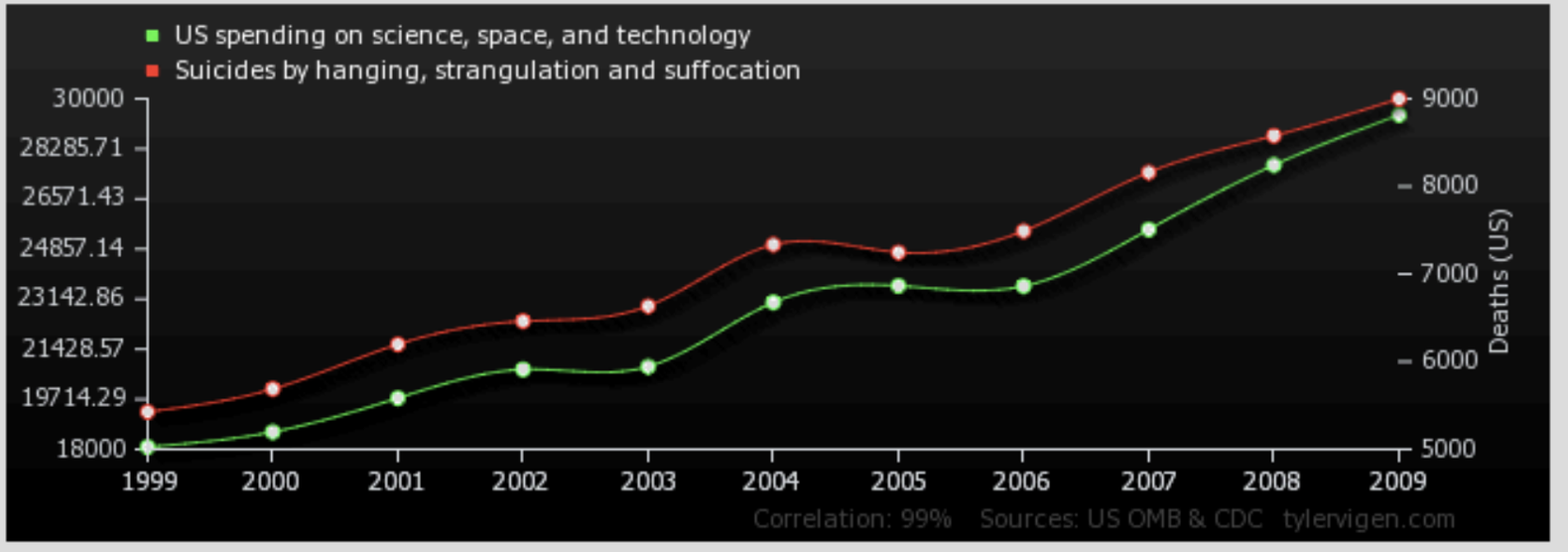
CAUSATION AND CORRELATION

CAUSATION AND CORRELATION

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

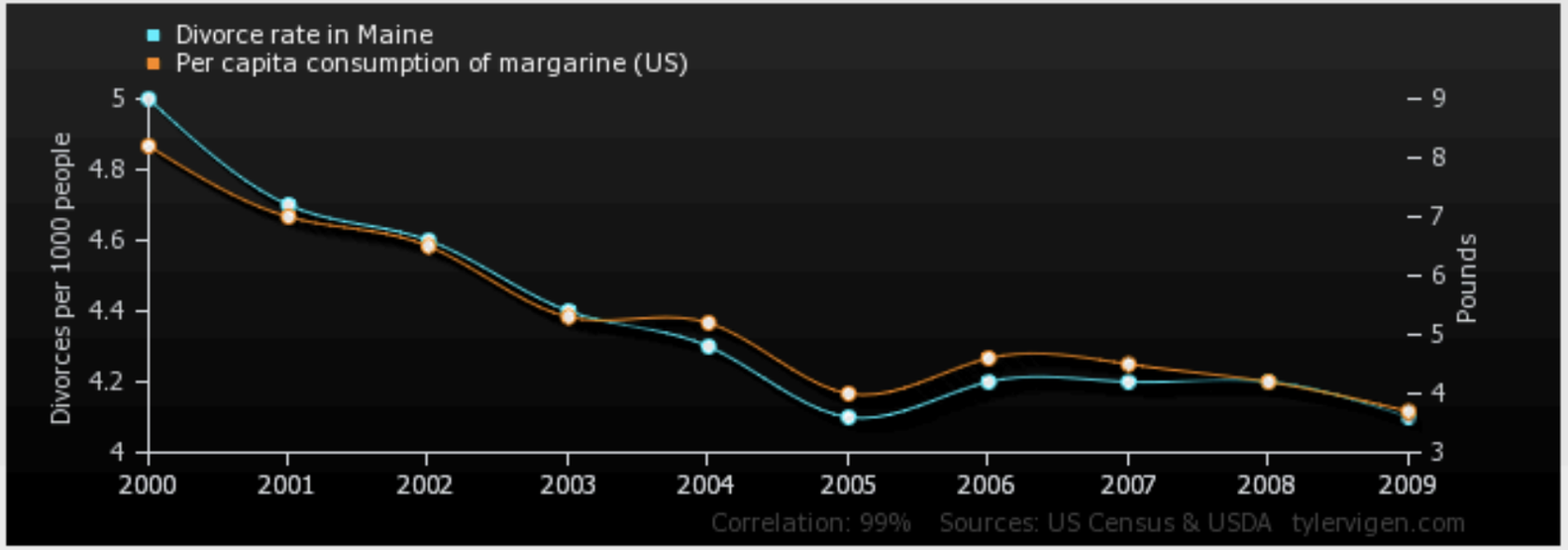
Correlation: 0.992082

CAUSATION AND CORRELATION

Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

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.

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.

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.

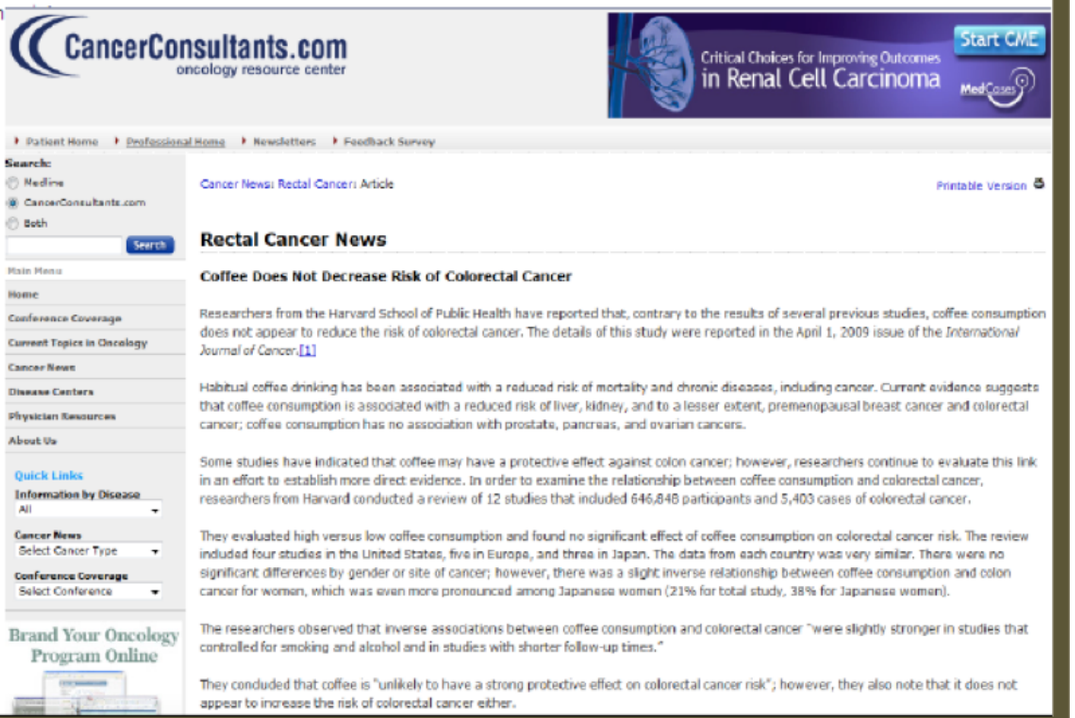
But unfortunately the effect was not seen in men, the medical research team said.

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.

Photos 1 of 1

Causal claims are often
inconsistent
and contradictory!



August 31, 2007

WebMD
Better information. Better health.

SEARCH

HOME | HEALTH A-Z | DRUGS & TREATMENTS | WOMEN'S HEALTH | MEN'S HEALTH

WebMD Home › Health News

WebMD newsletter
WebMD Daily
Your must-read health news source.
Enter Email Address
SUBMIT

Health News

Drinking and Dementia: Is There a Link?

FONT SIZE A A A

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*).

BBC NEWS

You are in: Health

Front Page
World
UK
UK Politics
Business
Sci/Tech
Health
Background
Briefings
Medical notes
Education
Entertainment
Talking Point
In Depth
AudioVideo

BBC SPORT
BBC Weather

SERVICES
Daily E-mail
News Ticker
Mobiles/PDAs

Feedback
Help

Low Graphics

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

CAUSATION AND CORRELATION

- Why is this?

CAUSATION AND CORRELATION

- Sensational headlines?
- There is neglect of a robust data analysis.
- There is also often a lack of understanding of the difference between *causation* and *correlation*.

