ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

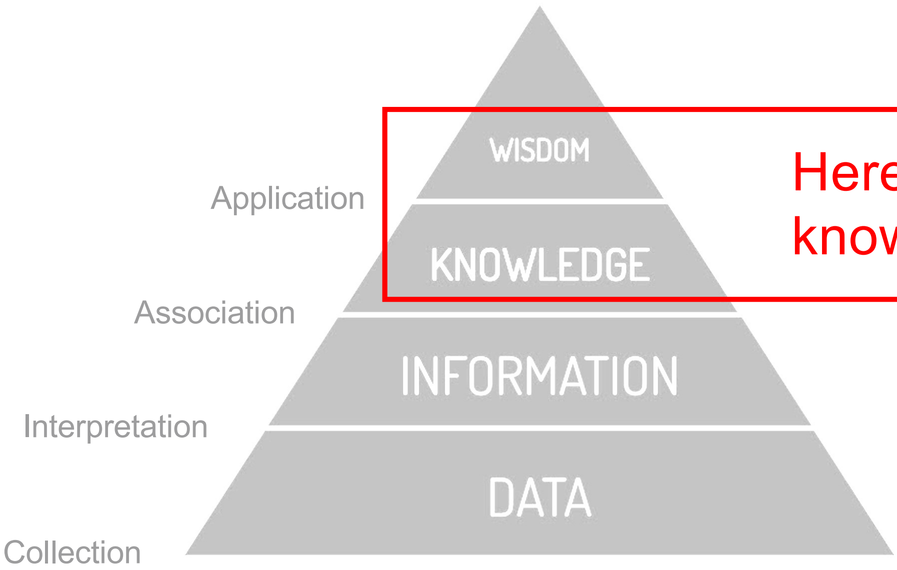
Lecture 9-10

Probability, Bayes Theorem, and Causal Analysis

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)
AI-Experience-Lab (reflect9.github.io/ael)

Things to Learn

1. Probability
2. Bayes Theorem
3. Causal Analysis



DIKW Model

Here we convert information into knowledge via **statistical inference**

Previously we have learned EDA and data visualization to convert data into meaningful information

Probability

What is Probability?

Probability of an event is calculated by **dividing the event's frequency** by the **overall observations (# trials).**

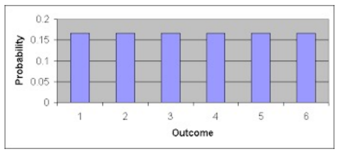
When an event is certain to happen then the probability of occurrence of that event is 1 and when it is certain that the event cannot happen then the probability of that event is 0.

E.g. Single Coin Toss


```
P(Head) = 1/2 = 0.5
P(Tail) = 1/2 = 0.5
```

E.g. Single dice roll

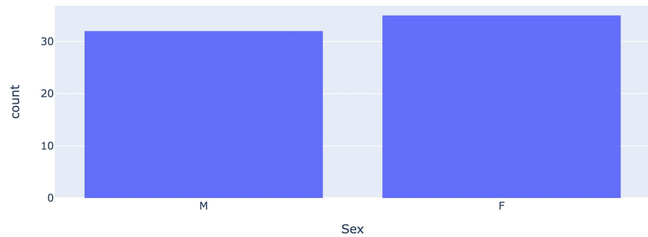
```
P(1) = 1/6
P(2) = 1/6
P(3) = 1/6
...
```



Uniform distribution

E.g. What is the probability of randomly choosing a male / a female user in the 'freshman_kgs.csv' dataset?

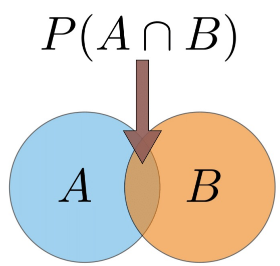
```
P(M) = #Male / #Records = 32 / 67 = 0.48
P(F) = #Female / #Records = 35 / 67 = 0.52
```



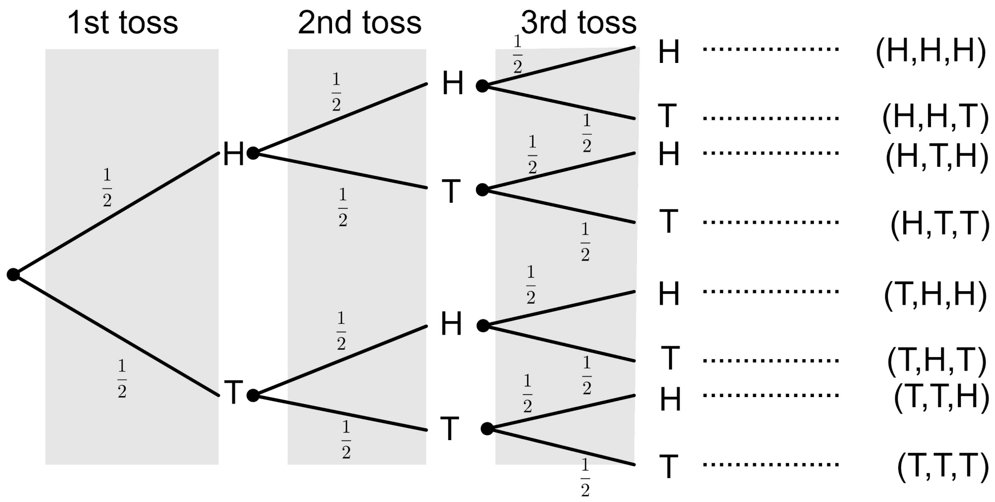
What is Compound Probability?

= How likely two (or more) events occur at the same time

= P(A,B) = P(A) * P(B)



E.g. Triple Coin Tosses



P(H,H,H) = ½ * ½ * ½ = 1/8

1/8 for every case

...

What is Conditional Probability?

P(A|B) = “The probability of A given B” = “If we observed B, how likely would we observe A?”

If a person is coughing, how likely the person would be sick?

P(Sick | Cough) = P(Cough,Sick) / P(Cough)
= (4 out of 10) / (6 out of 10) = 0.4 / 0.6 =
0.6666

If a person is sick, how likely the person would cough?

P(Cough | Sick) = P(Cough,Sick) / P(Sick)
= (4 out of 10) / (5 out of 10) =
= 0.4 / 0.5 = 0.8

Conditional Probabilty is NOT symmetrical

P(Sick | Cough) != P(Cough | Sick)

	Cough	No Cough	all	Marginal Probabilities : The probability of an event irrespective of the outcomes of other random variables.
Sick	4	1	5	P(Sick) = 5/10 = 0.5 50% of the population are sick
Not Sick	2	3	5	P(NotSick) = 5/10 = 0.5 50% of the population are not sick
all	6	4	10	P(Sick) + P(NotSick) = 1.0
	P(Cough) = 6/10 = 0.6 60% of the population are coughing	P(No Cough) = 4/10 = 0.4 40% of the population are not coughing	P(Cough) + P(NoCough) = 1.0	

Conditional Probability is the core logic of most data-driven features including recommendation engine, personalization, classification, and so on

Common Operations

RULE OF PRODUCT

P(A, B) = P(A | B) * P(B)

E.g.

P(Cough, Sick) = P(Cough | Sick) * P(Sick)
= (4/5) * (5/10)
= 4/10

From the table we can directly see **P(Cough,Sick) = 4/10**

RULE OF NEGATION

P(not A) = 1- P(A)

E.g.

P(NoCough) = 1 – P(Cough)
= 0.4

"Everyone is either coughing or not coughing"

"Everyone is either sick or not sick"

	Cough	No Cough	all	Marginal Probabilities : The probability of an event irrespective of the outcomes of other random variables.
Sick	4	1	5	P(Sick) = 5/10 = 0.5 50% of the population are sick
Not Sick	2	3	5	P(NotSick) = 5/10 = 0.5 50% of the population are not sick
all	6	4	10	P(Sick) + P(NotSick) = 1.0
	P(Cough) = 6/10 = 0.6 60% of the population are coughing	P(No Cough) = 4/10 = 0.4 40% of the population are not coughing	P(Cough) + P(NoCough) = 1.0	

Bayes Theorem

Bayes Theorem

Thomas Bayes: I am a 18th-century British mathematician. I have never published the theorem, but Pierre-Simon Laplace discovered and published after my death

P(A, B) = P(B, A) ← **Rule of Symmetry**

P(A|B) $\boxed{\text{* P(B)}}$ = P(B|A) * P(A) ← **Rule of Product**

P(A|B) = P(B|A) * P(A) $\boxed{\text{/ P(B)}}$

Remember?

Conditional Probabilty is NOT symmetrical

P(Sick | Cough) != P(Cough | Sick)

P(Sick | Cough) = P(Cough | Sick) * P(Sick) / P(Cough)
= (4/5) * (5/10) / (6/10)
= 0.4 / 0.6 = 66.6666%

Assume this is what we want to get

	Cough	No Cough	all
Sick	4	1	5
Not Sick	2	3	5
all	6	4	10

Marginal Probabilities
: The probability of an event irrespective of the outcomes of other random variables.

P(Sick) = 5/10 = 0.5
50% of the population are sick

P(NotSick) = 5/10 = 0.5
50% of the population are not sick

P(Sick) + P(NotSick) = 1.0

P(Cough) = 6/10 = 0.6

60% of the population are coughing

P(No Cough) = 4/10 = 0.4

40% of the population are not coughing

P(Cough) + P(NoCough) = 1.0

We can directly get the same result from the contingency table. In fact, filling numbers in the contingency table from probabilities is basically the same job as plugging probabilities into the Bayes Theorem.

In many other examples,
cough = mamogram
sick = cancer

E.g. Does cough indicate sickness?

A tool to update our **prior** knowledge with **observations** so that we get **posterior** knowledge

P(Sick) = 0.01

Prior
External sources told us that 1% of the entire population might get sick

P(Cough | Sick) = 0.95

Likelihood
95% of people who got confirmed to be sick were coughing.

This is the data we collect and analyze, which must be easier to acquire than the posterior

P(Sick | Cough) = 0.95 * 0.01 / 0.4 = 0.02375

Posterior (what we need)
If we see a person coughing, how likely is he/she sick? It must be higher than 1%. Using Bayes Theorem, we could get an updated knowledge “**If a person is coughing, with 2.375% chance he/she would be sick.**”

P(Cough | Sick) = 0.95

	Cough	No Cough	all
Sick	95	5	100
Not Sick	?	?	9900
all	4000	6000	10000

P(Sick) = 0.01

P(Sick | Cough)
= 95 / 4000 = 0.02375

We can get the same result by filling in the contingency table from probabilities

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

P(Cough) = 0.4

Marginalized
According to our survey, 40% of the randomly sampled population are coughing. The more people are coughing, the less likely coughing would be a good indicator of the sickness.

*A marginized probability is usually **calculated** rather than directly measured.*

E.g. Boy or Girl

Let us consider a school composed of 60% boys (B) and 40% girls (G), in which all boys have a short haircut (S) while the percentages of girls with long hair (L) and girls with short hair are equal (50/50).
We met a student with a short haircut. What is the probability that the student is a girl?

PRIOR

P(Girl) = 0.4

LIKELIHOOD

P(Short | Girl) = 0.5

POSTERIOR

P(Girl | Short) = P(S|G) * P(G) / P(S)
= 0.5 * 0.4 / 0.8
= 0.25

	Long	Short	all
Boy	0	60	60
Girl	20	20	40
all	20	80	100

P(Short | Girl) = 0.5

P(Girl) = 0.4

Among 80 students with short hair, 20 are girls. Therefore,

P(Girl | Short)
= 20 / 80
= 0.25

E.g. Will ice cream truck come today?

Let **A** represent the event that the ice cream truck is coming and **B** be the event of the weather.
Then we might ask **what is the probability of seeing the ice cream truck on a cloudy day?**

PRIOR

P(IceCream) = 0.3

LIKELIHOOD

P(Sunny | IceCream) = 0.6
P(Cloudy | IceCream) = 0.3
P(Rainy| IceCream) = 0.1

POSTERIOR

P(IceCream | Cloudy)
= P(Cloudy | IceCream) * P(IceCream) / P(Cloudy)
= 0.3 * 0.3 / 0.25
= 0.36

P(Cloudy | IceCream) = 0.3

	Rainy	Cloudy	Sunny	all
Ice cream	3	9	18	30
No Ice cream	?	?	?	70
all	25	25	50	100

P(IceCream) = 0.3

To calculate the posterior, we need the **marginalized probability** of cloudy weather P(Cloudy) – which is quite easy to get from weather statistics. Let's say the weather is cloudy with 25% chance in the region

E.g. Spam Filter

Let’s say a user has a list of confirmed spam emails, which is 80% of the entire received emails. 40% of the spams contain “free”. What is the probability of an email to be spam given that the email contains “free”.

PRIOR

$$P(\text{Spam}) = 0.8$$

LIKELIHOOD

$$P(\text{“Free”} \mid \text{Spam}) = 0.4$$

POSTERIOR

$$\begin{aligned}
&P(\text{Spam} \mid \text{“Free”}) \\
&= P(\text{“Free”} \mid \text{Spam}) * P(\text{Spam}) / P(\text{“Free”}) \\
&= 0.4 * 0.8 / (0.2*0.2 + 0.8*0.4) \\
&= 0.32 / 0.36 \\
&= 0.889
\end{aligned}$$

	No “Free”	“Free”	all
No Spam	?	?	20
Spam	48	32	80
all	64	36	100

P(“Free” | Spam) = 0.4

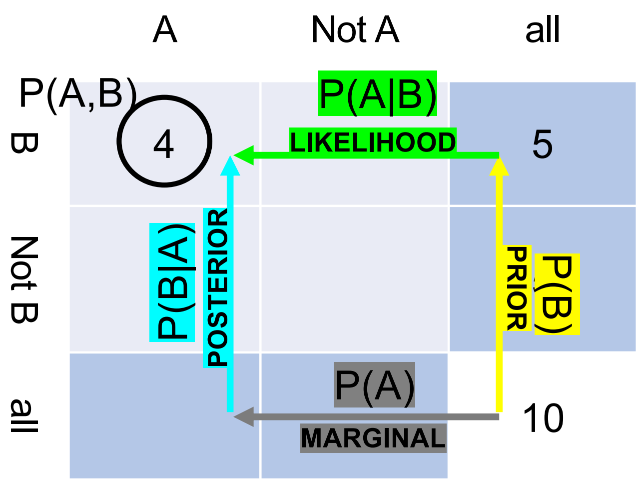
P(Spam) = 0.8

From a general email corpus, we know that “Free” would appear with 36% of documents.

P(Spam | “Free”)
= 32 / 36
= 0.889

Summary of Bayes Theorem

- **Bayes Theorem** is a method to determine **conditional probabilities**
  - Converts the likelihood (i.e. observation; P(A|B)) to the posterior (i.e. prediction; P(B|A))



P(A|B) * P(B) = P(B|A) * P(A)

Both paths reach the same compound probability
(P(A,B) = P(B,A); i.e. How likely A and B would occur)

Bayes Theorem helps calculate a missing probability based on the others.

