

Lots of activities in the main room  
today! No separate slide deck. 😊

# How to Lie with Statistics

CSE 312 Summer 21  
Lecture 23

# Announcements

## Upcoming Deadlines :

- Review Summary 3 – Friday, Aug 13 (TONIGHT!)
- Final Released – Friday, Aug 13 (TONIGHT!)
- Problem Set 7 – Monday, Aug 16
- Final Key Released – Tuesday, Aug 17
- Final Interviews – Wednesday - Friday, Aug 18 - 20

Office Hours will go until Wednesday

Use Ed for finals discussions exclusively! No discussion in Office Hours.

More logistics posted on Ed as a pinned post later today.

# How to Lie with Statistics – Darrell Huff

Published in 1954, over 500000 copies sold

Doesn't teach how to lie with statistics, but how we are/can be lied to using statistics

In the current age, we are lied to by the media, by politicians, and marketers.

- Often make decisions due to it: "4 out of 5 dentists recommend..."

Today's lecture is heavily inspired by the book and similar examples available on the internet.

If you like this lecture, please check out INFO 270  
(<https://www.callingbullshit.org/>)

# What is Statistics?

A way to make sense of information from data

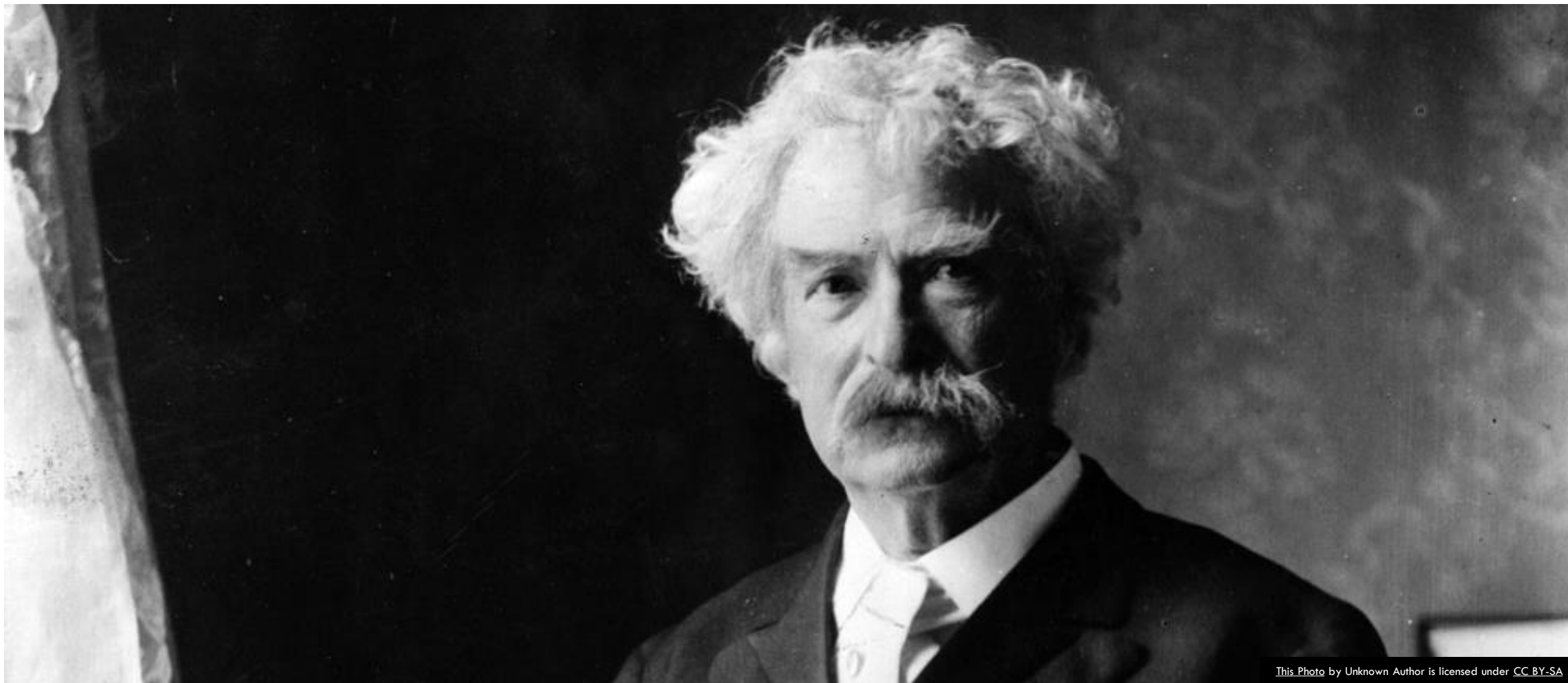
Framework for thinking, for reaching insights, and solving problems.

Numbers alone mean very little without context

Statistics is a marriage of:


- Math
- Science
- Art

“Facts are stubborn things, but statistics are pliable.”  
— Mark Twain



# Friday the 13<sup>th</sup>!



Neil deGrasse Tyson 

@neiltyson



"Friday the 13th" happens just once or twice a year.

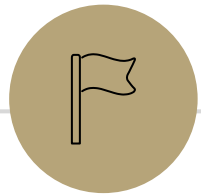
Exactly as rare as...

"Thursday the 12th" or "Saturday the 14th."

Or "Friday the 6th." Or "Friday the 20th." Or "Friday the 27th."

10:21 PM · Aug 12, 2021 · TweetDeck

**3,040** Retweets   **391** Quote Tweets   **29.1K** Likes



---

**Sampling gone wrong (bias)**

---

# Sampling Gone Wrong (Bias)

“The Literary Digest” Magazine wanted to predict the 1936 election.

- Alfred Landon vs Franklin D Roosevelt
- Sent 10 million surveys and received 2.4 million responses
- The people contacted were:
  - Subscribers of the “Literary Digest”
  - Owners of cars and telephones

Electoral Votes	Prediction	Actual
Landon	370	
Roosevelt	161	



## Topics of the day

**LANDON, 1,293,669; ROOSEVELT, 972,897**

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: “Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?” A telephone message only the day before these lines were written: “Has the Repub-

lican National Committee purchased THE LITERARY DIGEST?” And all types and varieties, including: “Have the Jews purchased THE LITERARY DIGEST?” “Is the Pope of Rome a stockholder of THE LITERARY DIGEST?” And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

**Problem**—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager \$100,000 on the accuracy of our Poll. We wired him as follows:

“For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the ‘lap of the gods.’

“We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1936:

“Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.”

In studying the table of the voters from the statistics and the material in this article are the property of Funk & Wagnalls Company and have been copyrighted by it; neither the whole nor any part thereof may be reprinted or published without the special permission of the copyright owner.



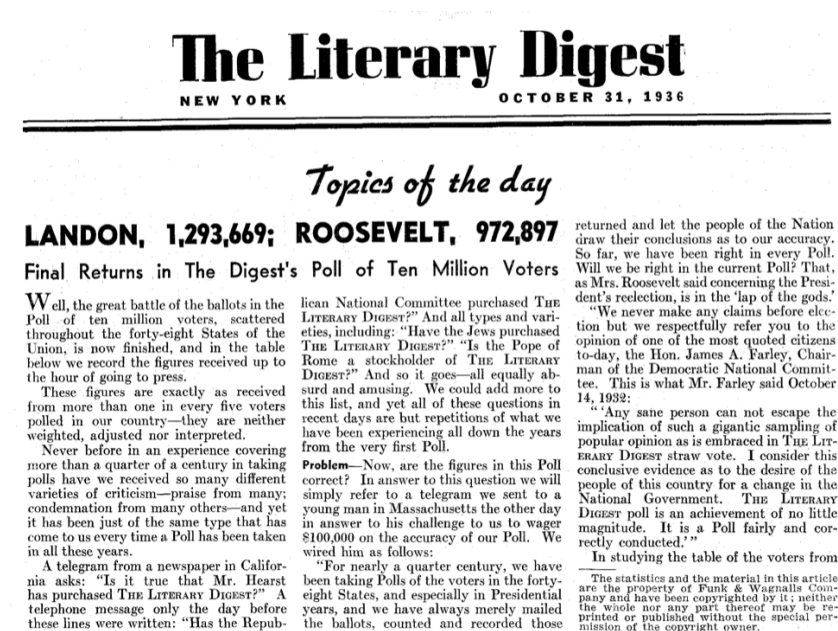
# Sampling Gone Wrong (Bias)

“The Literary Digest” Magazine wanted to predict the 1936 election.

- Alfred Landon vs Franklin D Roosevelt
- Sent 10 million surveys and received 2.4 million responses
- The people contacted were:
  - Subscribers of the “Literary Digest”
  - Owners of cars and telephones

Electoral Votes	Prediction	Actual
Landon	370	8
Roosevelt	161	523

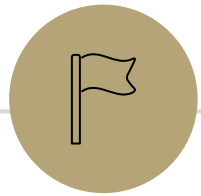
What went wrong?



# Sampling Gone Wrong (Bias)

- Not Representative
  - Voluntary Response Bias
    - Only 24% of respondents answered the poll
  - Not the Right Populations
    - Was biased towards people with more money, education, information, alertness than the average American
- Not Random
  - Convenience Sampling
    - Only people whose contact information was available
    - Standing outside a church and asking, “Do you believe in God?”, and then using the result of this sample to represent the beliefs of the entire US population.

More samples is NOT a solution for a bad sampling technique



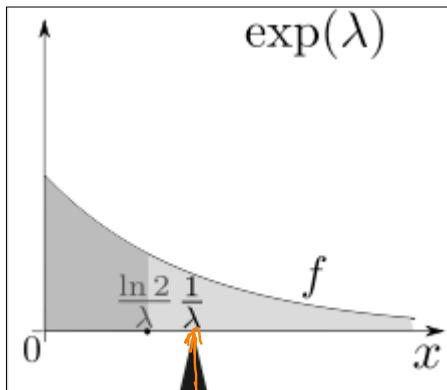
## The “Well-Chosen” Average

# The “Well-Chosen” Average

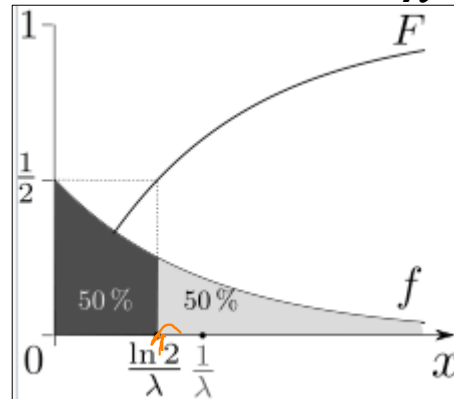
- **Mean:** Average of all values weighted by probability or density
- **Median:** The point  $m$  where  $\frac{1}{2}$  values are larger and  $\frac{1}{2}$  are smaller
- **Mode:** The point with the highest probability or density

Let  $X \sim \text{Exp}(\lambda)$ .

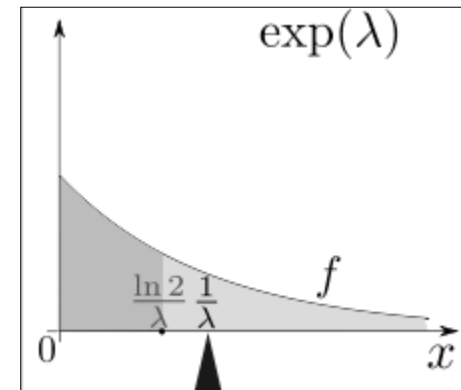
$$\mathbb{E}[X] = \frac{1}{\lambda}$$



$$\text{median}(X) = \frac{\ln 2}{\lambda}$$



$$\text{mode}(X) = 0$$



# The “Well-Chosen” Average

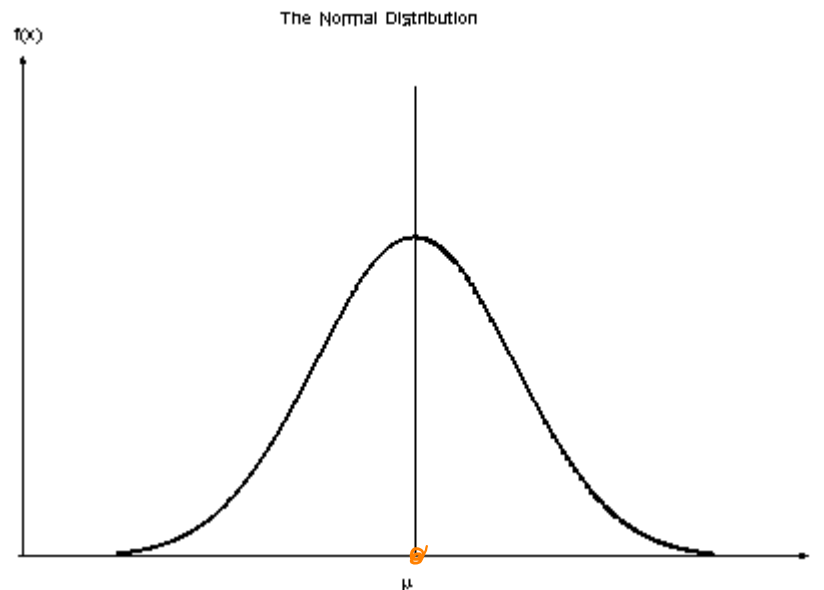
- **Mean:** Average of all values weighted by probability or density
- **Median:** The point  $m$  where  $\frac{1}{2}$  values are larger and  $\frac{1}{2}$  are smaller
- **Mode:** The point with the highest probability or density

Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

$$\mathbb{E}[X] = \mu$$

$$\text{median}(X) = \mu$$

$$\text{mode}(X) = \mu$$



# Are haircuts more expensive in Vancouver or Toronto?

Saloon	Vancouver	Toronto
1	\$20	\$15
2	\$20	\$25
3	\$22	\$25
4	\$24	\$29
5	\$25	\$35
6	\$28	\$45
7	\$400	\$65

What do you think?

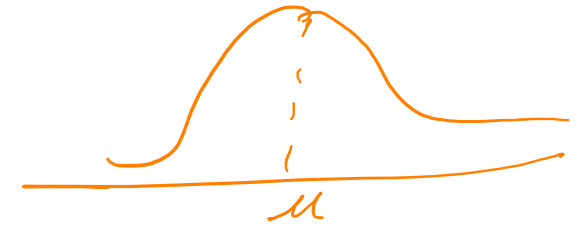
# Are haircuts more expensive in Vancouver or Toronto?

Saloon	Vancouver	Toronto
1	\$20	\$15
2	\$20	\$25
3	\$22	\$25
4	\$24	\$29
5	\$25	\$35
6	\$28	\$45
7	\$400	\$65
Mean	\$77	\$36
Median	\$24	\$29
Mode	\$20	\$25

Handwritten annotations in orange: A box around \$400, a circle around \$77, a circle around \$65, a vertical line between the Vancouver and Toronto columns, and the numbers 23 and 20 written between the lines.