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

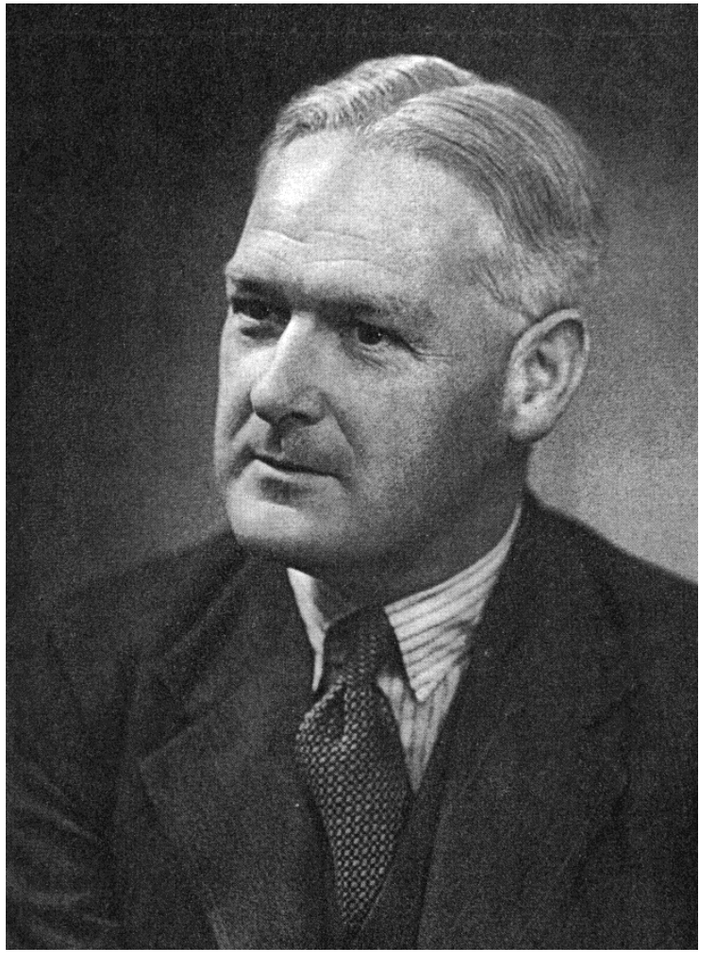
CAUSAL CRITERIA

- 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**.

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.
- We also might not understand *other* factors influencing the association.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

In pairs, think of a simple example where two variables -

1. may be correlated, and have a plausible/coherent indication of a causal relationship.
2. may be correlated, but do not have a plausible causal relationship.

DELIVERABLE

Answers to the above questions

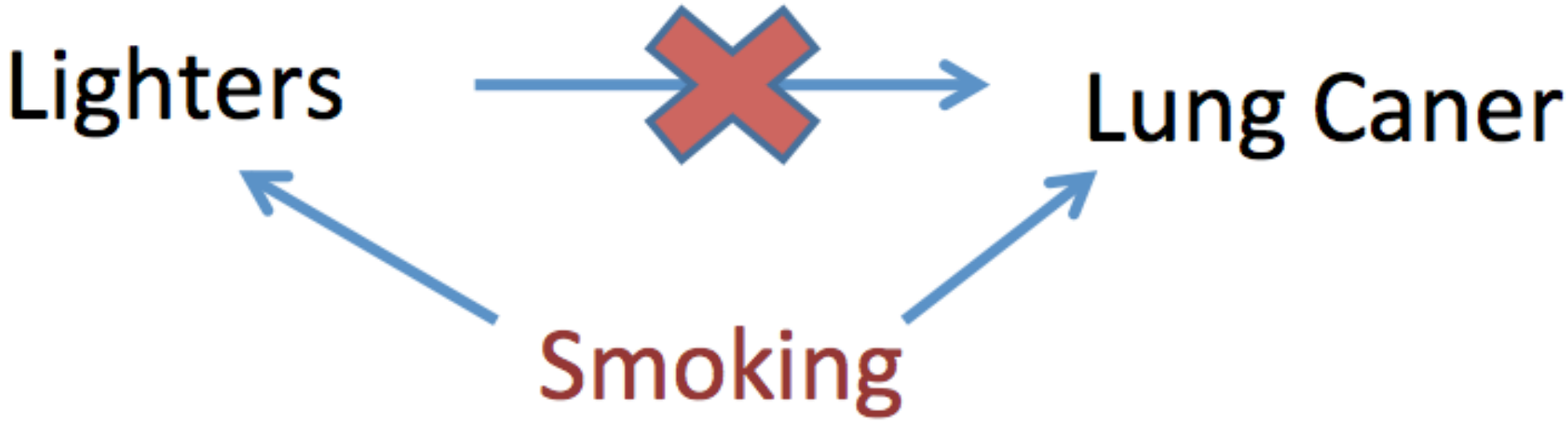
CONFOUNDING VARIABLES

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!