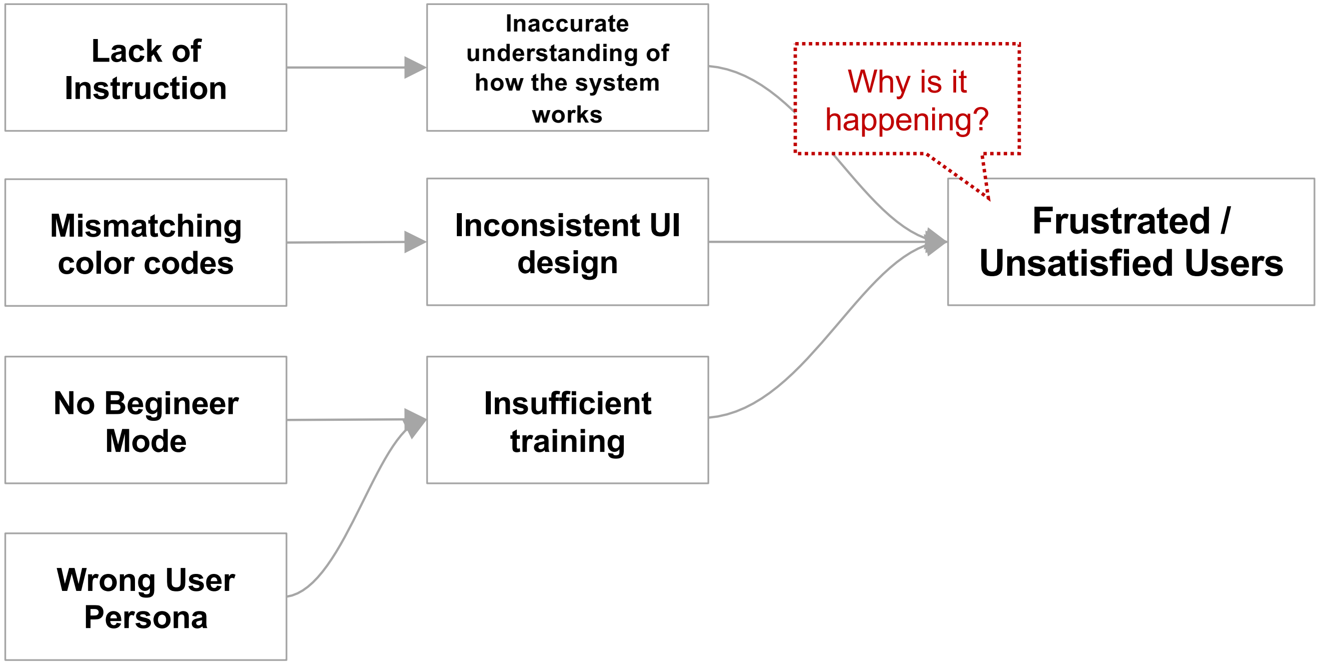
0.95 * 0.01 = P(B|A) * 0.4

P(B|A) = 0.95 * 0.01 / 0.4
= 0.02375

Causal Analysis

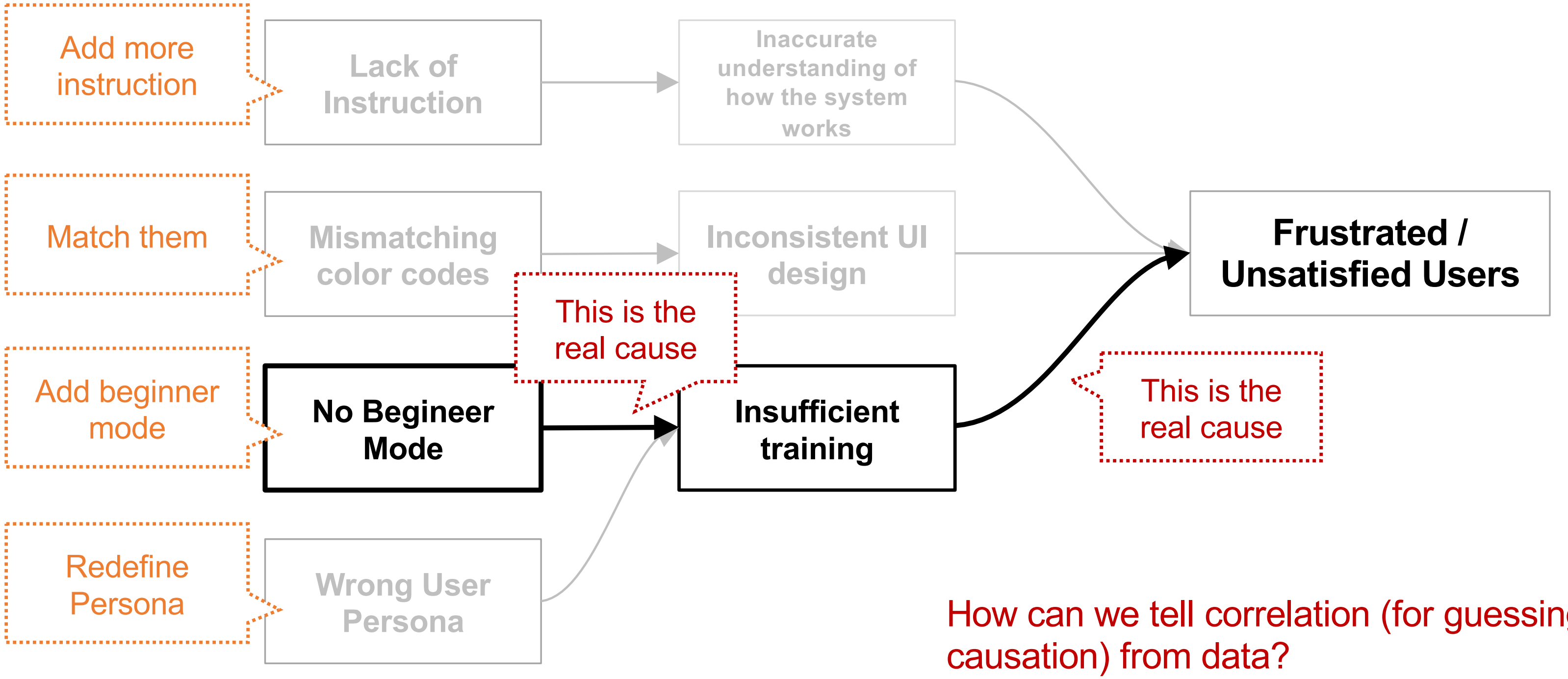
Causal Analysis in Design Process

Through observations and log analysis designer look for **reasons behind participants' behaviors**.



