

# STATISTICS FUNDAMENTALS, PART 2

*Abbas Chokor, Ph.D.*

*Staff Data Scientist, Seagate Technology*

---

# OUR PROGRESS SO FAR

---

---

## UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

<del>What is Data Science</del>	<del>Lesson 1</del>
<del>Research Design and Pandas</del>	<del>Lesson 2</del>
<del>Statistics Fundamentals I</del>	<del>Lesson 3</del>
Statistics Fundamentals II	Lesson 4
Flexible Class Session	Lesson 5

---

## UNIT 2: FOUNDATIONS OF DATA MODELING

Introduction to Regression	Lesson 6
Evaluating Model Fit	Lesson 7
Introduction to Classification	Lesson 8
Introduction to Logistic Regression	Lesson 9
Communicating Logistic Regression Results	Lesson 10
Flexible Class Session	Lesson 11

---

## UNIT 3: DATA SCIENCE IN THE REAL WORLD

Decision Trees and Random Forests	Lesson 12
Natural Language Processing	Lesson 13
Dimensionality Reduction	Lesson 14
Time Series Data I	Lesson 15
Time Series Data II	Lesson 16
Database Technologies	Lesson 17
Where to Go Next	Lesson 18
Flexible Class Session	Lesson 19
Final Project Presentations	Lesson 20



---

## LAST CLASS

---

# WHAT DID WE LEARN?

- Review basic pandas functions using lab2, a new dataset, and unit project
- Use NumPy and Pandas libraries to analyze datasets using basic summary statistics
- Create basic data visualizations to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization
- ID variable types and complete dummy coding by hand

---

## LAST CLASS


---

# ANNOUNCEMENTS


- ❖ Happy Hour
- ❖ Fill your exit ticket!



More about  
project 1



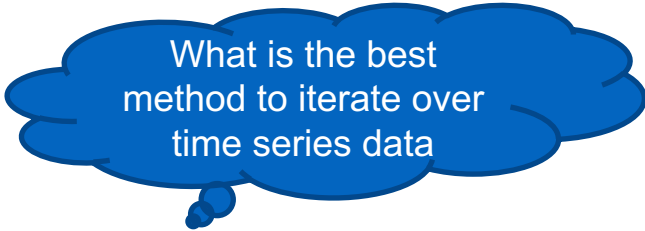
are our keyfobs  
supposed to help  
with parking?



just need  
to practice



Others?



What is the best  
method to iterate over  
time series data



Best Places for  
Project Data

---

## AFTER LAST CLASS

---

# Do you know how to do the following?

- Compute basic statistics, such as: mean, std, var, max, min, etc.
- Difference between bias and variance
- Distinguish between skewness and kurtosis
- Plot histogram, boxplot, etc.



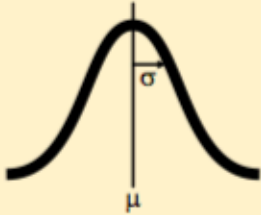

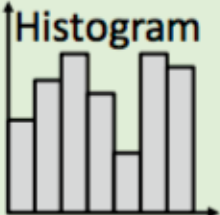
**You got all 4 objectives?**



**Not all of them...**

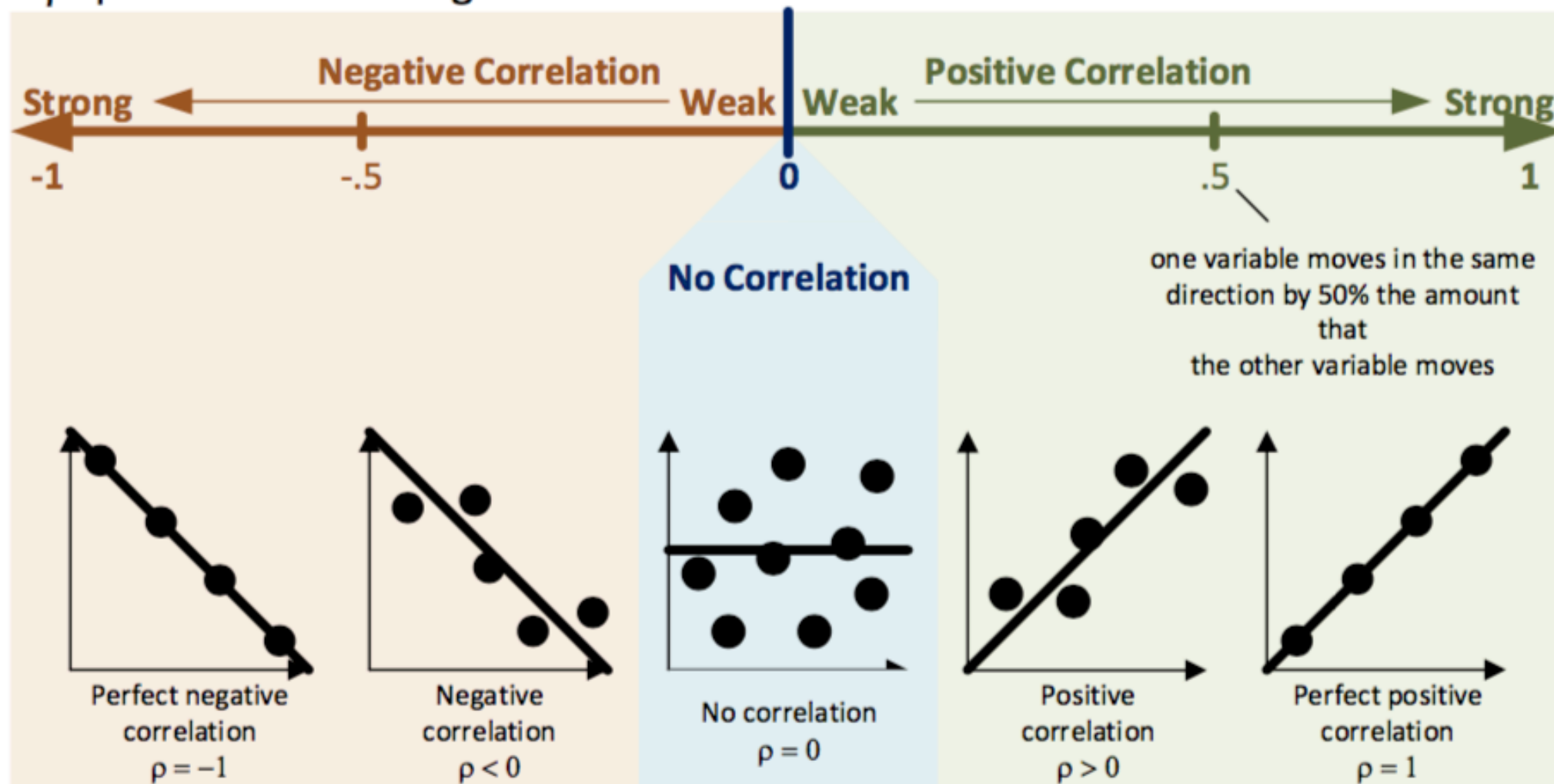
Let's form groups of 1's and 2's ...

# STATS SUMMARY

Measure of Centrality	Mean	Median	Mode
• In the dataset?	☹️	😊	😊
• Easy of compute	😊	😊	☹️
• Resistant to outliers?	☹️	😊	😊
Measure of Dispersion	😊 (Variance, Standard Deviation)	😊 (Interquartile Range)	☹️
Extensive used in mathematical models?	😊	☹️	☹️
Graphical Methods		Boxplot 	Histogram 

# CORRELATION

$\rho$  quantifies the strength and direction of movements of two random variables



---

# PYTHON & PANDAS

---

<i>Measure of Centrality</i>	<code>.mean()</code>	<code>.median()</code>	<code>.mode()</code>
<i>Measure of Dispersion</i>	<code>.var()</code> , <code>.std()</code>	<code>.min()</code> , <code>.max()</code> <code>.quantile()</code>	
<i>Summary</i>	<code>.describe()</code>		
<i>Graphical Methods</i>		<code>.plot(kind = 'box')</code>	<code>.plot(kind = 'hist')</code>
<i>Correlation Matrix</i>	<code>.corr()</code>		
<i>Scatter plot</i>	<code>DataFrame.plot(kind = 'scatter', x = 'SerieName', y = 'SerieName')</code>		
<i>Scatter matrix</i>	<code>pd.tools.plotting.scatter_matrix(DataFrame)</code>		



---

## STATISTICS FUNDAMENTALS, PART 2

---

# LEARNING OBJECTIVES

- Review of statistics and go over unit project 1
- 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)

---

**QUIZ**

---

**Let's Test Your  
Statistics ...**

---

# ANSWER 10 QUESTIONS IN 15-20 MINUTES

---

1. Import the wine data and show first few rows
  2. Transform categorical data and drop columns as needed
  3. Get the correlations between variables. Any strong correlations?
  4. Calculate kurtosis of your data.
  5. Which variable has the most positive kurtosis? Plot its density plot. For density plot, # refer to:  
<http://pandas.pydata.org/pandas-docs/version/0.15.0/visualization.html>
  6. Calculate skewness of your data.
  7. Which variable has the most null kurtosis? Plot its density plot. For density plot, # refer to:  
<http://pandas.pydata.org/pandas-docs/version/0.15.0/visualization.html>
  8. Get the basics statistics of your data
  9. Plot histogram of density, pH, and quality (subplots)
  10. Plot boxplots of density, pH, and quality
- Bonus: plot scatter matrix (look at pandas scatter matrix for volatile acidity, citric acid, residual sugar)

---

# UNIT PROJECT 1

---

## **Read and evaluate the following problem statement:**

Determine how likely free-tier customers are to convert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer usage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

### **1. What is the outcome?**

Answer: Likelihood of conversion to paid customer

### **2. What are the predictors/covariates?**

Answer: age, gender, location, profession, days since last log in, and activity score 1 = active user, 0= inactive user

### **3. What timeframe is this data relevant for?**

Answer: Jan-Apr 2015

### **4. What is the hypothesis?**

Answer: That customers who were more recently active are more likely to convert to the paid tier

# UNIT PROJECT 1

Let's get started with our dataset

```
In [1]: %matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
import pylab as pl
import numpy as np

df = pd.read_csv("../assets/admissions.csv")
df.head()
```

Out[1]:

	admit	gre	gpa	prestige
0	0	380	3.61	3
1	1	660	3.67	3
2	1	800	4.00	1
3	1	640	3.19	4
4	0	520	2.93	4

1. Create a data dictionary

Answer: To be completed when the dataset is finalized

Variable	Description	Type of Variable
Admit	0 = not admitted 1 = admitted	categorical
GRE	GRE score 200-800	continuous
GPA	GPA 0-4.0	continuous
Prestige	1= not prestigious 2 = low prestige 3= good prestige 4= high prestige	categorical

We would like to explore the association between admission into grad school and the prestige of undergraduate institutions.

1. What is the outcome?

Answer: admission into grad school

2. What are the predictors/covariates?

Answer: Prestige, GRE, GPA

---

# UNIT PROJECT 1

---

## 3. What timeframe is this data relevant for?

Answer: The timeframe for this data isn't immediately clear. This is something that could be researched further by contacting the original collectors of the data or researching the dataset's history. It could also be acknowledged as a potential limitation of the dataset.

## 4. What is the hypothesis?

Answer: Students that more prestigious undergraduate schools will have higher admissions rates into graduate school.

Using the above information, write a well-formed problem statement.

## Problem Statement

Determine if there is an association between graduate school admission and the prestige of a student's undergraduate school using data from the UCLA admissions data set.

---

# UNIT PROJECT 1

---

Using the lab from class as a guide, create an exploratory analysis plan.

## 1. What are the goals of the exploratory analysis?

Answer:

1. Determine if there is any missing data
2. Examine the distributions of the variables to determine if any of the variables need be transformed

## 2a. What are the assumptions of the distribution of data?

Answer: normality

## 2b. How will determine the distribution of your data?

Answer: histograms

## 3a. How might outliers impact your analysis?

Answer: They could skew the associations in the direction of the outlier.

## 3b. How will you test for outliers?

Answer: Box plots are one good way.

## 4a. What is colinearity?

Answer: when two variables are capturing similar variance in the data

## 4b. How will you test for colinearity?

Answer: create a correlation matrix

## 5. What is your exploratory analysis plan?

Using the above information, write an exploratory analysis plan that would allow you or a colleague to reproduce your analysis 1 year from now.

Answer:

1. Check for missing data and remove observations.
2. Check for colinearity.
3. Check for normal distribution.

---

## STATISTICS FUNDAMENTALS, PART 2

---

# LEARNING OBJECTIVES

- ~~Review of statistics and go over unit project 1~~
- 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.

## A few cups of coffee may lower colon cancer risk

Posted: 01 August 2007 17:08 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 95,000 people in Japan aged from 40 to 69 between the early 1990s and 2002, of whom 726 men and 437 women suffered colon cancer.



Photos 1 of 1

**Causal claims are often inconsistent and contradictory!**

**Search:**

Medline  
CancerConsultants.com  
Both

Cancer News: Rectal Cancer Article

Printable Version

**Main Menu**

Home  
Conference Coverage  
Current Topics in Oncology  
Cancer News  
Disease Centers  
Physician Resources  
About Us

**Quick Links**

Information by Disease  
All

Cancer News  
Select Center Type

Conference Coverage  
Select Conference

Brand Your Oncology Program Online

### 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*.<sup>[1]</sup>

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 646,648 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.

August 31, 2007

**WebMD**  
Better information. Better health.

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

WebMD Home > Health News

**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*).

life'sDHA  
Sample Daily Menu  
What's your Daily DHA?  
Find out Now

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

**BBC NEWS**

You are in: Health  
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

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

---

# 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.

---

# 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

---

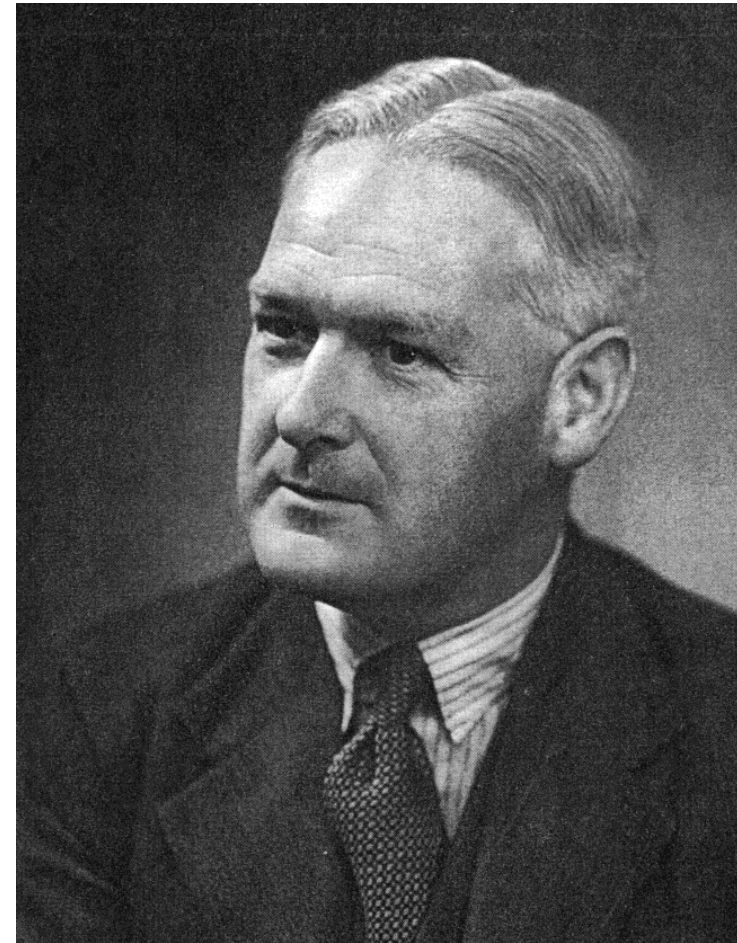
- Causal criteria is one approach to assessing causal relationships.
- However, it's *very hard to define* universal causal criteria.
- One attempt that is commonly used in the medical field is based on work by Bradford Hill.

---

# 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.Temporalitiy
  - 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

---

## ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. What is the difference between causation and association?

## DELIVERABLE

Answers to the above questions

---

# ACTIVITY: KNOWLEDGE CHECK

---



## EXERCISE

### **Correlation:**

- A statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
- A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

### **Causation:**

- One event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.
- This is also referred to as cause and effect.

---

## 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 lighters 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?

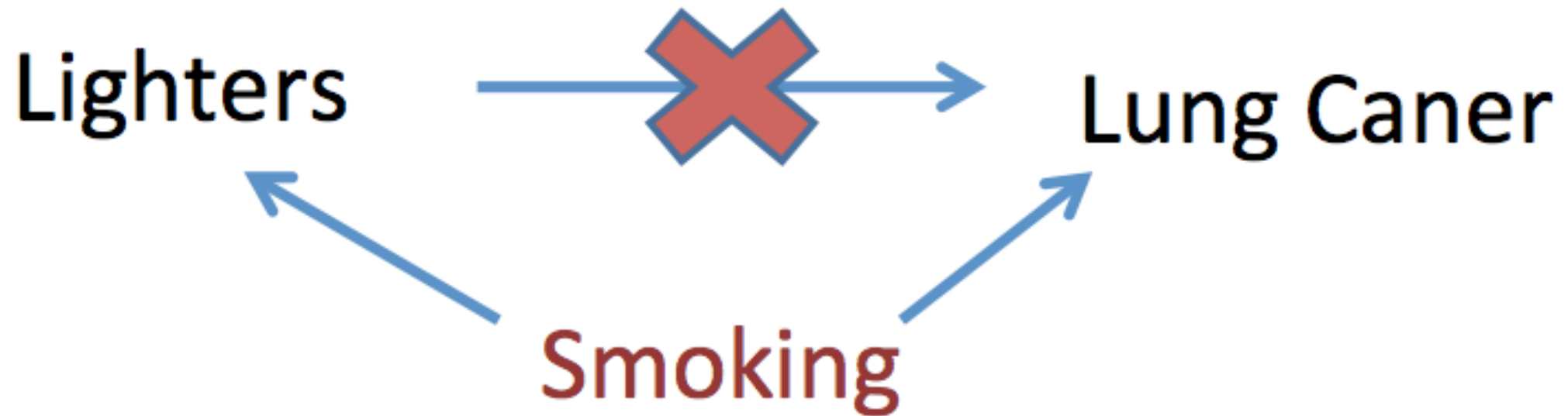


---

# CONFOUNDING

---

▸ No!

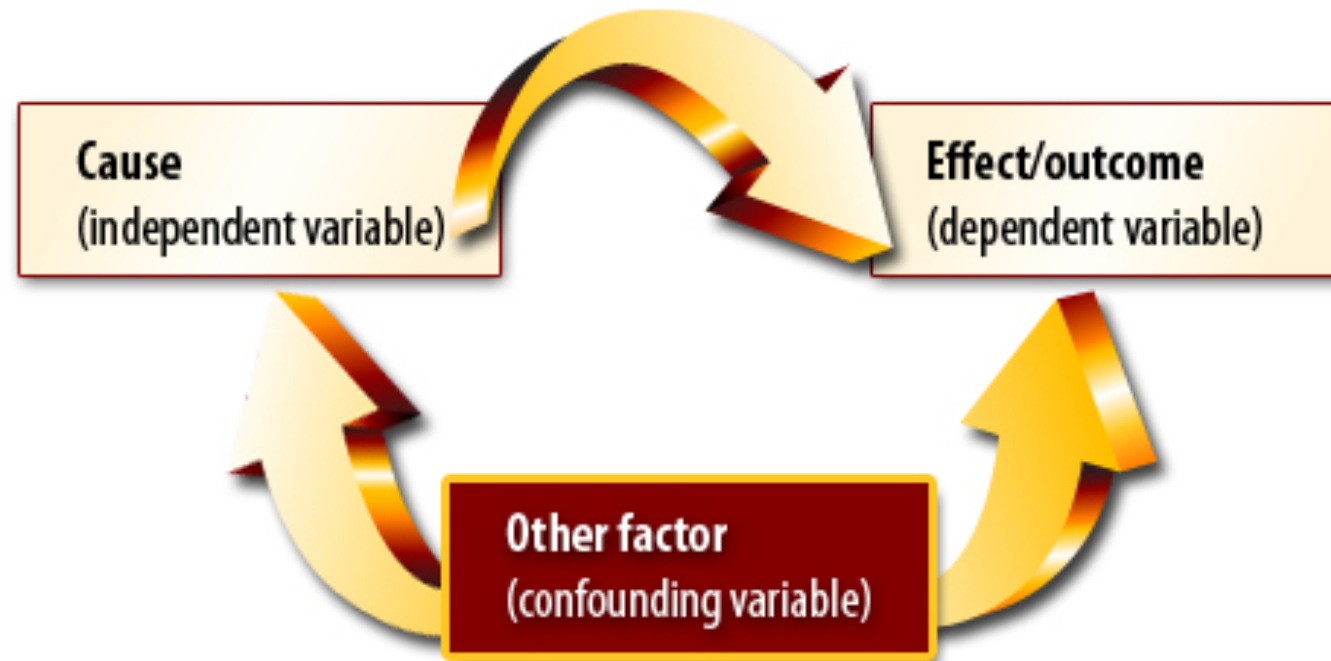


---

# CONFOUNDING

---

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



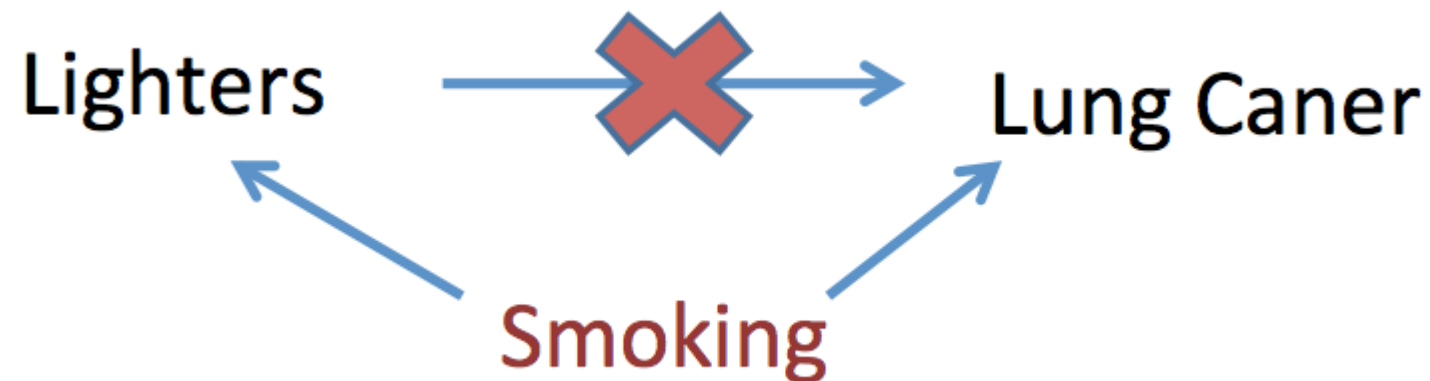
In statistics, a confounder is a variable that influences both the dependent variable and independent variable causing a spurious association.

---

# 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

---

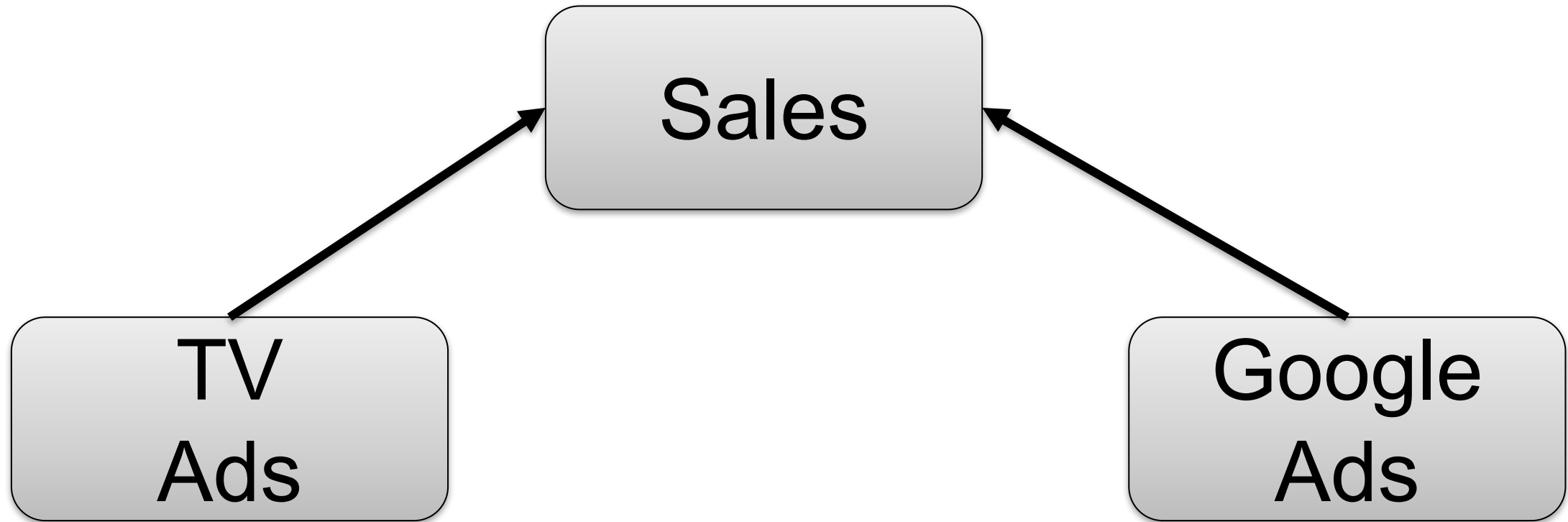
- The exposure/predictor is TV ads, associated with the outcome: sales.
- We can measure the strength to demonstrate a strong association.
- What other factors may increase sales?
- What other types of ads?

---

# 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 Age, BMI, Diabetes.
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?

---

## 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.

---

## STATISTICS FUNDAMENTALS, PART 2

---

# LEARNING OBJECTIVES

- ~~Review of statistics and go over unit project 1~~
- ~~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

---

# HYPOTHESIS TESTING

---

# HYPOTHESIS TESTING

---

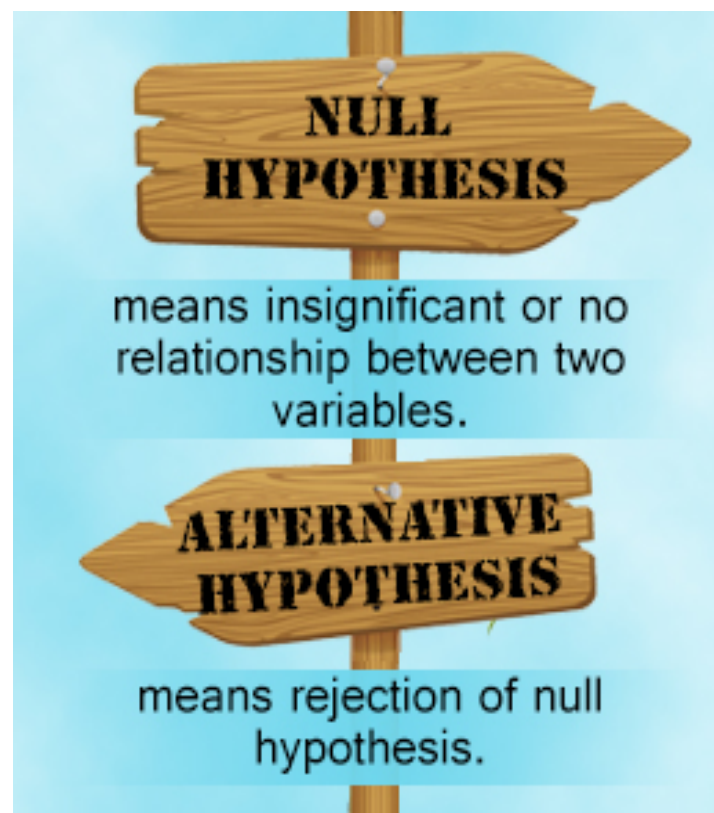
- How can we tell the difference between two groups of observations (e.g. smokers vs. non-smokers)?
- Imagine we are testing the health of smokers vs. non-smokers. At a cursory glance, our results may show that smokers are marginally healthier than non-smokers.
- Are they healthier due to random chance 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.
- Once you have your hypotheses, you can check whether the data supports rejecting the null hypothesis or failing to reject the hypothesis.

10 min

Break

---

**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

---

## ANSWER THE FOLLOWING QUESTIONS



EXERCISE

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

## DELIVERABLE

Answers to the above questions

---

## STATISTICS FUNDAMENTALS, PART 2

---

# LEARNING OBJECTIVES

- ~~Review of statistics and go over unit project 1~~
- ~~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