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?

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.
- 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.

HYPOTHESIS TESTING

HYPOTHESIS TESTING

- How can we tell the difference between two groups of observations (e.g. smokers vs. non-smokers)?

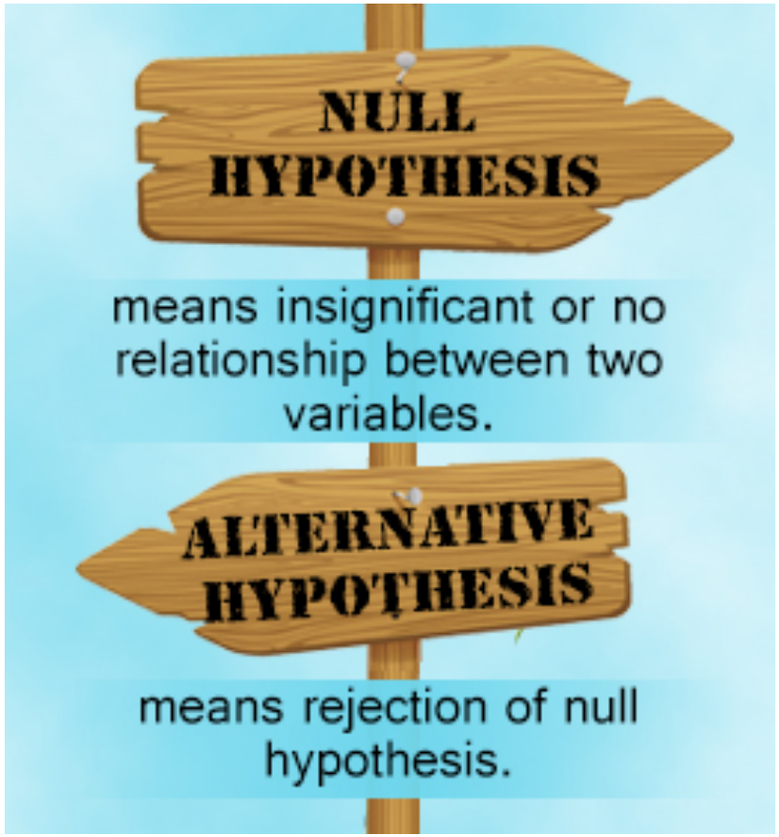
- Imagine we are testing the health of smokers vs. non-smokers. At a cursory glance, our results may show that smokers are marginally healthier than non-smokers.

- 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.

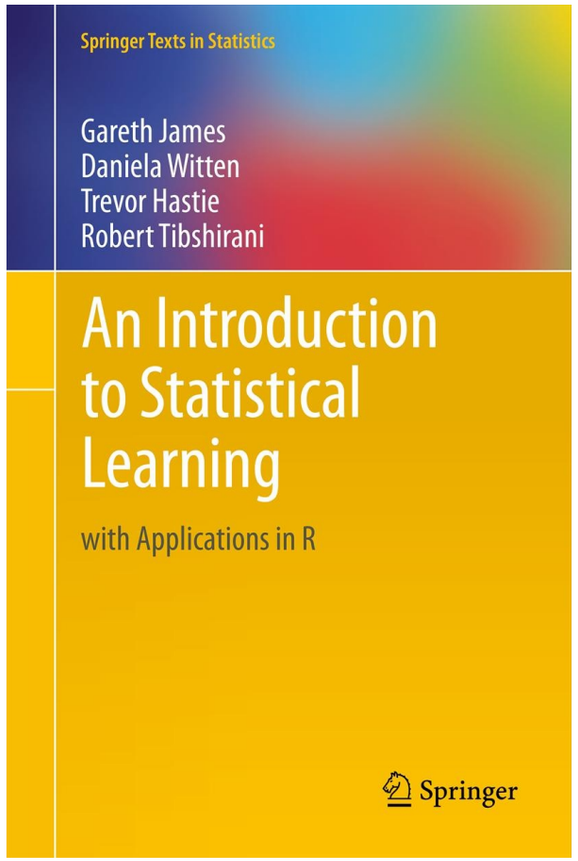
- Keep this in mind so you don't overstate your findings.

DEMO

HYPOTHESIS TESTING CASE STUDY

DATA SOURCE

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



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

ANSWER THE FOLLOWING QUESTIONS

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

DELIVERABLE

Answers to the above questions

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.

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

ANSWER THE FOLLOWING QUESTIONS

1. What does a p-value 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

LAB REVIEW

LAB REVIEW

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

BEFORE NEXT CLASS

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET