Learning Objectives

After this lesson, you should be able to:

- Explain the difference between causation and correlation
- Identify a normal distribution within a dataset using summary statistics and visualization
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (t-tests, p-values, t-values, confidence intervals)



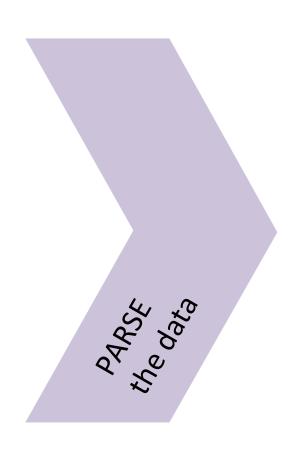
Review



Review

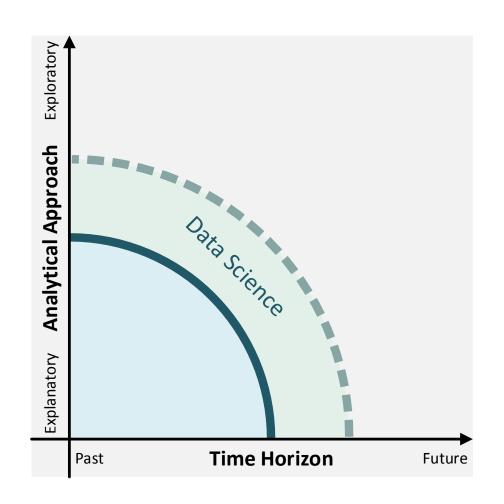
3 Parse the Data

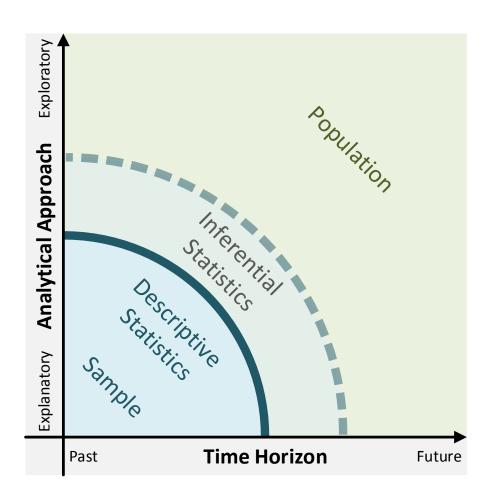
Parse the Data



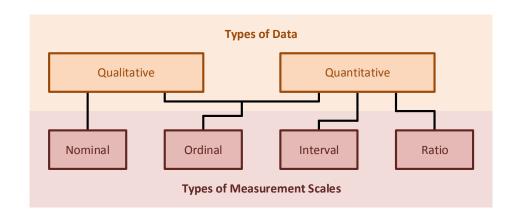
- Parse the Data
 - Read any documentation provided
 with the data (session 2)
 - Perform exploratory data analysis(session 3/4)
 - Verify the quality of the data(sessions 2/3)

Where Data Science fits with Descriptive and Inferential Statistics





Types of Data and Types of Measurement Scales

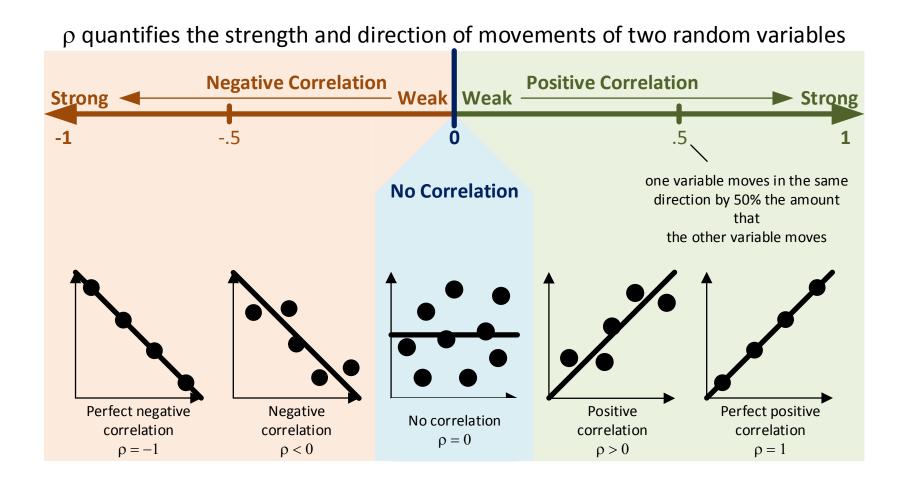


	Nominal	Ordinal	Interval	Ratio
Categorize?	✓	✓	✓	✓
Rank-order?	*	✓	✓	✓
+; -?	×	×	✓	✓
*; /?	×	×	×	✓

Descriptive Statistics

Measure of Centrality	Mean	Median	Mode
Measurement Scales	Interval - Ratio	Interval - Ratio	Nominal - Ratio
In the dataset?	8	©	©
 Easy of compute 	©	©	8
 Resistant to outliers? 	8	©	©
Measure of Dispersion	© (Variance, Standard Deviation)	☺ (Interquartile Range)	8
Extensive used in mathematical models?	©	8	8
Graphical Methods	$\int_{\mu}^{\overline{\sigma}}$	Boxplot ××	Histogram

Correlation



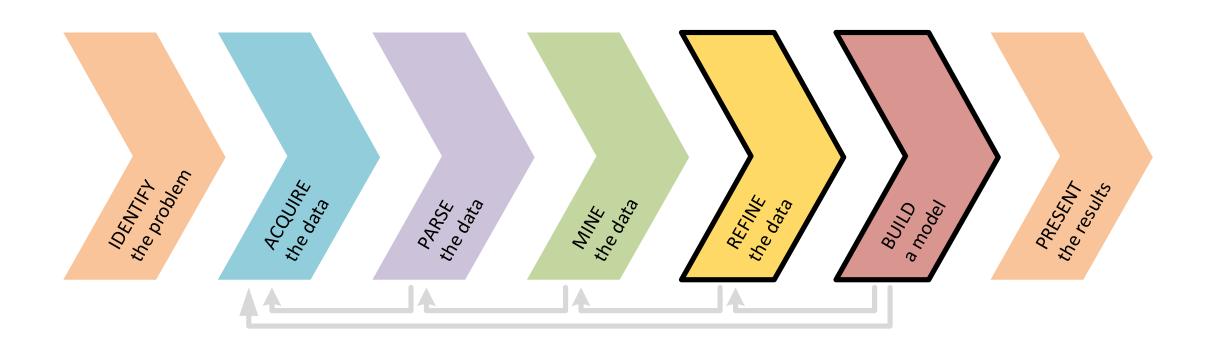
Python and pandas

Measure of Centrality	.mean()		.median())	.mode()
Measure of Dispersion	.var	<pre>.var(), .std() .quantile()</pre>		* *	
Summary		.descr	ribe()		
Graphical Methods		<pre>.plot(kind = 'box')</pre>		<pre>.plot(kind = 'hist')</pre>	
Correlation Matrix	.corr()				
Scatter plot	DataFrame.plot(kind = 'scatter', x = 'SerieName', y = 'Serie		me', y = 'SerieName')		
Scatter matrix	<pre>pd.tools.plotting.scatter_matrix(DataFrame)</pre>		DataFrame)		
.columns, .set_inde: .drop()	umns, .set_index(), .unique() .va .drop() .isnull(),		t(), .sum(), lue_counts(), .notnul(), ona()	nբ	o.sort(), .apply()



Today

Today we will shift our focus on the inferential statistics sections of REFINE the data and BUILD a model



