Inferential Statistics for Model Fit

Ivan Corneillet

Data Scientist



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)



Announcements and Exit Tickets



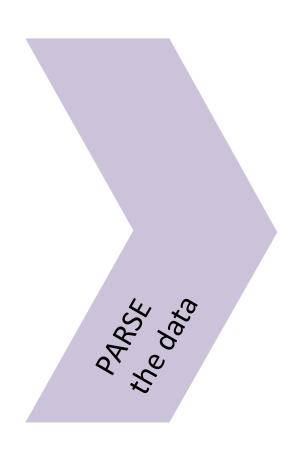
Review



Review

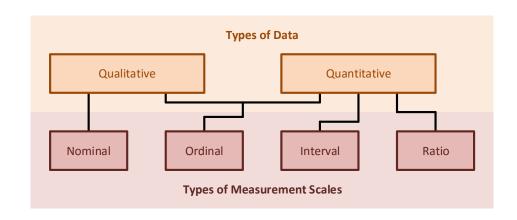
3 Parse the Data

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)

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	©	©
Ease 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

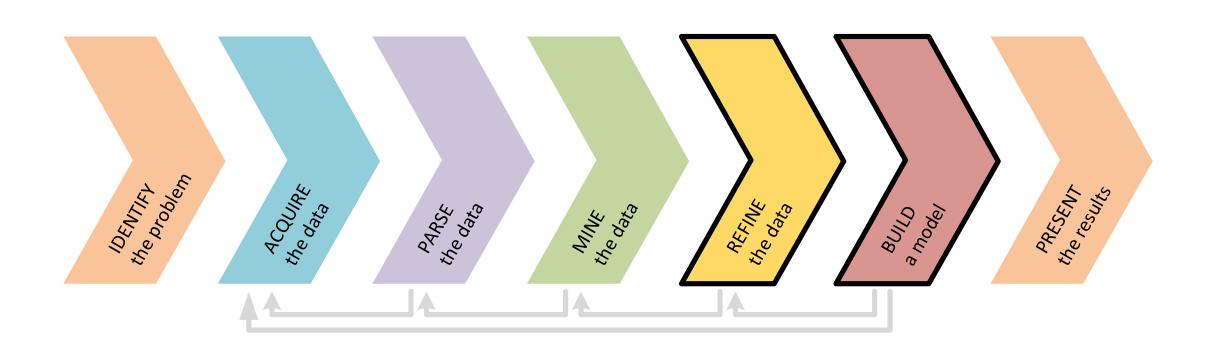
Correlation

 ρ quantifies the strength and direction of movements of two random variables **Negative Correlation Positive Correlation** Strong Weak Weak Strong -1 -.5 one variable moves in the same **No Correlation** direction by 50% the amount that the other variable moves Perfect negative Negative Positive Perfect positive No correlation correlation correlation correlation correlation $\rho = 0$ $\rho = -1$ $\rho < 0$ $\rho > 0$ $\rho = 1$



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 and Data Analysis	Research Design	Data Visualization in pandas	Statistics	Exploratory Data 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's happening today:

- Announcements and Exit Tickets
- Review
- G Refine the Data and G Build a Model
 - Causation and Correlation
 - If correlation doesn't imply causation, then what does?
 - Confounding
 - Do you really need causality or is correlation enough?
 - Data Mining, "Fooled by Randomness", and Spurious Correlations
 - Inferential Statistics | Motivating Example

- The Normal Distribution
 - \rightarrow 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
- Exit Tickets

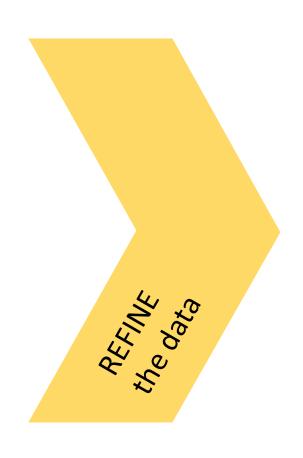


Q & A



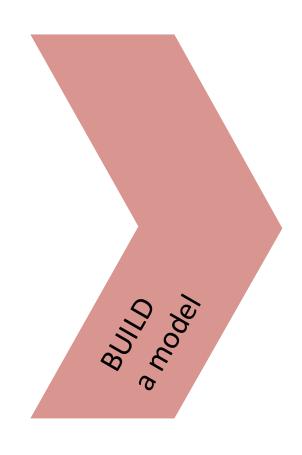
Refine the Data Build a Model

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

If an association is observed, the first question to ask should always be... is it real? E.g., Coffee and Colon Cancer





Coffee Does Not Decrease Risk of Colorectal Cancer

Categories: Calon Cancer, News, Rectal Cancer.

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

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

Some studies have indicated that coffee may have a protective effect against colon cancer, however, researchers continue to evaluate this link in an effort to establish more direct evidence. In order to examine the relationship between coffee consumption and colorectal cancer, researchers from Harvard conducted a review of 12 studies that included 646,848 participants and 5,403 cases of colorectal cancer.

They evaluated high versus low coffee consumption and found no significant effect of coffee consumption on colorectal cancer risk. The review included four studies in the United States, five in Europe, and three in Japan. The data from each country was very similar. There were no significant differences by gender or site of cancer, however, there was a slight inverse relationship (reduction in risk) between coffee consumption and colon cancer for women, which was even more pronounced among Japanese women (21% for total study, 38% for Japanese women).

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. *International Journal of Cancer*. 2009; 124: 1862-1868.

If an association is observed, the first question to ask should always be... is it real? E.g., Alcohol and Dementia Risk



WHAND MOMO Drinking and Dementia: Is There a Drinking and Dementia: Is There a Link? Link? Study Shows Drinkers With Genetic Study Shows Drinkers With Genetic Predisposition to Alzheimer's Disease at Predisposition to Alzheimer's Disease at Higher Risk Higher Risk By Galacter Streets Lifestyle influences WHOMD Health Resid Authorized Association visit president for restrict and extensive affects Sept. 2, 2004. Driving alsohol in middle age musiners are the risk of BRT floor. PhD, exheat the sentiment. This last, WeSMD that even care the more than people who are greetically predisposed to develop. Hough the data to suggest a protective femality by light to explorate Authorized planeau, according to fredings from a Scandinarian study. In the lay the studies receiving a friending and airlings received and far from conclusive. Researcher's from Montholie's Karolinska bed indevegoried that Gene contient situalitemently strink had a threefold increase in this. . This says there are many after things penals with family topones of Not the findings also show a protective effect for infrequent drinkers. As hence it as other age-related demonstral care do no reduce that this who did not have the greatly risk factor. One this testistation and invitable looping that blood property board and and common of Proppert of inters in the study were twice as likely to experience reliable upder control, maintaining a hardine wealth, getting planty of control cognitive decrines taken in the accommoded decimary. and eating well. Other tips can be found in the "Mointain Your Brain" parties of the Atthebest 's Association selection (www.accorpt-The finalings are reperied in the Sept. 4 more of the BNU framewive the British Medical Assendi. "We have much better exhibitor that these illestate factors contribute "Cartier strattes indicated that light to moderate drinking macheproteinflive. But this study shows that the picture is much more - PREVIOUS PAGE 1 2 complex." researcher Mile Kivisetto, MEI 9763, MIS WARMED. "The more people with this susceptibility gave presit, the more their mix for domand is increased." Apolinoprotein F The glady included but 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 Microsoft commercials People were considered infrequent drawers if they study ground less thus once a county and treasured drinkers if they shook never at times a The income have also book bloom parenter to determine until to their participants were carriers of the applicacy of evil 5 personers. 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 researchers reported that deeserd a risk appeared to be directly related to drinking frequency omong study participants who were carriers of the gene. 'Our correct data indicate that howest applied owning has harreful affects on the brain, and this may be serve unusualized if there is genetic susceptibility," the researchers wrote. We therefore its not must be proporting people to direk here alcohol in the belief that they or a professing thermselves applicat formatifia-1 | 2 HERTPAGE -

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

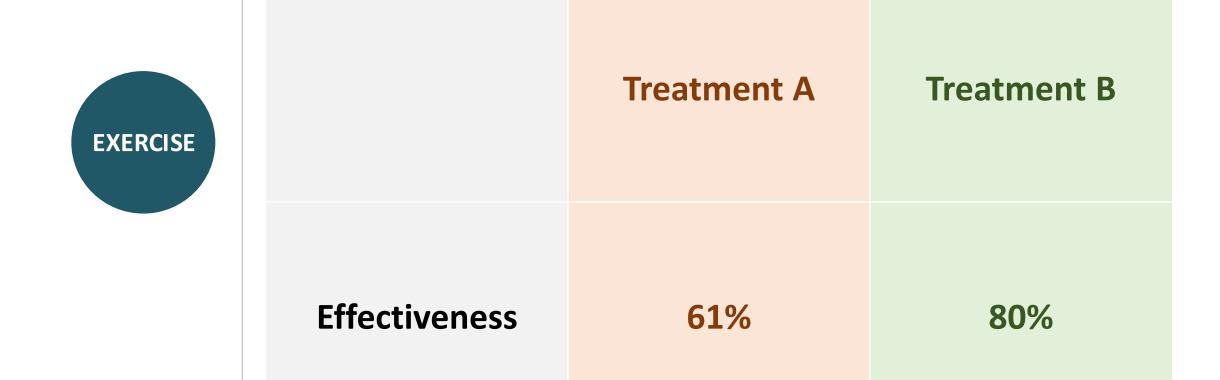
"Partial evidence may be worse than no evidence if it leads to an illusion of knowledge, and so to overconfidence and certainty where none is justified. It's better to know that you don't know" – Michael Nielsen



Refine the DataBuild 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?



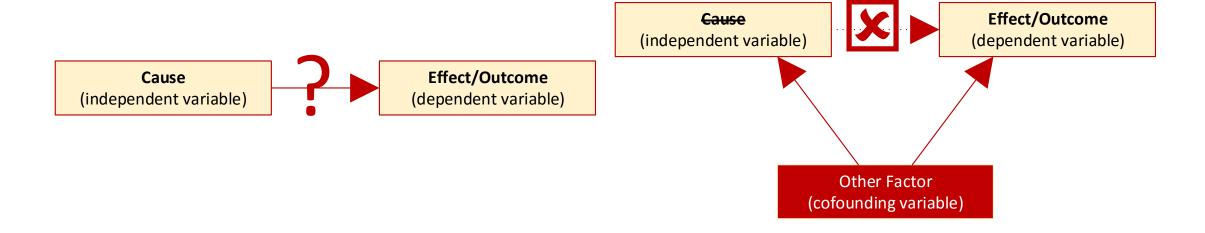
The gotcha	Treatment A	Treatment B
Patients with large kidney stones	94%	85%
Those with small kidney stones	7%	0%
Overall	61%	80%



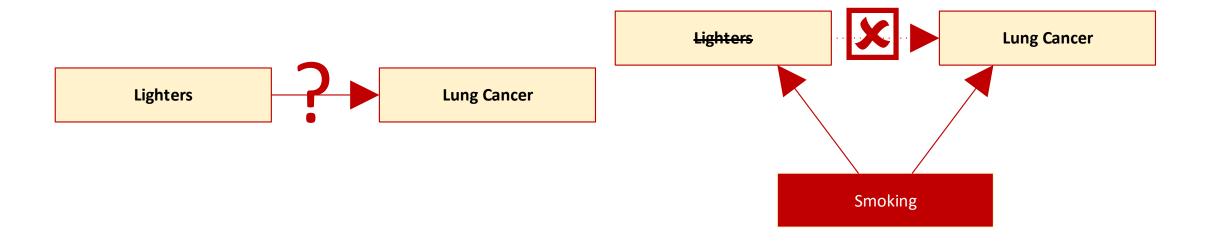
Refine the DataBuild a Model

Confounding

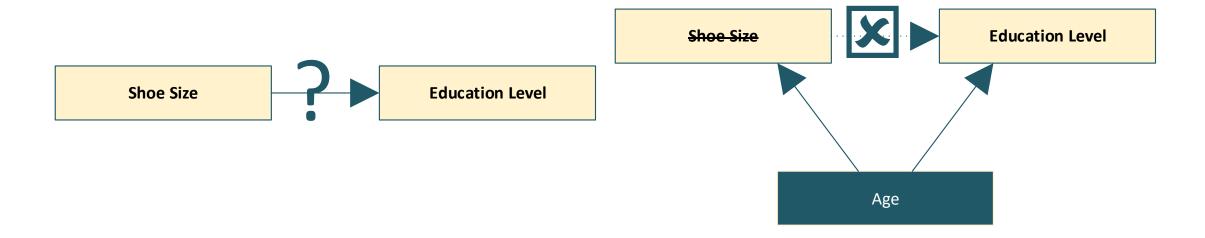
Confounding



Lighters causing lung cancer?



Shoe size as a proxy of education?

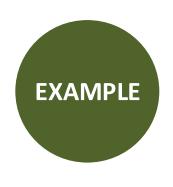




Refine the Data Build a Model

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

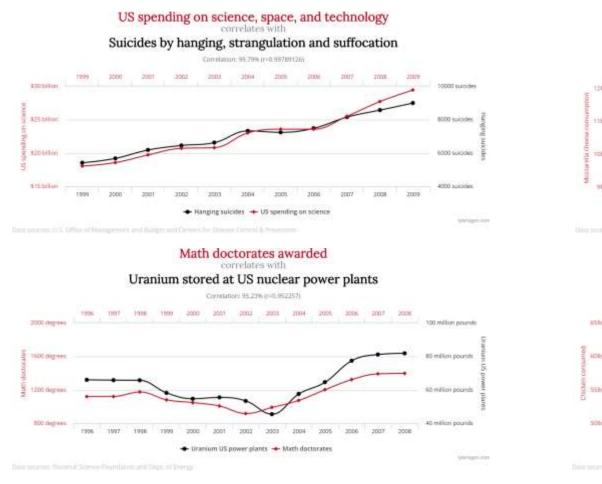
To generate personal recommendations, the service retrieves from the table the similar items lists corresponding to the items known to be of interest to the user. These similar items lists are appropriately combined into a single list, which is then sorted (based on combined similarity scores) and filtered 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.

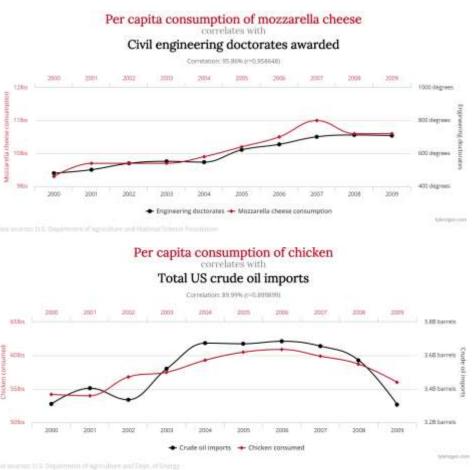


Refine the Data Build a Model

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

Spurious Correlations

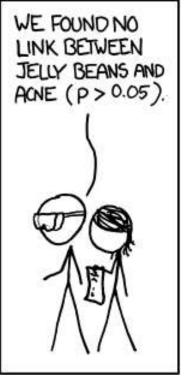


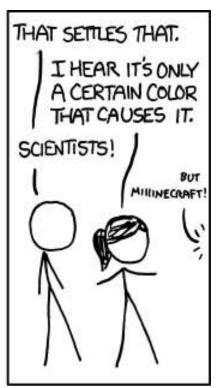


Source: tylervigen.com

Data Mining

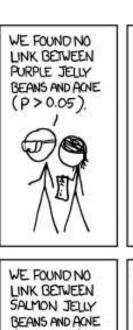




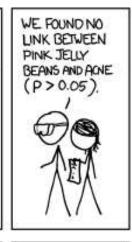


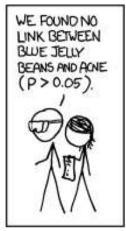
Source: xkcd.com

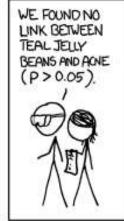
Data Mining (cont.)

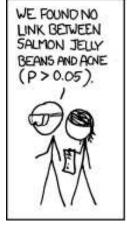




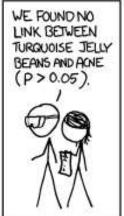


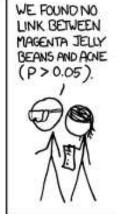


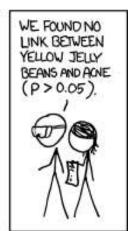






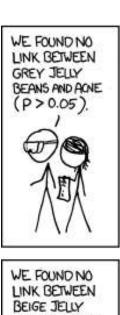


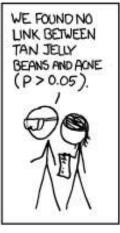


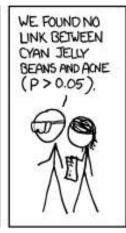


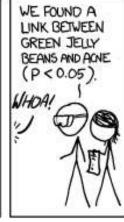
Source: xkcd.com

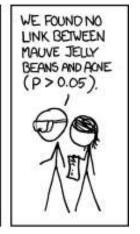
Data Mining (cont.)

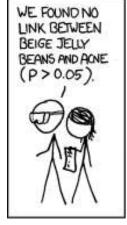


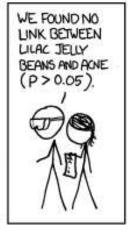


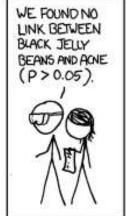


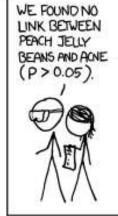


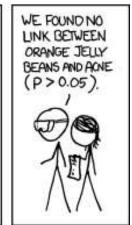






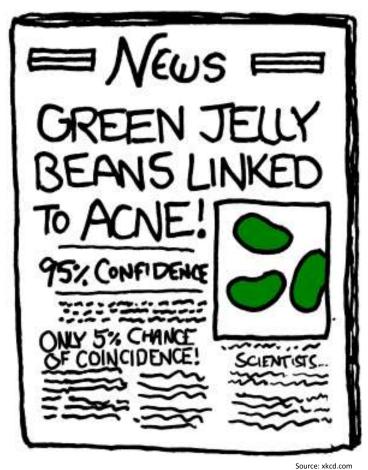






Source: xkcd.com

Data Mining (cont.)





Refine the Data Build a Model

Codealong | 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		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 Exploratory Data 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

Codealong | Your first Machine Learning Models

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

$$SalePrice = \beta_1 \cdot M1$$

```
X = df[ ['M1'] ] # X, the feature matrix, is a DataFrame
y = df.SalePrice # y, the response vector, is a Series
model = smf.OLS(y, X).fit()
```

How do we interpret these results?

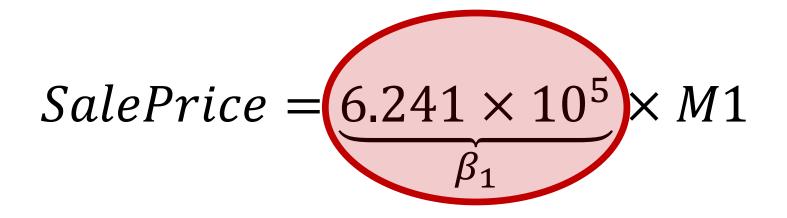
OLS Regression Results

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:		Log-L. Hhoods	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	3894.95	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'] ] # X, the feature matrix, is a DataFrame
y = df.SalePrice # y, the response vector, is a Series
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: -0.001 Model: 0.06941 Least Squares Method: F-statistic: Prob (F-statistic): 0.792 Date: Time: No. Observations: AIC: 3.207e+04 1000 BIC: **Df Residuals:** 999 3.208e+04 Df Model: Covariance Type: nonrobust

	coef	std err	t	P> t	[95.0%	Conf. Int.]
M2	3.195e+04	1.21e+b	0.263	0.792	-2.06e	+05 2.7e+05
Om	nibus:	1664.600	Duri	oin-Wa	tson:	0.971
Pro	b(Omnibus):	0.000	Jarque-Bera (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?

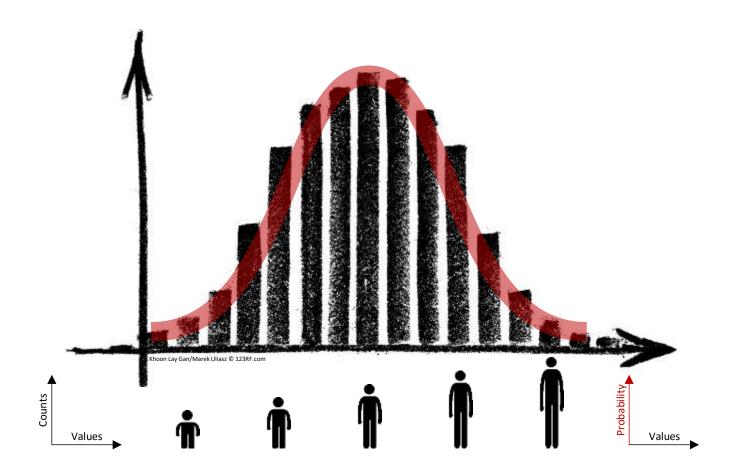
	coef	std err	t	P> t	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05
M2	3.195e+04	1.21e+05	0.263	0.792	-2.06e+05 2.7e+05