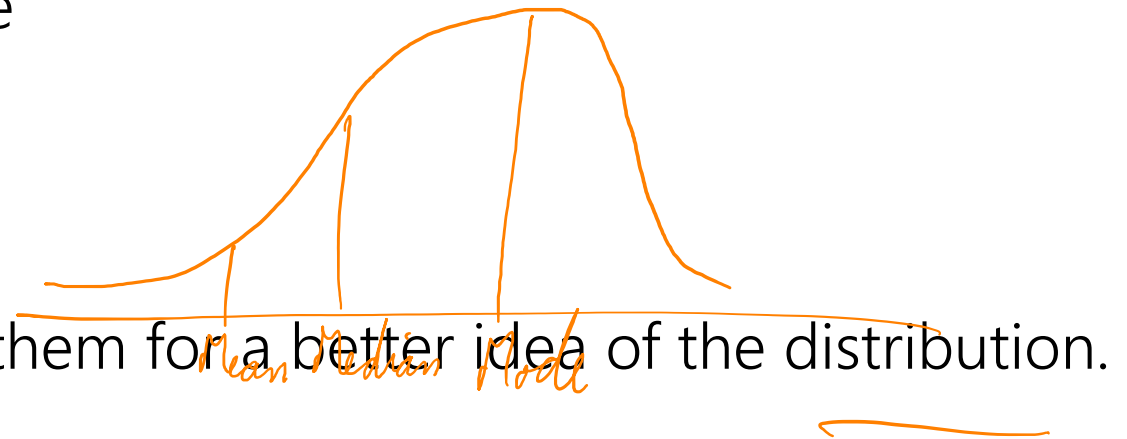
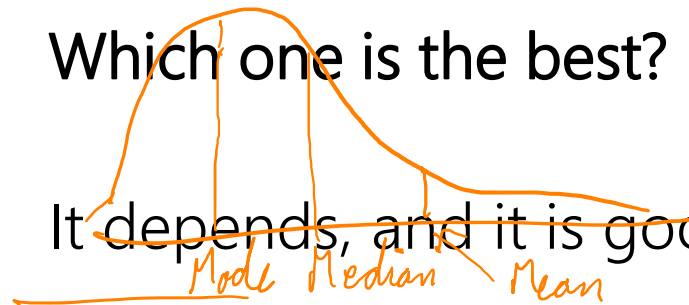
What do you think now?

# The "Well-Chosen" Average



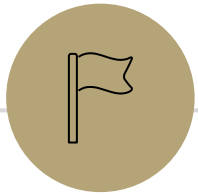
- **Mean:** Heavily affected/influenced by outliers. Any extreme value(s) may make this measure terrible
- **Median:** About half the values are higher than this, and half are lower than this
- **Mode:** Most frequently occurring value

Which one is the best?



It is good to know all - mean, median, and, mode - for a better idea of the distribution.





## **Small Sample Size**



# Sample Size Too Small



Senserdime (toothpaste company) claims 86% of dentists recommend their product.

Sounds very impressive.

Would you buy a Senserdime toothpaste?

# Sample Size Too Small



Senserdime (toothpaste company) claims 86% of dentists recommend their product.

Sounds very impressive.

86% out of how many dentists?

- $\frac{6}{7} = 86\%$
- $\frac{30}{35} = 86\%$
- $\frac{600}{700} = 86\%$

# Sample Size Too Small



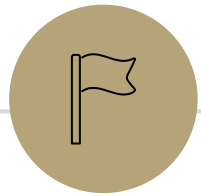
Senserdime (toothpaste company) claims 86% of dentists recommend their product.

Sounds very impressive.

86% out of how many dentists?

- $\frac{6}{7} = 86\% \rightarrow [0.7664, 0.9479]$
- $\frac{30}{35} = 86\% \rightarrow [0.8166, 0.8977]$
- $\frac{600}{700} = 86\% \rightarrow [0.8481, 0.8662]$

These are the 95% confidence intervals for the above



## Misleading results

---

# Colgate 2007 Ad Campaign

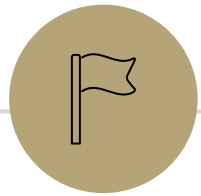
In 2007, Colgate advertised that more than 80% of dentists recommended their toothpaste.

How would you read this Ad Campaign?

- More than 80% dentists recommend Colgate over other toothpaste brands
- OR
- More than 80% of dentists recommend Colgate among other toothpaste brands

# Colgate 2007 Ad Campaign

- More than 80% dentists recommend Colgate **over** other toothpaste brands
  - ▣ This may imply that only 20% of dentists recommend toothpaste that are from brands other than Colgate
- More than 80% of dentists recommend Colgate **among** other toothpaste brands
  - ▣ This means that more than 20% of dentists recommend toothpaste that are from brands other than Colgate where a dentist can recommend more than 2 brands



**Correlation → Causation?**

---



# Correlation → Causation?



- People who use Senserdime generally have less cavities than those who use generic brands
  - Can we say "Senserdime prevents cavities"?