Today, we are covering how inferential statistics is used in model fitting

Research Design	Research Design	Data Visualization in		Exploratory Data
and Data Analysis		pandas	Inferential Statistics for Model Fit	Analysis in <i>pandas</i>
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

Here's what happening today:

- Unit Project 2 due today
- Announcements and Exit Tickets
- Review
- • Refine the Data and **6** Build a Model
 - Causation and Correlation
 - If correlation doesn't imply causation, then what does?
 - Confounding
 - Hill's Criteria for Causation
 - Do you really need causality or is correlation enough?
 - Data Mining, "Fooled by Randomness", and Spurious Correlations
 - Inferential Statistics | Motivating Example

- The Normal Distribution
 - The 68 90 95 99.7 Rule
- Hypothesis Testing
 - Two-Tail Hypothesis Test
 - t-value
 - p-value
 - Confidence Intervals
- Lab Inferential Statistics for Model Fit
- Review
- Final Project 1 (due next session on 5/24)

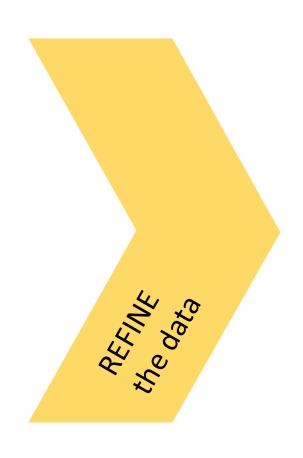


Q & A



Refine the Data
Build a Model

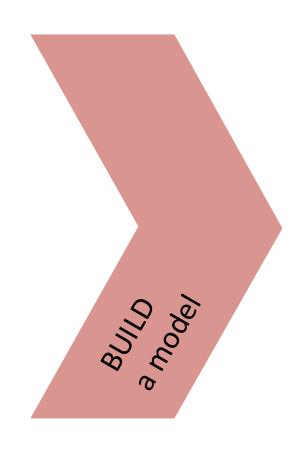
6 Refine the Data



Refine the Data

- Identify trends and outliers (session 3)
- Apply descriptive (sessions 3/4) and
 inferential statistics (session 5)
- Document (session 2) and transform data (units 2-3)

6 Build a Model



Build a Model

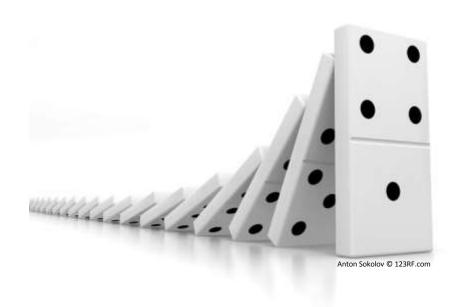
- Select appropriate model (units2-3)
- Build model (units 2-3)
- Evaluate (sessions 6/7) and refine model (units 2-3)



Refine the Data Build a Model

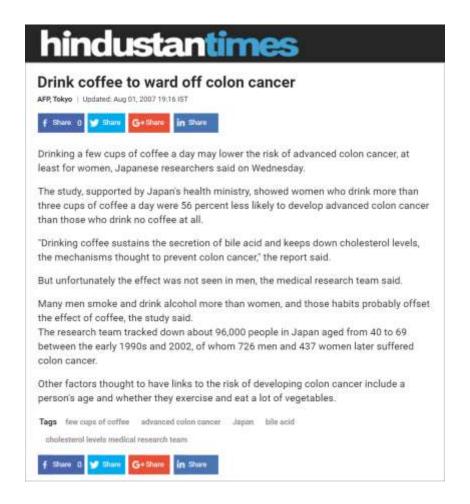
Causation and Correlation

Causation and Correlation



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

E.g., Coffee and Colon Cancer





Home » Coffee Does Not Decrease Risk of Colorectal Cancer Categories: Calon 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.

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).

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.

Reference:

[1] Je Y, Liu W, Giovannucci E. Coffee consumption and risk of colorectal cancer: A systematic review and meta-analysis of prospective cohort studies. *Informational Journal of Cancer*. 2009; 124: 1862-1868.

E.g., Alcohol and Dementia Risk



Alzheimer's disease is the most common cause of dementia, accounting for 50% of all

The society's research director Dr Richard Harvey said: "This interesting new study

"It is very much the case of a little of what you fancy appears to do you good."

confirms the results of previous research which has suggested that light to moderate

"It is particularly impressive that just 1-3 drinks per day can reduce the risk of vascular

"Clearly, however, excessive alcohol consumption is not good for our long term health and

All those taking part in the research were aged 55+ and did not have dementia at the start

Vascular dementia accounts for about 20% of cases.

alcoholic consumption is actually good for our health.

The Alpheimer's Society has welcomed the survey findings

increases the risk of serious diseases such as cirrhosis of the liver.

WHAND

Drinking and Dementia: Is There a Link?

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

By Galacter Streets WHOMD Health Resid

care the more than project only are generically precisioned to develop . Hough the data to suggest a protective benefit for light to excise size Authorized planeau, according to fredings from a Scandinarian study. In this lay, the studies receiving shrinking and airlings increasing are far

Remark the information of the Managard State of the Company of the infrequent drinkers have a twofold increase in the risk of deerenbuilty. Trusbudy is suggesting that people who don't drink according that drink Game constant within the quantity of this had a three field increase in this.

Tries anys, there are yours within things people with family highwave of

who did not have the generic risk factor. Low risk tectoraism; and invitaling isosping that blood procure, blood sugar, and community cognitive decrines taken in the as infrequent distincts.

The finalistic are improved in the Sept. 4 more of the BMJ increasivities British Medical Assensitio

"Cartier strattes indicated that light to moderate drinking macheprotective, but this study shows that the picture is much more complex." researcher Mile Elvisotts, MEI 796's, MRI WARMED. "The more people with this susceptibility gave break, the more their risk for domand is increased."

Apolinoprotein F

The study included just more than 1,000 more and women followed for an average of 20 years, who were between the ages of this and 79 at follow-up, At encolleged, the participants provided deballs about their Michigan consumption.

People were considered inframed distress. If they pract appropriates thus once a county and treasured drinkers if they shook never at times a

The insuranters also book blust satisfies to determine which shafts participants were common of the applicagnoists is greatings. The generation to an established rise factor for demonstra in settings, and as many as one in high Americans are corrient. Hybridis said.

The Karolinska resourcitory reported that dementia risk appeared to be directly related to drinking frequency omong study participants who were carriers of the gere.

"Our correct data indicate that howest accide to mong has harreful." affects on the brain, and this may be serve unusualized if there is genetic susceptibility." the researchers wrote. "We therefore its not ment to property people to which here alcohol in the belief that they ore protecting themselves against permettly

1 | 2 HEXT PAGE -

MOMO

Drinking and Dementia: Is There a Link?

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

Lifestyle Influences

Authorized Association virg president for residual and actoristic affairs Sept. 2, 2004. Driving alsohol in middle age musinerosay the risk of BRT floor, PhD, exheat the sentiment. This last, WeSMD that even from conclusive.

Not the findings also show a protective effect for infrequent drinkers. As hence it as other age-related demonstral cardo to refuse that this Proppert of inters in the study were twice as likely to experience relial upder control, maintaining a hardine on this party of asserbaand outing well. Other thes can be found in the "Moletain Your Brain" parties of the Atthebest 's Association selection (www.accorpt-

> "We have much better eximinar that there if exture factors contribute to Alaheriner's," Traes your

> > - PREVIOUS PAGE 1 1 2

Causation and Correlation (cont.)