Causal Analysis in Design Process

Through observations and log analysis designer look for **reasons behind participants' behaviors**.



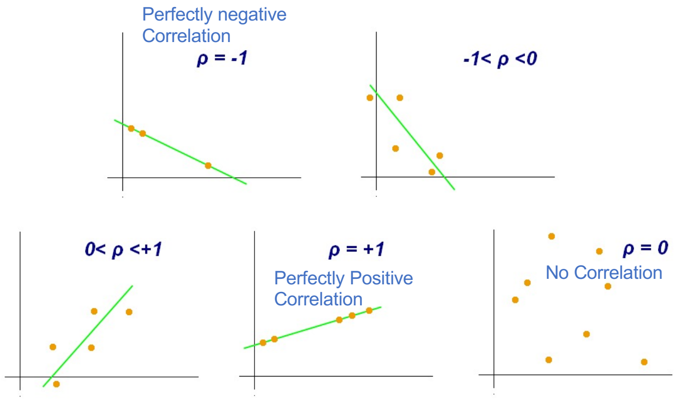
How can we tell correlation (for guessing causation) from data?

Designers try to replace bad design decisions with good ones.

Metrics for Correlation

Correlation is the best way to guess causal relationships. How can you tell column A and B are correlated?

Pearson Correlation

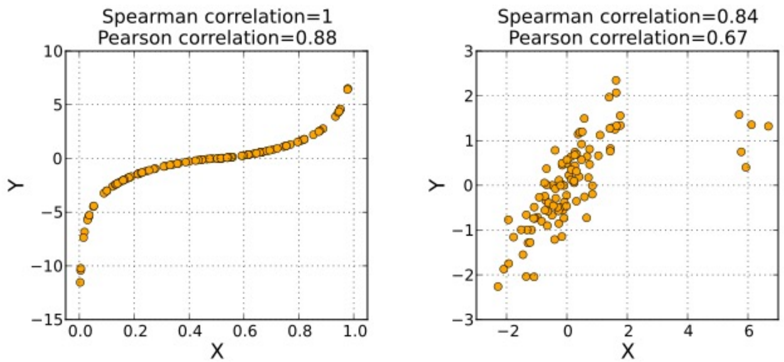



```python
import numpy as np
x_simple = np.array([-2, -1, 0, 1, 2])
y_simple = np.array([4, 1, 3, 2, 0])
my_rho = np.corrcoef(x_simple, y_simple)

print(my_rho)
```

Both columns must be **numeric**;
Relationship must be **linear** (i.e. straight line)

Spearman's Rank Correlation




```python
from scipy import stats
stats.spearmanr([1,2,3,4,5], [5,6,7,8,7])
```

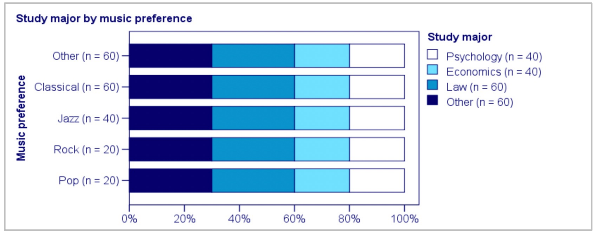
Comparing ranks on both X and Y axes;
Robust for **non-linear relationships;** Applicable for **ordinal** values

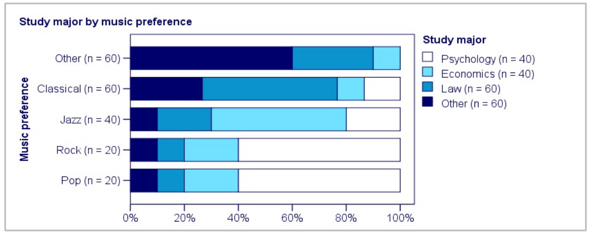
Pearson Correlation is helpless for non-linear patterns like above

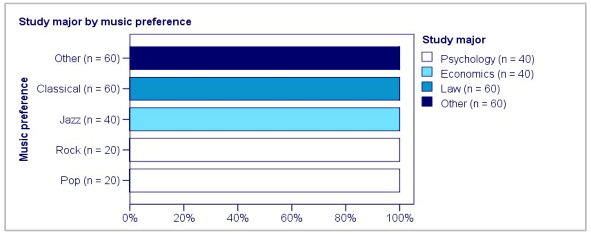
Correlation Metrics for Categorical values

Cramer's V Is there a meaningful correlation (association) between two columns?

E.g. Student's Music Preference by their majors







$\boldsymbol{\phi_c = 0,}$ when the coefficient is 0. The two factors (music preference and student majors) are independent. This means that music preference "**does not say anything**" about study major.

$\boldsymbol{\phi_c = 0.43.}$ The two factors (music preference and student majors) have moderate association. This means that music preference "**says something interesting**" about study major. For instance, 60% of all students who prefer pop music study psychology. Those who prefer classical music mostly study law.

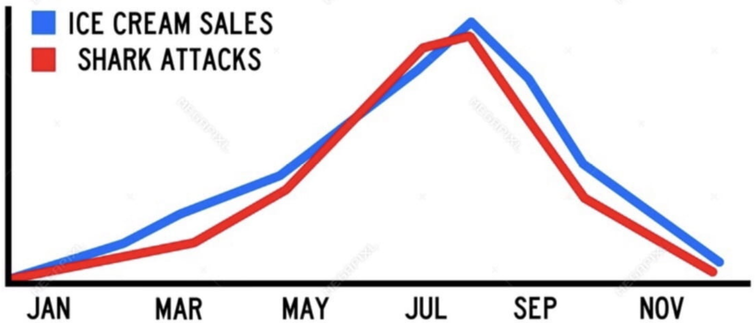
$\boldsymbol{\phi_c = 1.}$ The two factors (music preference and student majors) have perfect association. This means that music preference "**tells the student's major with confidence**". Do notice, however, that it doesn't work the other way around: we can't tell with certainty someone's music preference from his study major (e.g. Psychology students are divided into Rock and Pop). Nevertheless, this is not necessary for perfect association.

$\boldsymbol{\phi(A, B) = \phi(B, A)}$ Independence is not directional

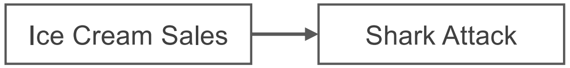
$\boldsymbol{P(A|B) \neq P(B|A)}$ Conditional probability is directional

https://www.spss-tutorials.com/cramers-v-what-and-why/

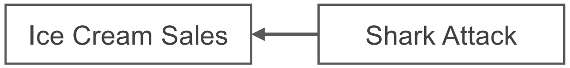
Correlation ≠ Causation



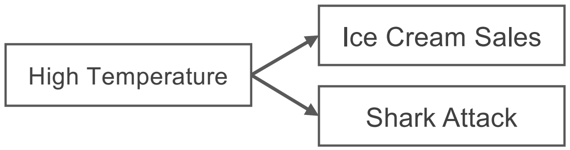
(A)



(B)



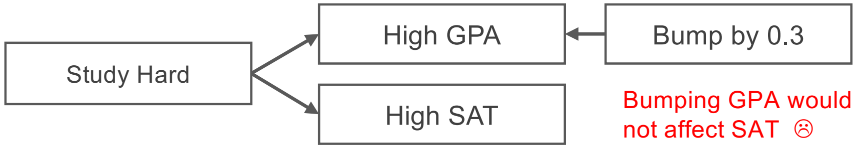
(C)



Which one is correct?

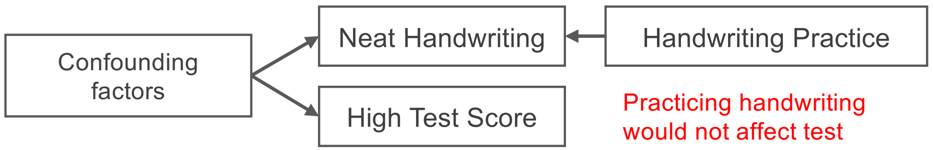
More Examples

"We just did an analysis for every student that took the SAT and there is a very clear pattern. Students with higher GPAs tended to score higher on the SAT. In order to increase our overall SAT scores at EKHS, the administration has decided to give all students with less than a 2.0 GPA a gift: bumping their GPA up by 0.3"
- Stats medic



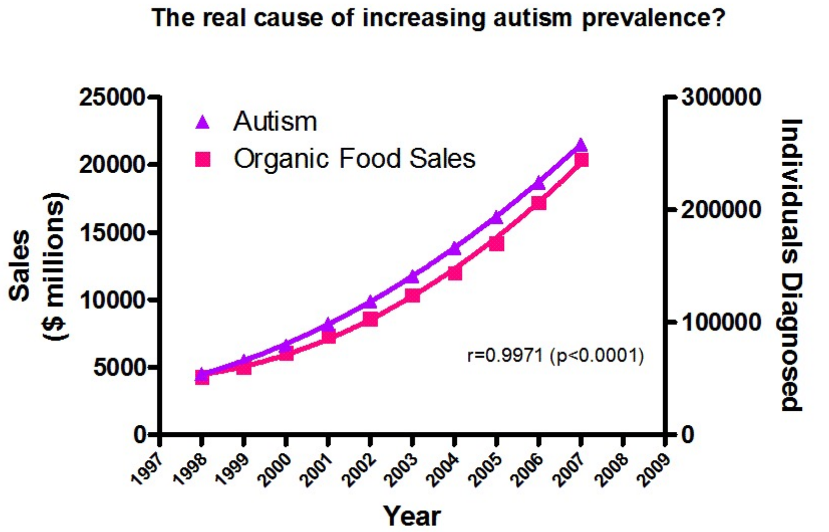
Bumping GPA would not affect SAT ☹

"Last night when I was grading your tests, I decided to give each of you a handwriting grade (1-10 with 10 being best). After grading all the tests I made a scatterplot to see if there is a relationship between quality of handwriting and test scores. There is a strong, positive, linear relationship. Therefore, we will be spending 20 minutes now every Friday practicing our handwriting"
- Stats medic



Practicing handwriting would not affect test score ☹

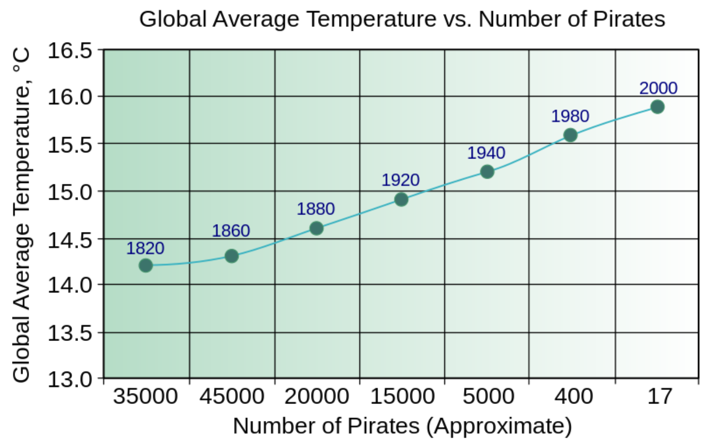
Correlation ≠ Causation



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

storks and birth rate in Denmark

priests in America and alcoholism



in the start of the 20th century it was noted that there was a strong correlation between 'Number of radios' and 'Number of people in Insane Asylums'

How silly these examples are!
But there are nontrivial cases that we daily get confused with :P

Simpson’s Paradox

UC Berkeley gender bias [edit]

One of the best-known examples of Simpson's paradox comes from a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.[13][14]

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8442	**44%**	4321	**35%**

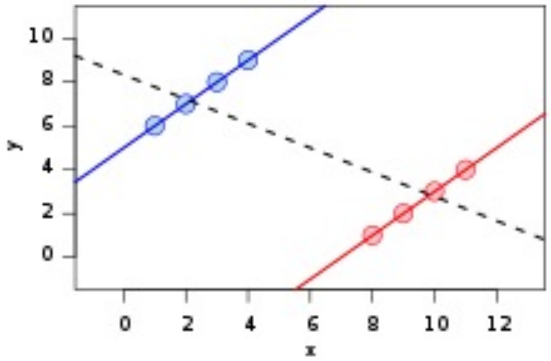
Men (44%) were more likely to get admitted than women (35%). Is this an evidence of gender inequality?

However, when examining the individual departments, it appeared that 6 out of 85 departments were signi against men, while 4 were significantly biased against women. In total, the pooled and corrected data shov statistically significant bias in favor of women".[14] The data from the six largest departments are listed belc departments by number of applicants for each gender italicised.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	***825***	62%	108	**82%**
B	585	63%	***560***	63%	25	**68%**
C	918	35%	325	**37%**	***593***	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	***393***	24%
F	714	6%	373	6%	341	**7%**
Total	4526	39%	2691	45%	1835	30%

In fact, women applied to more competitive departments.

The research paper by Bickel et al. concluded that women tended to apply to more competitive departments with lower rates of admission, even among qualified applicants (such as in the English department), whereas men tended to apply to less competitive departments with higher rates of admission (such as in the engineering department).[14]



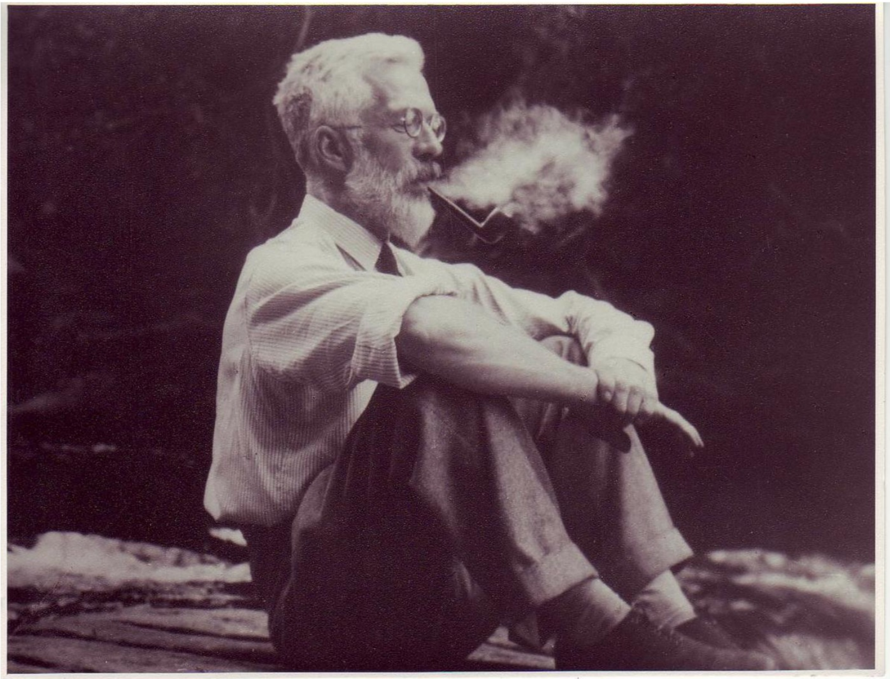
A trend appears in several groups of data but disappears or reverses when the groups are combined

Analysts gotta break down a combined group and look into individuals before making a hasty conclusion.

Why the Father of Modern Statistics Didn't Believe Smoking Caused Cancer

By Ben Christopher

Share Like 1.5K Tweet



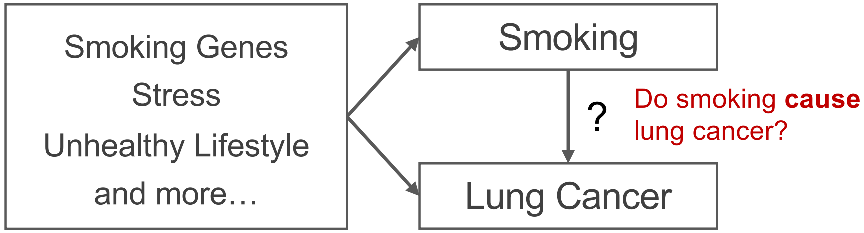
Ronald A. Fisher, father of modern statistics, enjoying his pipe.

https://priceonomics.com/why-the-father-of-modern-statistics-didnt-believe/

Counter arguments:

"Many people smoke their whole lives and never get lung cancer."