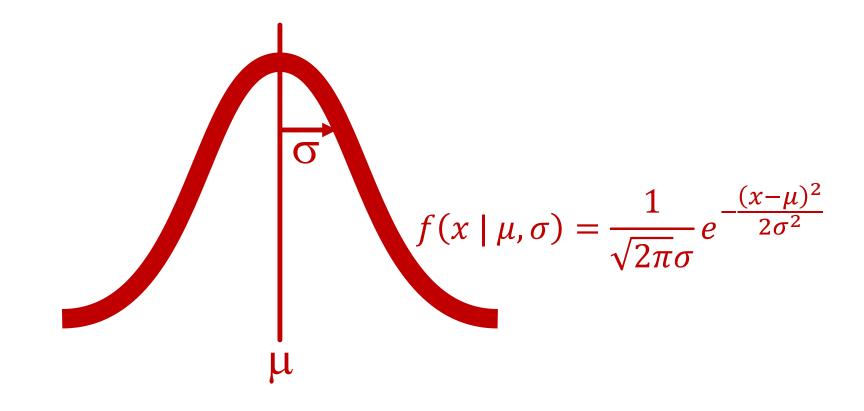
Refine the DataBuild a Model

The Normal Distribution

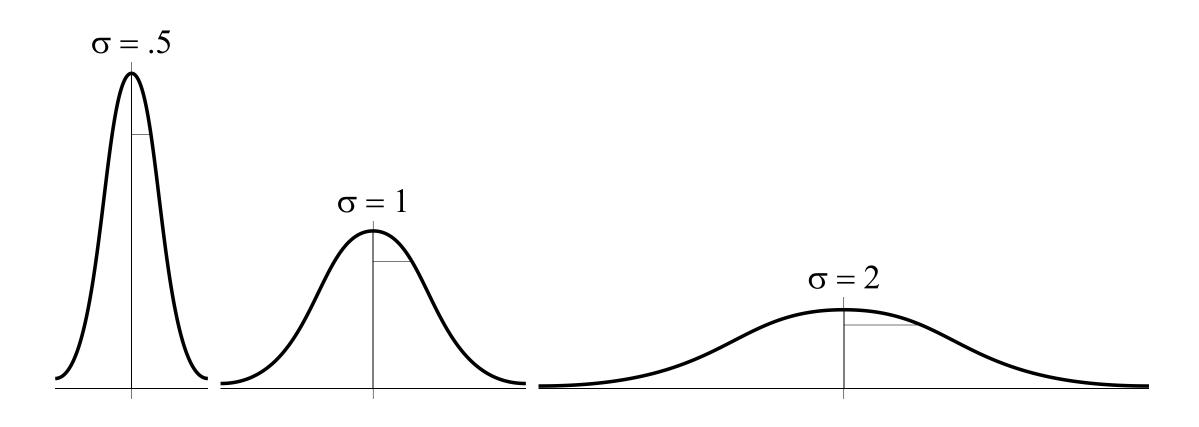
People's height follows a bell shape distribution. (For men, the average height is around 70 inches (5'10) with a standard deviation of 4 inches; few people are shorter than 67 inches; few are as tall as 73 inches)



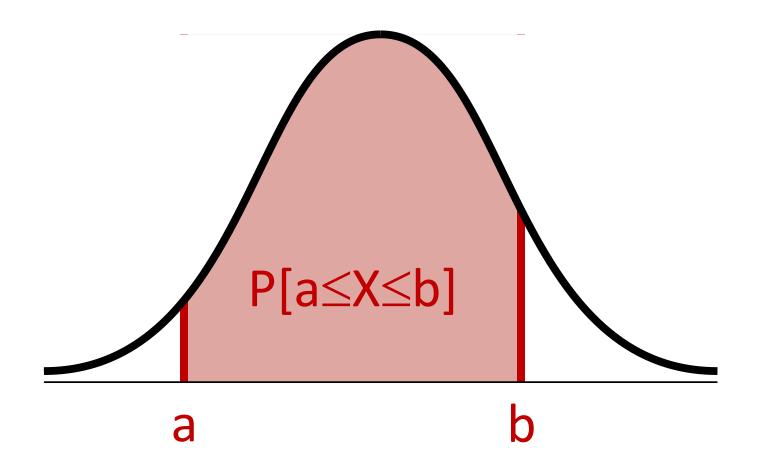
The Normal Distribution



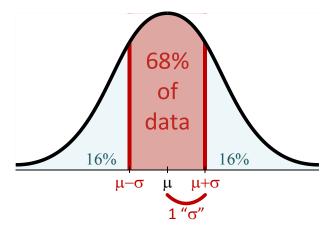
This is a probability density function: The area under the curve is always 1 (for any σ) (cont.)

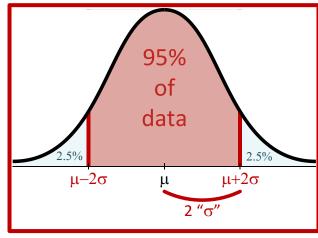


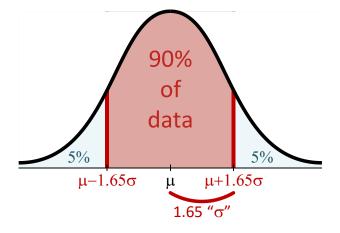
How to read a probability density function?

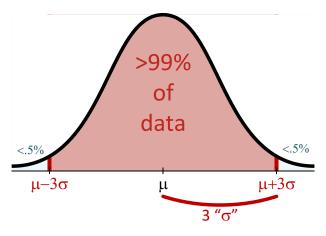


The 68 - 90 - 95 - 99.7 Rule











Refine the Data Build a Model

Activity | Knowledge Check

Activity | Knowledge Check



DIRECTIONS (5 minutes)

- 1. Adult women have an average height of 65 inches (5'5) and standard deviation of 3.5 inches. What are the lower and upper bounds for the middle 68%, 90%, 95%, and 99.7%?
- 2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Activity | Knowledge Check (cont.)



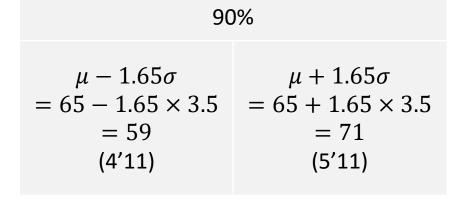
$$68\%$$

$$\mu - \sigma \qquad \mu + \sigma$$

$$= 65 - 3.5 \qquad = 65 + 3.5$$

$$= 62 \qquad = 69$$

$$(5'2) \qquad (5'9)$$



95%
$$\mu - 2\sigma \qquad \mu + 2\sigma \\
= 65 - 2 \times 3.5 \qquad = 65 + 2 \times 3.5 \\
= 58 \qquad = 72 \\
(4'10) \qquad (6'0)$$

99.7%
$$\mu - 3\sigma \qquad \mu + 3\sigma \\
= 65 - 3 \times 3.5 \qquad = 65 + 3 \times 3.5 \\
= 55 \qquad = 76 \\
(4'7) \qquad (6'4)$$



Refine the Data Build a Model

Hypothesis Testing

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

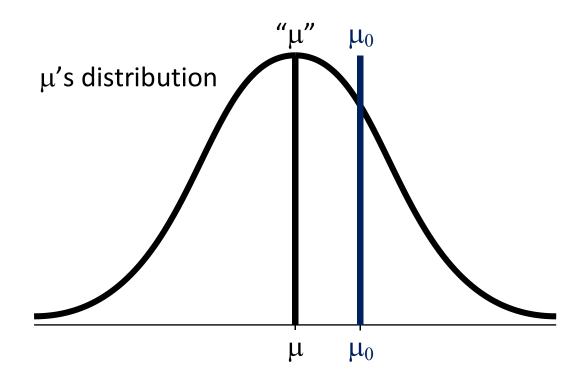
• The *null hypothesis* (H_0) represents the status quo; that the mean of the population is equal to a specific value:

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

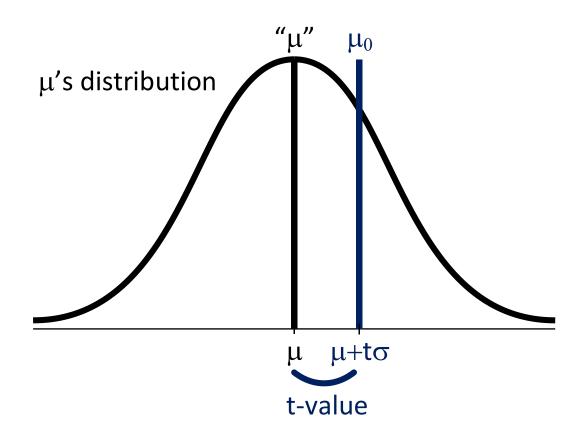
• The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis and holds true if the *null hypothesis* is found to be false:

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

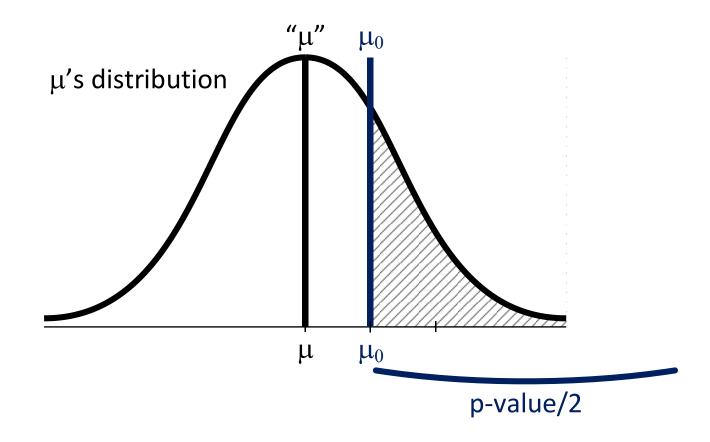
Two-Tail Hypothesis Test (cont.)



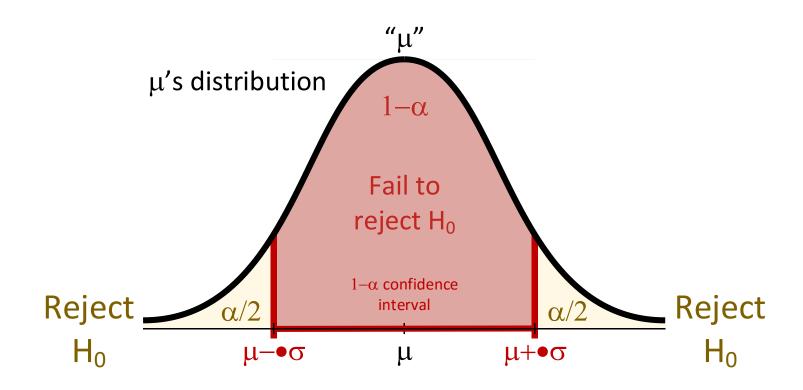
t-value measures the difference to μ_0 in σ . t-values of large magnitudes (either negative or positive) are less likely. The far left and right "tails" of the distribution curve represent instances of obtaining extreme values of t, far from μ