- Why is this?
 - Sensational headlines
 - No robust data analysis
 - Lack of understanding of the difference between causation and correlation
 - → "caused" ≠ "measured" or "associated"
 - Correlation does not imply causation

- Understanding this difference is critical in the data science workflow, especially when **Identifying** the problem and **Acquiring** the data
 - We need to fully articulate our question and use the right data to answer it, including any confounders
- Additionally, this comes up when **Presenting** our results to stakeholders



Refine the Data Build a Model

If correlation doesn't imply causation, then what does?

Source: Michael Nielser



Refine the DataBuild a Model

Activity | In 1964, which group from the US House of Representatives most voted for the Civil Rights Act?

Activity | Simpson's Paradox (a.k.a., Yule–Simpson effect, reversal paradox, or amalgamation paradox): A trend appears in different groups of data but disappears or reverses when these groups are combined



	Democrats	Republicans
North	94%	85%
South	7%	0%
Overall	61%	80%

Activity | Simpson's Paradox: The Arithmetic



	Democrats	Republicans
North	145 / 154 = 94%	138 / 162 = 85%
South	7 / 94 = 7%	0 / 10 = 0%
Overall	152 / 248 = 61%	138 / 172 = 80%

Simpson's Paradox: Takeaway

Determining causality on the basis
 of correlations is tricky and can
 even lead to contradictory
 conclusions

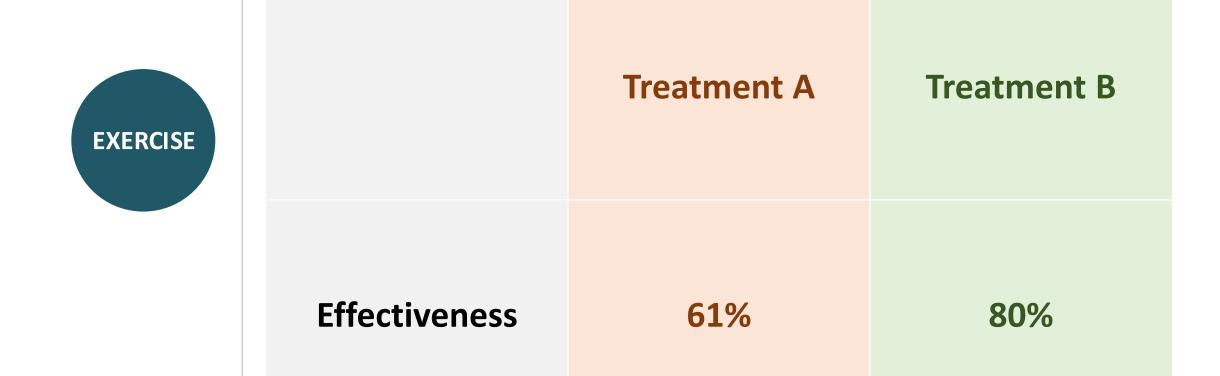
Partial evidence may be worse than no evidence if it leads to an illusion of knowledge, and so to overconfidence and certainty where none is justified. It's better to know that you don't know" – Michael Nielsen



Refine the Data Build a Model

Activity | Take 2

Activity | You suffer from kidney stones, and your doctor offers you two choices: treatment A or treatment B. Which will you chose?



Activity | The gotcha... Still going with treatment B?



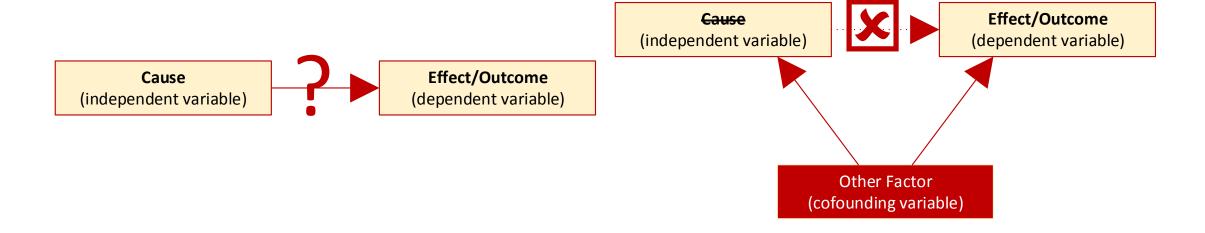
The gotcha	Treatment A	Treatment B
Patients with large kidney stones	94%	85%
Those with small kidney stones	7%	0%
Overall	61%	80%



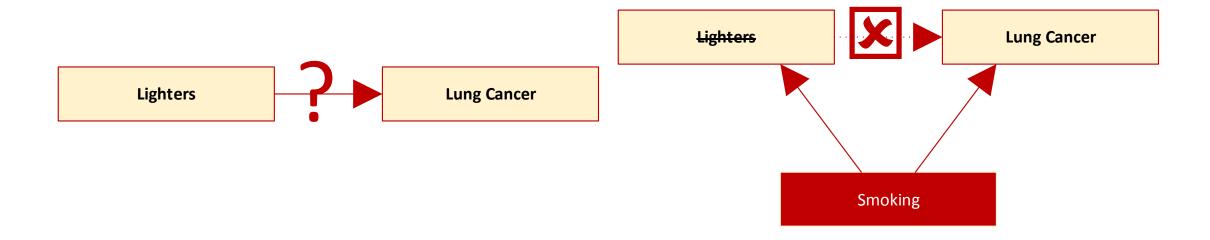
Refine the Data Build a Model

Confounding

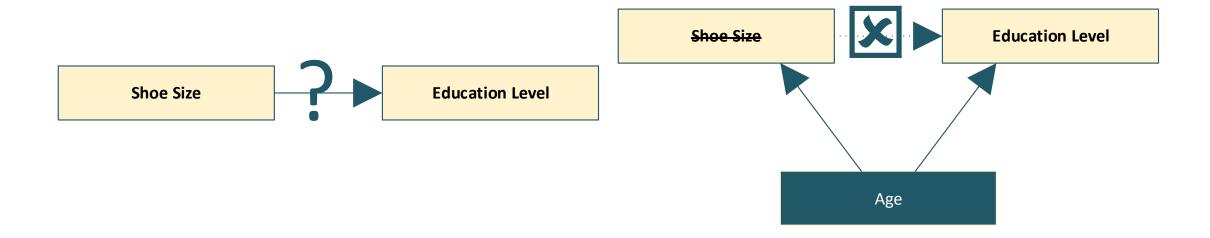
Confounding



Lighters causing lung cancer?



Shoe size as a proxy of education?





Refine the Data Build a Model

If you really need to establish causality...

The Bradford Hill Criteria (a.k.a., Hill's Criteria for Causation): a group of minimal conditions necessary to establish causality (commonly used in the medical field)

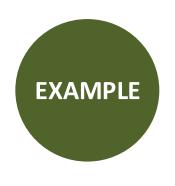
Strength	(or effect size): a small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal
Consistency	Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect
Specificity	Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship
Temporality	The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay)
Biological gradient	Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence
Plausibility	A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge)
Coherence	Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that " lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations"
Experiment	"Occasionally it is possible to appeal to experimental evidence"
Analogy	The effect of similar factors may be considered



Refine the Data Build a Model

Do you really need causality or is correlation enough?