"Some people get lung cancer without ever lighting up a cigarette."



There are confounding factors of smoking and cancer

Randomized Controlled Trials (RTC) is neither feasible nor ethical

: How could you assign people at random to smoke for decades?

Cornfield's Inequality ended the debate!

"Let's say lung cancer is 9 times common among smokers compared to non-smokers. If a confounding factor (e.g. Smoking Genes) completely account for lung cancer, Smoking genes need to be at least 9 times more common in smokers than in non-smokers, which sounds very unlikely to biologists."

I.e. Now opponents cannot naively argue "there could be confounding factors" for strong correlations like smoking and lung cancer.

https://web.augsburg.edu/~schield/MiloPapers/99ASA.pdf

The Lion Man

Stadel Cave, Baden-Württemberg, Germany, 40,000 years old.

The Lion Man is the earliest representation of imaginary creature (half man and half lion). No other species has the cognitive ability to reason about something impossible (i.e. counterfactual).

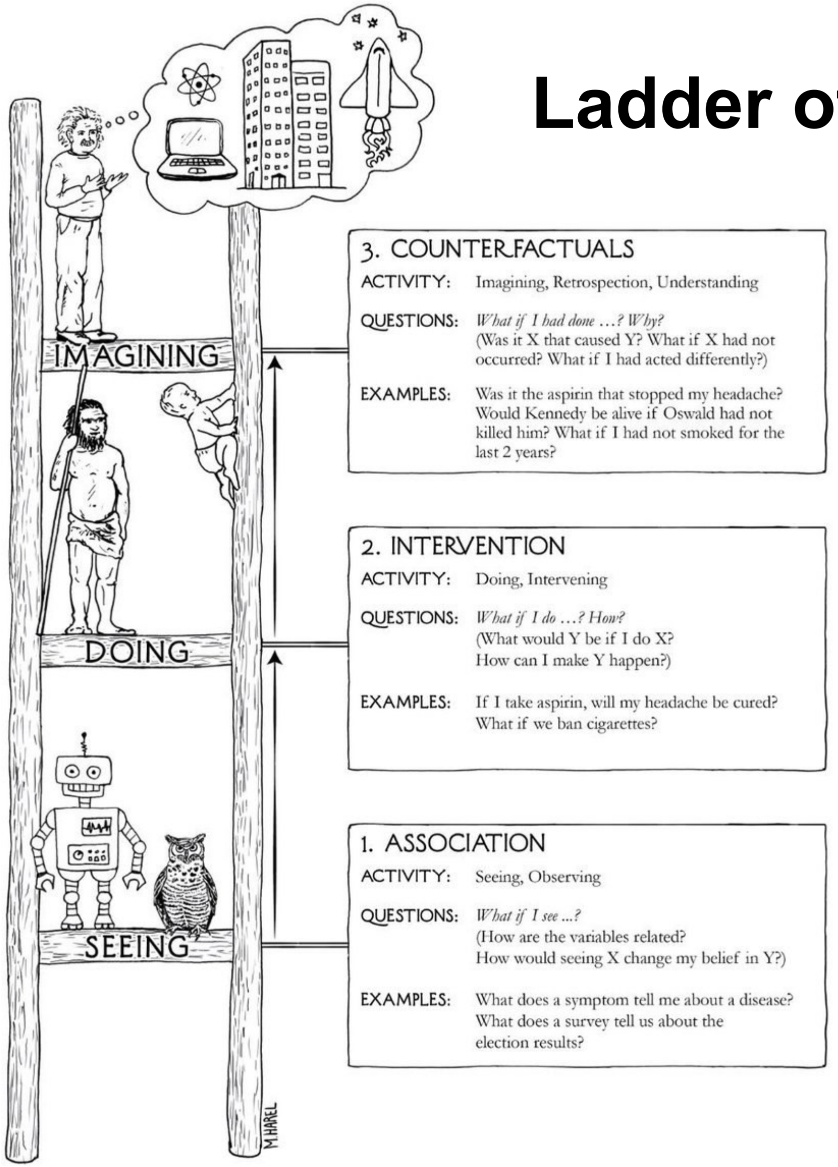
What if I were a half lion and half man?

Imagining counterfactuals is what designers daily do:

- Reflecting and improving on past actions
  What if we have done it differently?

- Flexiblity to switch between problem solving and discovery
  Why should we stick to the given problem? Let’s redefine it.

Ladder of Causation Judea Pearl



*E.g. "Was the inconsistency the actual cause of frustration? **What if we had created** two versions of the system for beginners and experts? It's too late to make such big changes though."*

*E.g. "**What if we redesign** the UI components consistent with other parts of the system? Will less users get frustrated?"*

*E.g. "During the usability test **we observed** that users got frustrated while dealing with inconsistent UI components."*

Conclusion

- **(Compound, Conditional) Probability**
  - How to use contigency table

- **Bayes Theorem**
  - Calculating posterior from other probabilities (prior, likelihood, marginal)

- **Causal Analysis**
  - Correlation metrics
    - **Pearson Correlation** for quantitative columns having linear relationship
    - **Spearman's Rank Correlation** for non-linear relationship and/or ordinal columns
    - **Cramer's V** for nominal (categorical) columns
  - Correlation is not Causation
  - Simpson's Paradox
  - Smoking and Lung Cancer (Confounding factors; Cornfield's Inequality)
  - Ladder of Causation (Association → Intervention → Counterfactuals)

(Today's content is the **primer** of the following lectures about **graph data** and **Journey Analysis**)

END