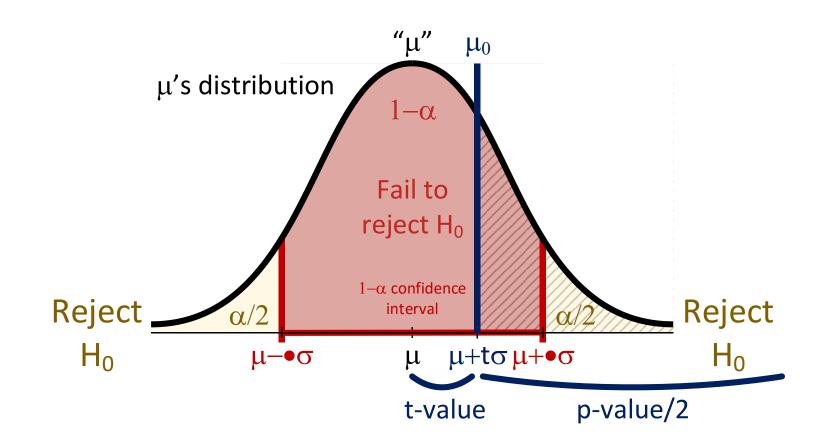
p-value determines the probability (assuming H_0 is true) of observing a more extreme test statistic in the direction of H_a than the one observed



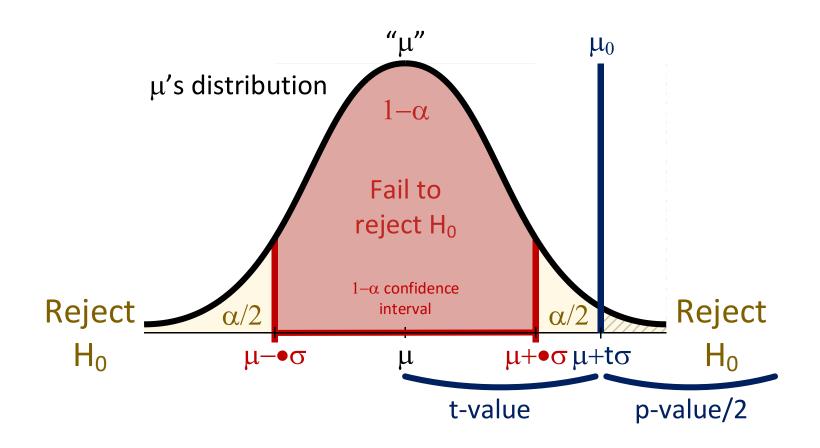
Two-Tail Hypothesis Test (simplified) (cont.)



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	
≥·	≤ <i>α</i>	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

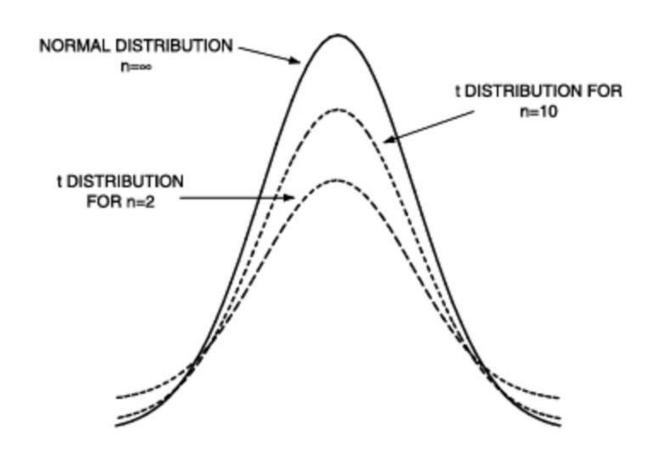
Answers to the above questions



Refine the DataBuild a Model

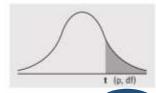
Student's t-distribution

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



Student's t-distribution table: as the sample size grows, the Student's t-distribution converges to a normal distribution

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	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



Refine the Data Build a Model

Further Readings

Further Readings

• ISLR

Assessing the Accuracy of the Coefficient Estimates (section 3.1.2, pp.

$$63 - 68)$$



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

Next Class

Introduction to Regression and Model Fit

Learning Objectives

After the next lesson, you should be able to:

- Define simple linear regression and multiple linear regression
- Build a linear regression model using a dataset that meets the linearity assumption
- Evaluate model fit
- Understand and identify multicollinearity in a multiple regression



Exit Ticket

Don't forget to fill out your exit ticket here

Slides © 2016 Ivan Corneillet Where Applicable Do Not Reproduce Without Permission