Amazon



- "the Amazon Voice"
 - Hand-crafted reviews and title recommendations
 - Considered one of the company's crown jewels and a source of its competitive advantage
- What if Amazon could recommend specific books to customers based on their individual shopping preferences?
 - Comparing people with other people was cumbersome
 - All it needed to do was find associations among product themselves
 - "item-to-item" collaborative filtering patent
 - Data-generated material generated vastly more sales and the "Amazon Voice" group was disbanded
 - Knowing *what*, not *why*, is good enough

Amazon | "Item-to-Item" Collaborative Filtering

Collaborative recommendations using item-toitem similarity mappings

US 6266649 B1

ABSTRACT

A recommendations service recommends items to individual users based on a set of items that are known to be of interest to the user, such as a set of items previously purchased by the user. In the disclosed embodiments, the service is used to recommend products to users of a merchant's Web site. The service generates the recommendations using a previously-generated table which maps items to lists of "similar" items. The similarities reflected by the table are based on the collective interests of the community of users. For example, in one embodiment, the similarities are based on correlations between the purchases of items by users (e.g., items A and B are similar because a relatively large portion of the users that purchased item A also bought item B). The table also includes scores which indicate degrees of similarity between individual items.

Publication number US6266649 B1
Publication type Grant

Application number US 09/157,198
Publication date Jul 24, 2001
Filing date Sep 18, 1998
Priority date Sep 18, 1998

Fee status ⑦ Paid

Also published as EP1121658A1, EP1121658A4,

WO2000017792A1

Inventors Gregory D. Linden, Jennifer A. Jacobi, Eric A.

Benson

Original Assignee Amazon.Com, Inc.

Export Citation BiBTeX, EndNote, RefMan

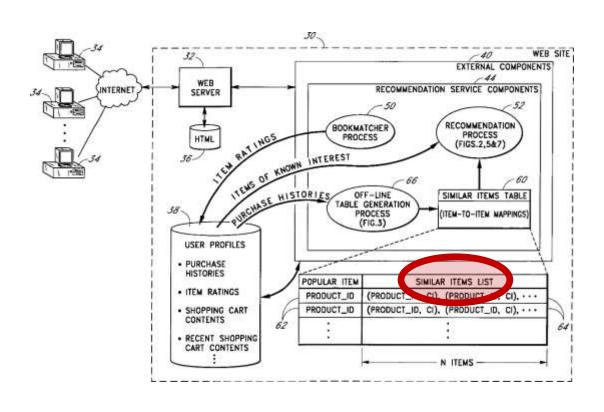
Patent Citations (22), Non-Patent Citations (39), Referenced by (1104),

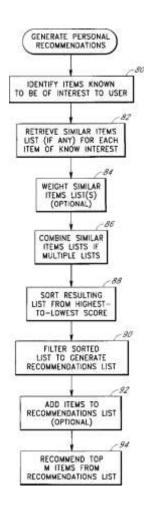
Classifications (23), Legal Events (9)

External Links: USPTO, USPTO Assignment, Espacenet

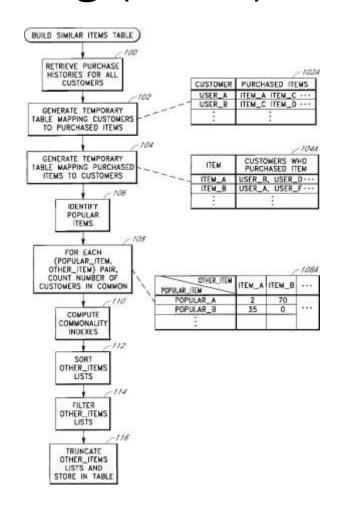
To generate personal recommendations, the service retrieves from the table the similar items lists corresponding to the items known to be of interest to the user. These similar items lists are appropriately combined into a single list, which is then sorted (based on combined similarity scores) and filtered to generate a list of recommended items. Also disclosed are various methods for using the current and/or past contents of a user's electronic shopping cart to generate recommendations. In one embodiment, the user can create multiple shopping carts, and can use the recommendation service to obtain recommendations that are specific to a designated shopping cart. In another embodiment, the recommendations are generated based on the current contents of a user's shopping cart, so that the recommendations tend to correspond to the current shopping task being performed by the user.

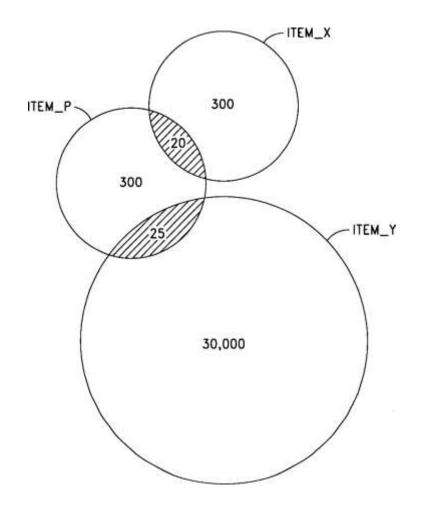
Amazon | "Item-to-Item" Collaborative Filtering (cont.)



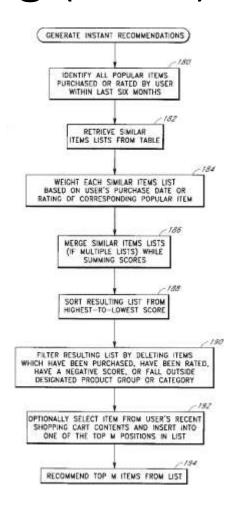


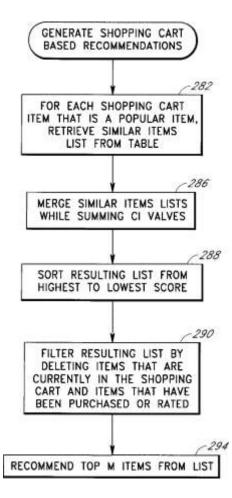
Amazon | "Item-to-Item" Collaborative Filtering (cont.)





Amazon | "Item-to-Item" Collaborative Filtering (cont.)



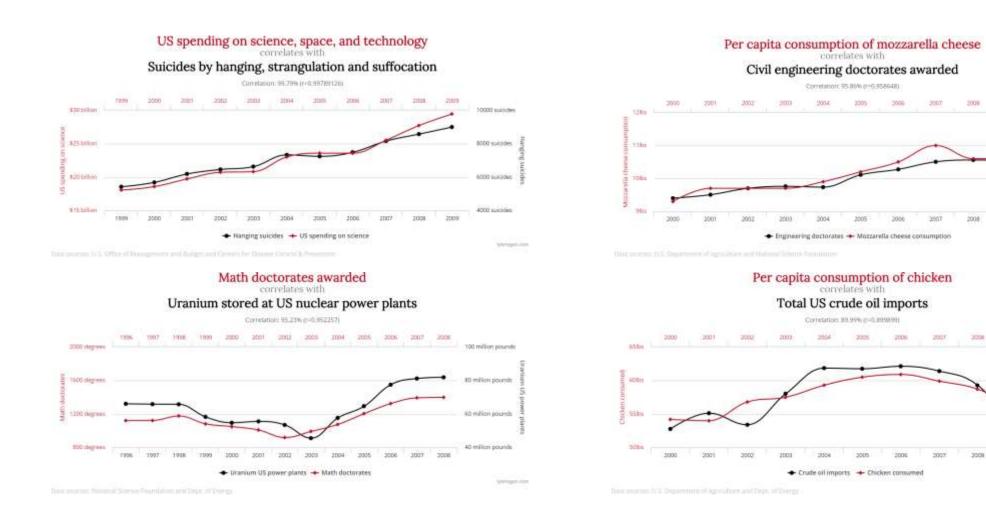




Refine the Data Build a Model

Data Mining, "Fooled by Randomness", and Spurious Correlations

Spurious Correlations



SOC degrees

600 Hegrees

3.85 tamete.

3.65 harrels

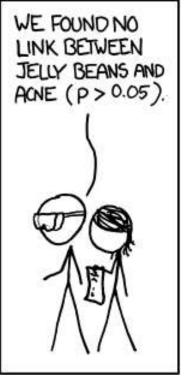
3.48 barrels

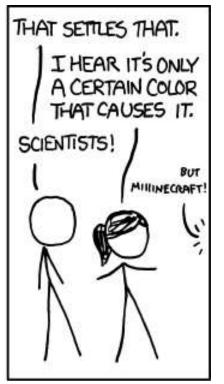
3.26 barrels

Source: tylervigen.com

Data Mining

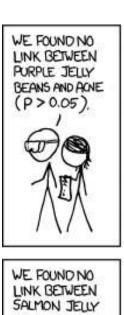




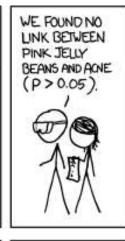


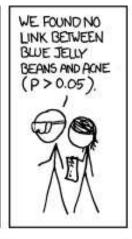
Source: xkcd.com

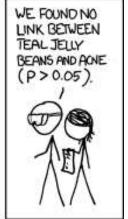
Data Mining (cont.)

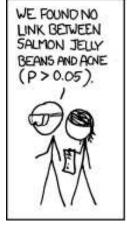


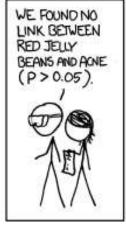


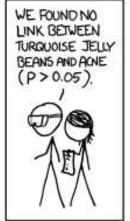


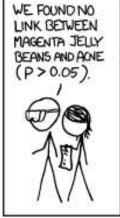


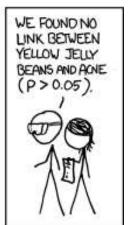






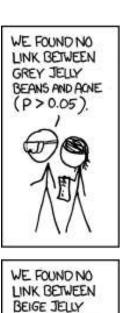


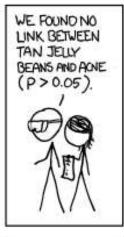


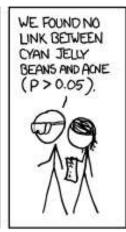


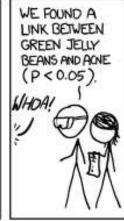
Source: xkcd.com

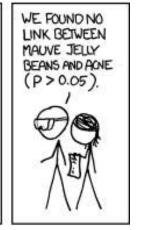
Data Mining (cont.)

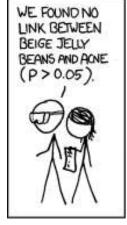


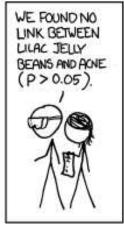


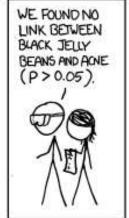


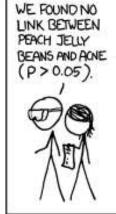


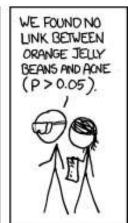






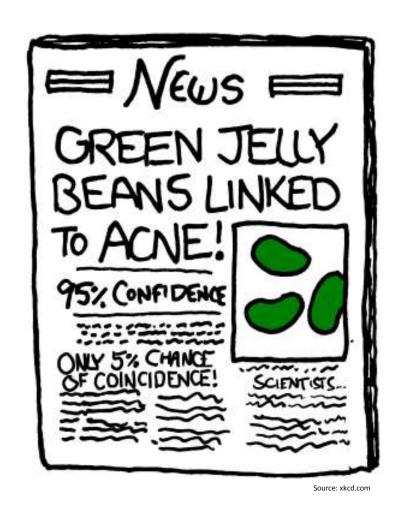






Source: xkcd.com

Data Mining (cont.)





Refine the DataBuild a Model

Inferential Statistics | Motivating Example

We are using our usual SF housing dataset but we added two new variables M1 and M2 to it

	Address	DateOfSale	SalePrice	IsAStudio	BedCount		Size	LotSize	BuiltInYea	M1	M2
ID											
15063471	55 Vandewater St APT 9, San Francisco, CA	12/4/15	710000	0	1		550	NaN	1980	1.099658	0.097627
15063505	740 Francisco St, San Francisco, CA	11/30/15	2150000	0	NaN		1430	2435	1948	3.687657	0.430379
15063609	819 Francisco St, San Francisco, CA	11/12/15	5600000	0	2	***	2040	3920	1976	8.975475	0.205527
15064044	199 Chestnut St APT 5, San Francisco, CA	12/11/15	1500000	О	1	-70	1060	NaN	1930	2.317325	0.089766
15064257	111 Chestnut St APT 403, San Francisco, CA	1/15/16	970000	0	2		1299	NaN	1993	1.380945	-0.152690
	***	***	***	***	***	***	***		***	***	***
2124214951	412 Green St APT A, San Francisco, CA	1/15/16	390000	1	NaN	444	264	NaN	2012	0.428094	-0.804647
2126960082	355 1st St UNIT 1905, San Francisco, CA	11/20/15	860000	0	1	***	691	NaN	2004	1.302833	0.029844
2128308939	33 Santa Cruz Ave, San Francisco, CA	12/10/15	830000	0	3		1738	2299	1976	1.608882	0.876824
2131957929	1821 Grant Ave, San Francisco, CA	12/15/15	835000	0	2		1048	NaN	1975	1.025920	-0.542707
2136213970	1200 Gough St, San Francisco, CA	1/10/16	825000	0	1		900	NaN	1966	1.383641	0.354282



Refine the Data Build a Model

Activity | Knowledge Check

Activity | Knowledge Check



DIRECTIONS (10 minutes)

- 1. Perform Data Exploratory Analysis on the these two "mystery" variables M1 and M2 and how they relate to SalePrice
- 2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions



Refine the Data Build a Model

Your first Machine Learning Models

Machine Learning Model #1 | SalePrice as a function of M1

$$SalePrice = \beta_1 \cdot M1$$

```
X = df[ ['M1'] ]
y = df.SalePrice

model = smf.OLS(y, X).fit()
```

How do we interpret these results?

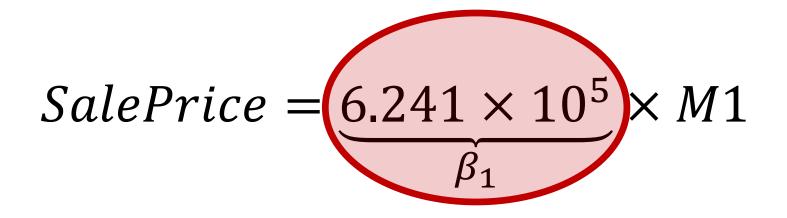
OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	2.567e+04
Date:		Prob (F-statistic):	0.00
Time:		Log-Likelihood:	-14393.
No. Observations:	1000	AIC:	2.879e+04
Df Residuals:	999	BIC:	2.879e+04
Df Model:	1		
Covariance Type:	nonrobust		
		_	

	coef	std err	t	P> t	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05

Omnibus:	1044.296	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	901486.247
Skew:	3.948	Prob(JB):	0.00
Kurtosis:	149.879	Cond. No.	1.00

SalePrice as a function of M1



• But how good is this model?

How good is this model?

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	2.567e+04
Date:		Prob (F-statistic):	0.00
Time:	23	Log-Libood:	14393.
No. Observations:	1000	AIC:	2.879e+04
Df Residuals:	999	BIC:	2.879e+04
Df Model:	1		5
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[95.0% Conf. Int.]
M1	6.241e+05	3894.95	160.228	0.000	6.16e+05 6.32e+05

Omnibus:	1044.296	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	901486.247
Skew:	3.948	Prob(JB):	0.00
Kurtosis:	149.879	Cond. No.	1.00

Machine Learning Model #2 | SalePrice as a function of M2

$$SalePrice = \beta_1 \cdot M2$$

```
X = df[ ['M2'] ]
y = df.SalePrice

model = smf.OLS(y, X).fit()
```

SalePrice = $3.195 \times 10^5 \times M2$. But again, how good is this model?