---

# VALIDATE YOUR FINDINGS

---

# VALIDATE YOUR FINDINGS

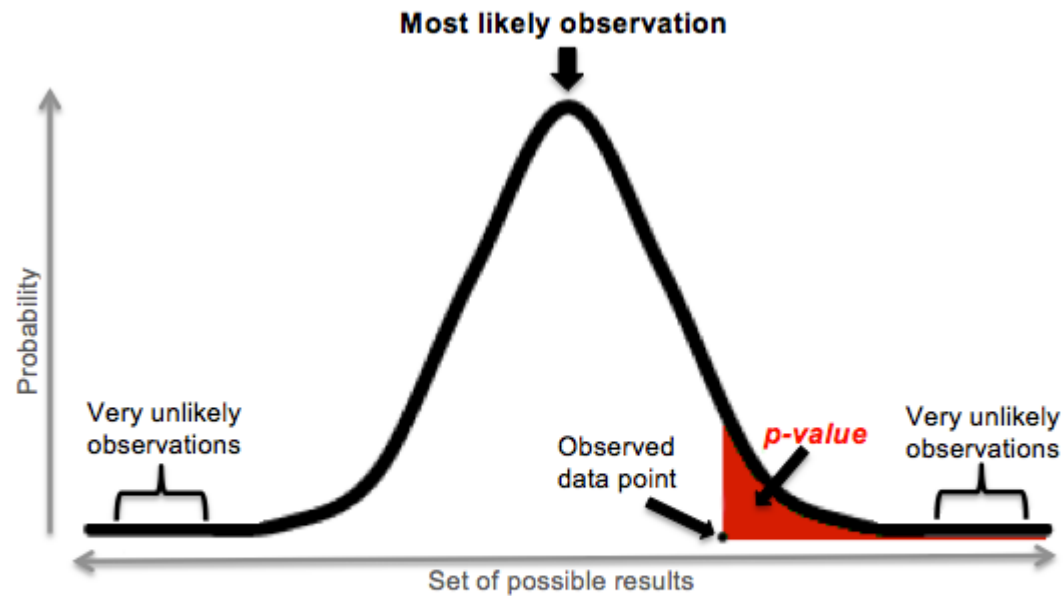
---

- 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

---

## ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. What does a 95% confidence interval indicate?

## DELIVERABLE

Answers to the above questions

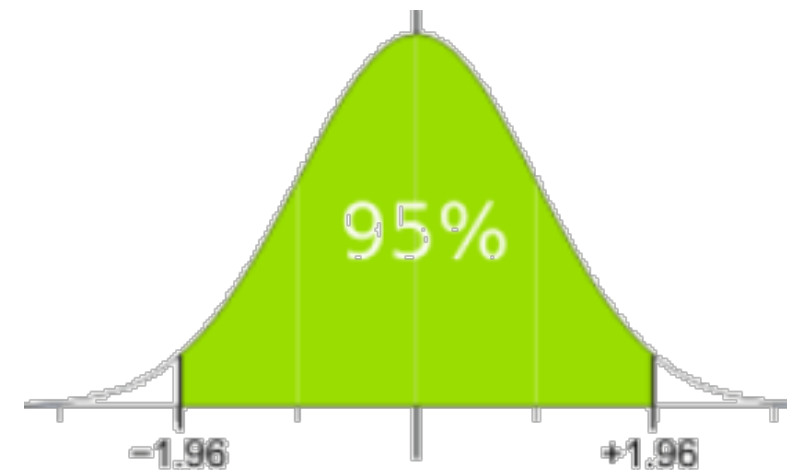
# ACTIVITY: KNOWLEDGE CHECK

## EXERCISE

### ANSWER THE FOLLOWING QUESTIONS

1. What does a 95% confidence interval indicate?

If we repeat this study 100 times, our point of estimate would lie in that range 95 times.



---

**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?

---

## STATISTICS FUNDAMENTALS, PART 2

---

# LEARNING OBJECTIVES

- ▶ ~~Review of statistics and go over unit project 1~~
- ▶ ~~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)~~

---

**LESSON**

---

# EXIT TICKET

**DON'T FORGET TO FILL OUT YOUR EXIT TICKET**

---

---

Q & A

---

**COURSE**

---

**BEFORE NEXT  
CLASS**

---

## BEFORE NEXT CLASS

---

# Start Working ...

- Project: Unit Project 2
- Think about Final Project ...



---

# OUR PROGRESS SO FAR

---

---

## UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

<del>What is Data Science</del>	<del>Lesson 1</del>
<del>Research Design and Pandas</del>	<del>Lesson 2</del>
<del>Statistics Fundamentals I</del>	<del>Lesson 3</del>
<del>Statistics Fundamentals II</del>	<del>Lesson 4</del>
Flexible Class Session	Lesson 5

---

## UNIT 2: FOUNDATIONS OF DATA MODELING

Introduction to Regression	Lesson 6
Evaluating Model Fit	Lesson 7
Introduction to Classification	Lesson 8
Introduction to Logistic Regression	Lesson 9
Communicating Logistic Regression Results	Lesson 10
Flexible Class Session	Lesson 11

---

## UNIT 3: DATA SCIENCE IN THE REAL WORLD

Decision Trees and Random Forests	Lesson 12
Natural Language Processing	Lesson 13
Dimensionality Reduction	Lesson 14
Time Series Data I	Lesson 15
Time Series Data II	Lesson 16
Database Technologies	Lesson 17
Where to Go Next	Lesson 18
Flexible Class Session	Lesson 19
Final Project Presentations	Lesson 20



---

**LESSON**

---

Q & A