OLS Regression Results 0.000 Dep. Variable: SalePrice R-squared: OLS Adj. R-squared: Model: -0.001 Least Squares 0.06941 Method: F-statistic: Prob (F-statistic): 0.792 Date: Time: AIC: 3.207e+04 No. Observations: 1000 **Df Residuals:** 999 BIC: 3.208e+04 Df Model: Covariance Type: nonrobust

Politi 195 0% Conf Int 1

	coei	stu en		1-14	[95.0%	Com. maj
M2	3.195e+04	1.21e+b	0.263	0.792	-2.06e	+05 2.7e+05
Om	nibus:	1664.600	Duri	bin-Wa	tson:	0.971
Pro	b(Omnibus):	0.000	Jaro	ue-Ber	a (JB):	986904.813
Ske	w:	10.532	Prot	o(JB):		0.00
Kur	tosis:	155.453	Con	d. No.		1.00

Today, we will start with the coefficients' statistics and answer the following question: From a statistical standpoint, are these coefficients "significant", i.e., do they make sense?

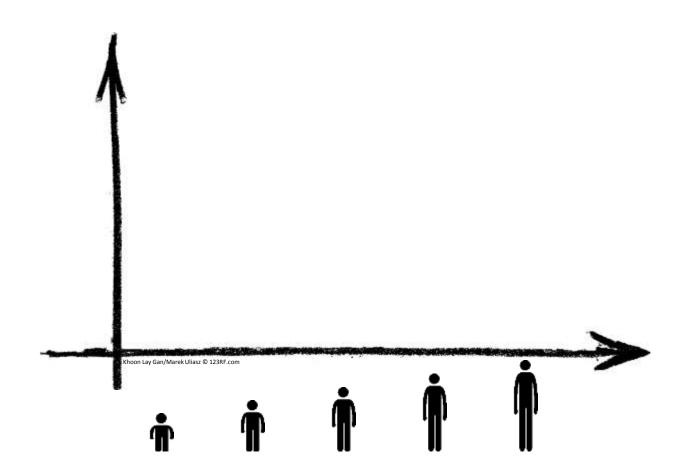
	coef	std err	t	P> t	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05
M2	3.195e+04	1.21e+05	0.263	0.792	-2.06e+05 2.7e+05



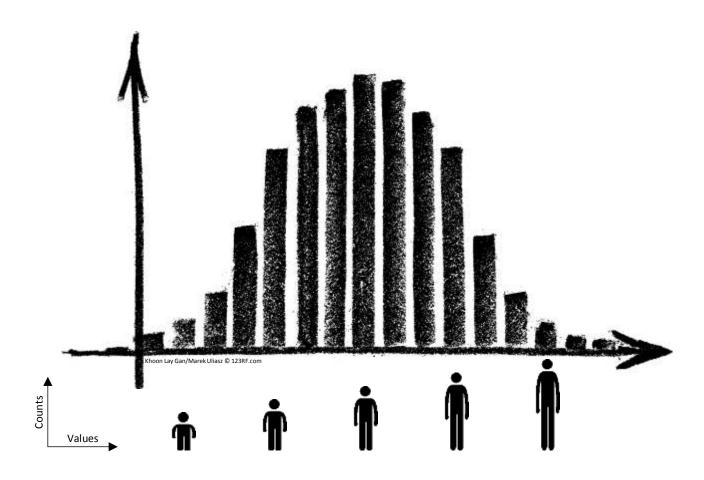
Refine the Data Build a Model

The Normal Distribution

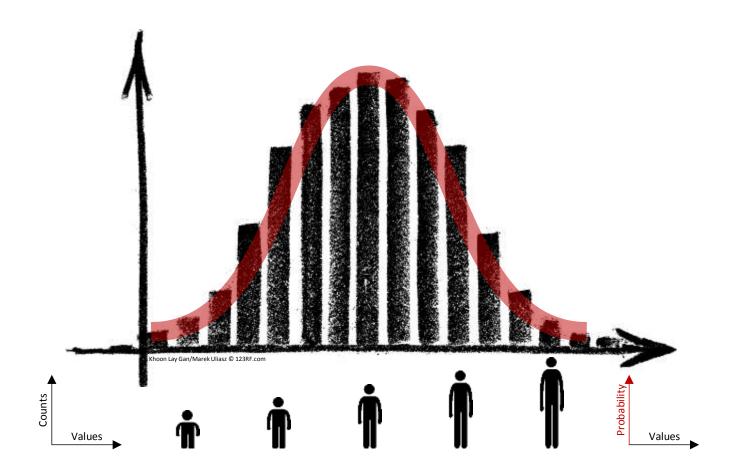
How is people's height distributed?



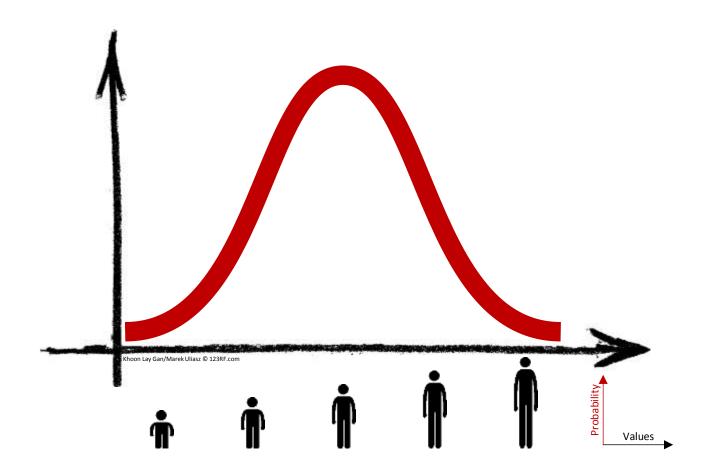
People's height follows a bell shape distribution. (For men, the average height is around 70 inches (5-10), with few people shorter than 67 inches, and few as tall as 73 inches)



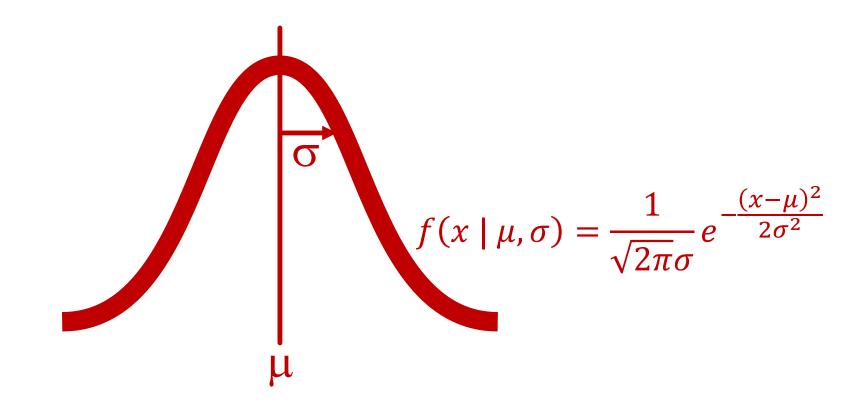
People's height follows a bell shape distribution (cont.)



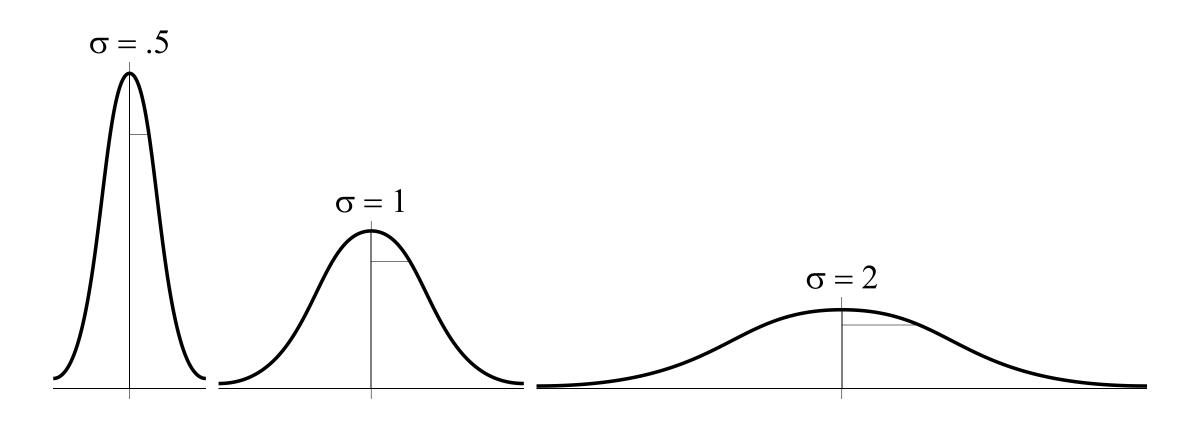
The Normal Distribution



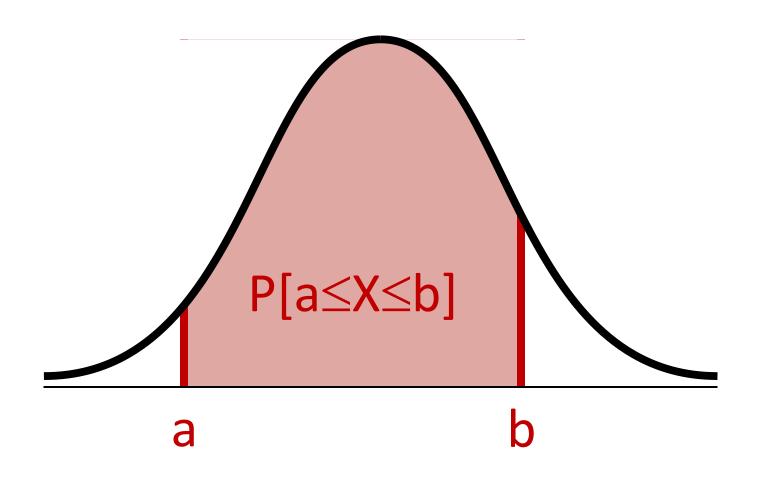
The Normal Distribution (cont.)



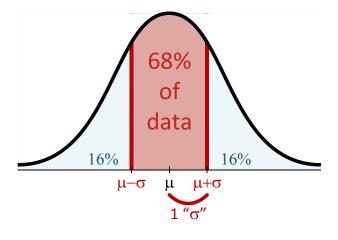
This is a probability density function: The area under the curve is always 1 (for any σ) (cont.)

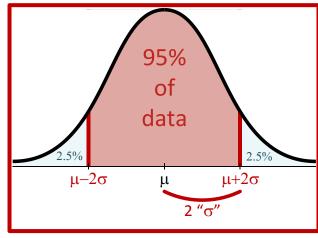


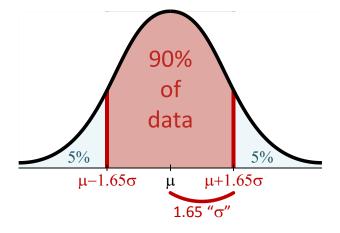
How to read a probability density function?

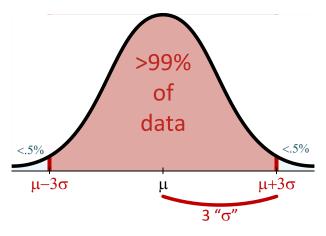


The 68 - 90 - 95 - 99.7 Rule









Hypothesis Testing

- A hypothesis is an assumption about the a population parameter. E.g.,
 - M1's coefficient is 6.241×10^5
 - M2's coefficient is 3.195×10^4
- In both cases, we made a statement about a population parameter that may or may not be true
- The purpose of hypothesis testing is to make a statistical conclusion about
 rejecting or failing to reject such statement

Two-Tail Hypothesis Test

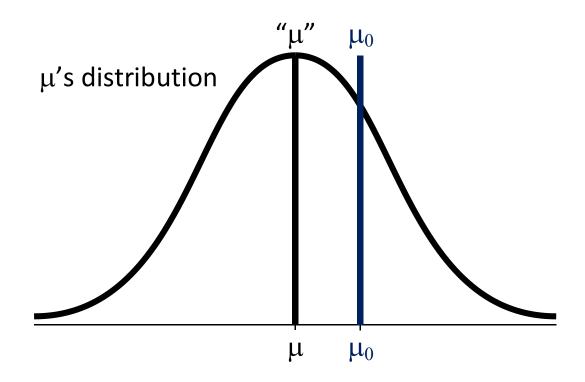
• The *null hypothesis* (H_0) represents the status quo; that the mean of the population is equal to a specific value:

$$H_0$$
: $\mu = \mu_0$

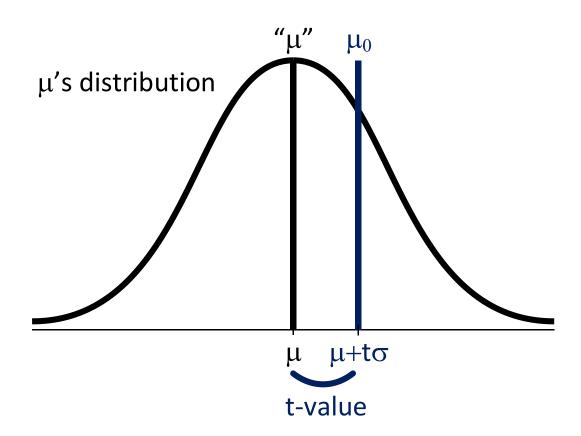
• The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis and holds true if the *null hypothesis* is found to be false:

$$H_a$$
: $\mu \neq \mu_0$

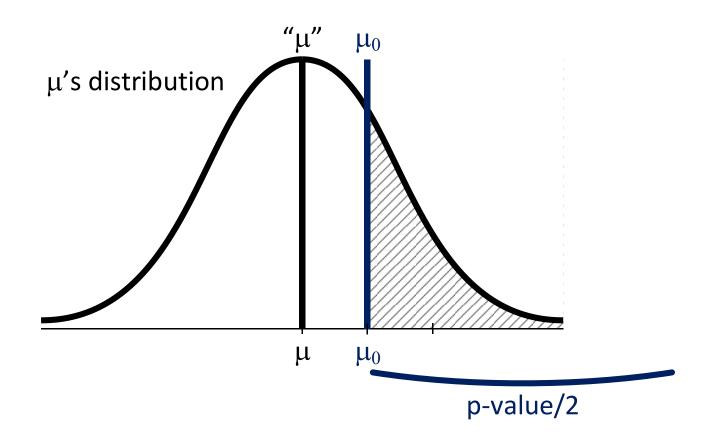
Two-Tail Hypothesis Test (cont.)



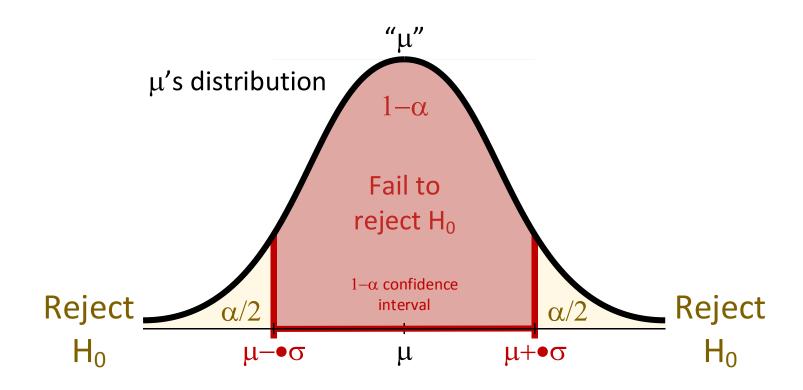
t-value measures the difference to μ_0 in σ . t-values of large magnitudes (either negative or positive) are less likely. The far left and right "tails" of the distribution curve represent instances of obtaining extreme values of t, far from μ



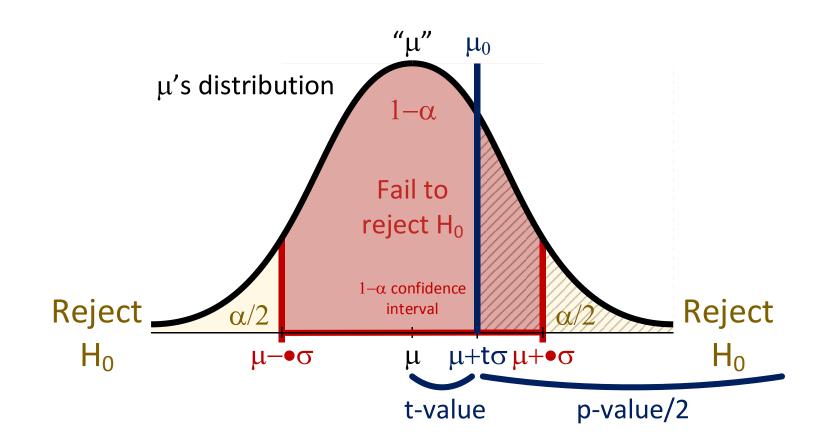
p-value determines the probability (assuming the H_0 is true) of observing a more extreme test statistic in the direction of H_a than the one observed



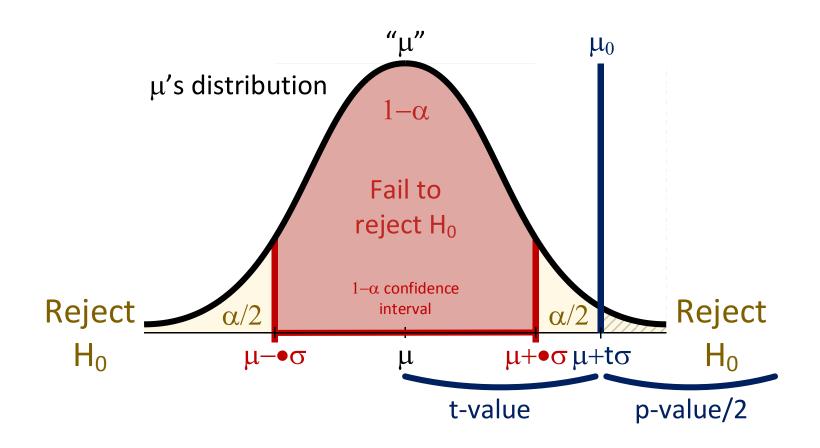
Two-Tail Hypothesis Test (simplified) (cont.)



Two-Tail Hypothesis Test (cont.)



Two-Tail Hypothesis Test (cont.)



Two-Tail Hypothesis Test (cont.)

t-value	p-value	$1-\alpha$ Confidence Interval $([\mu_0-\cdot\sigma,\mu_0+\cdot\sigma])$	H_0/H_a	Conclusion	
≥·	≤ α	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H ₀	$\mu \neq \mu_0$	
<.	> \alpha	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu=\mu_0$ (assume)	

Two-Tail Hypothesis Test ($\alpha = .05$) (cont.)

t-value	p-value	95% Confidence Interval $([\mu_0-2\sigma,\mu_0+$	H ₀ / H _a	Conclusion	
\geq " \sim 2"(*) (*) (check the t-table slide)	≤ .025	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H ₀	$\mu \neq \mu_0$	
< "~2"(*)	> .025	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu=\mu_0$ (assume)	



Refine the Data Build a Model

Activity | Knowledge Check

Activity | Knowledge Check



DIRECTIONS (10 minutes)

1. What are the *null* and *alternate hypothesis* for the M1 and M2 coefficients? (Hint: What makes these coefficients "statistically" significant?)

	coef	std err	t	P> t	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05
M2	3.195e+04	1.21e+05	0.263	0.792	-2.06e+05 2.7e+05

2. When finished, share your answers with your table

DELIVERABLE

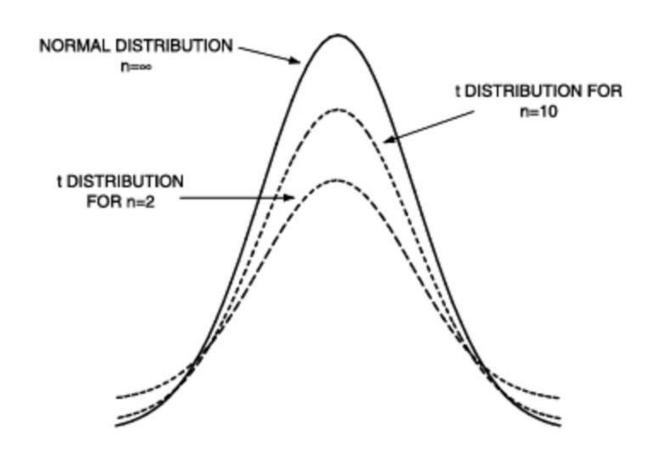
Answers to the above questions



Refine the Data Build a Model

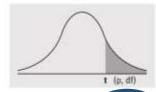
Student's t-distribution

FYI | We simplified things a bit... t-values use the Student's t-distribution, not the normal distribution



Student's t-distribution table (cont.)

Numbers in each row of the table are values on a t-distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208

CI			80%	90%	95%	98%	99%	99.9%
Z	0.253347	0.674490	1.281552	1.6448.	1.95996	2.32635	2.57583	3.2905
30	0.255605	0.682756	1.310415	1.697721	2.04227	45726	2.75000	3.6460
29	0.255684	0.683044	1.311434	1.699127	2 04523	2.46202	2.75639	3.6594
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
23	0256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
22	0256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
18	0.257123	0.688364	1.330391	1.734064	2.10092	2,55238	2.87844	3.9216
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
16	0257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405



Lab

Inferential Statistics for Model Fit



Review

Review

You should now be able to:

- Explain the difference between causation and correlation
- Identify a normal distribution within a dataset using summary statistics and visualization
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (t-tests, p-values, t-values, confidence intervals)



Q & A



Before Next Class

Before Next Class

- Understand the difference between vectors, matrices, *pandas Series*, and *pandas DataFrames*
- Understand the concepts of outliers and distance
- Effectively show correlations between an independent variable *X* and a dependent variable *Y*
- Be able to interpret t-values, p-values, and confidence intervals
- Install the *seaborn* Python package:
 - % conda install seaborn



Exit Ticket

Don't forget to fill out your exit ticket here