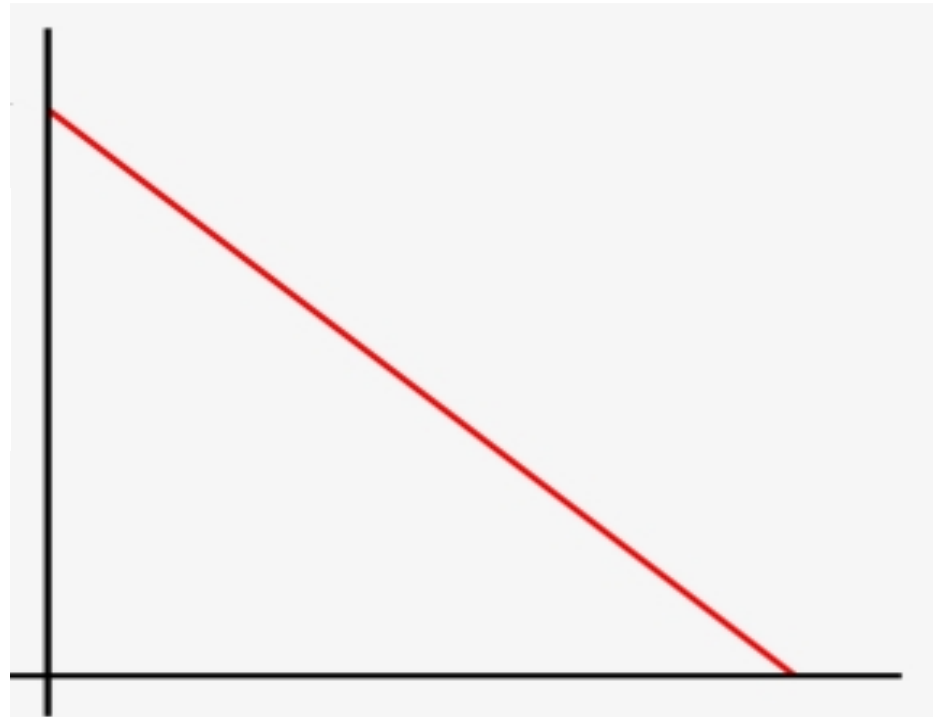
# Correlation → Causation?



- People who use Senserdime generally have less cavities than those who use generic brands
  - Can we say “Senserdime prevents cavities”?
  - Turns out that a tube of Senserdime costs \$1000.
    - This means that only wealthy people can afford it.
    - Wealthy people have access to good healthcare and hygiene
    - They are less likely to get cavities.
    - **Therefore, Senserdime did not do anything!**

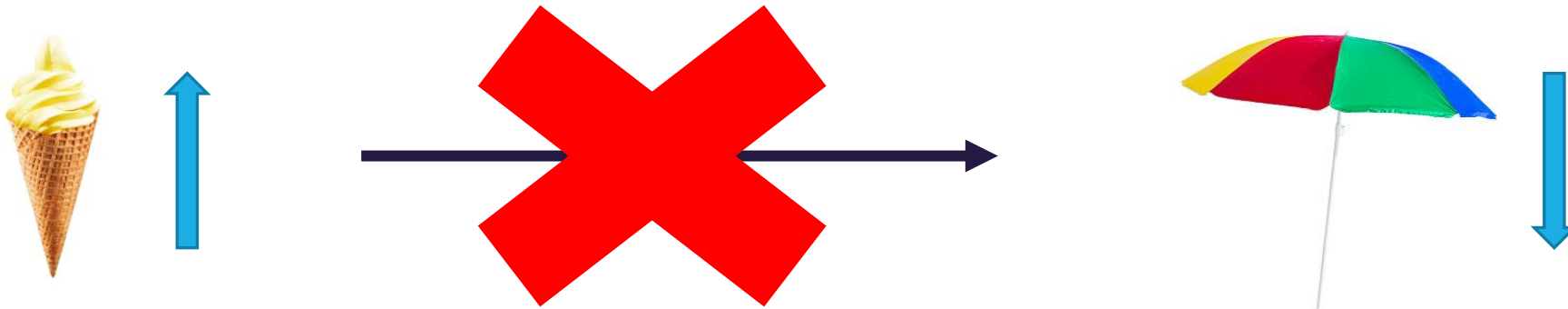
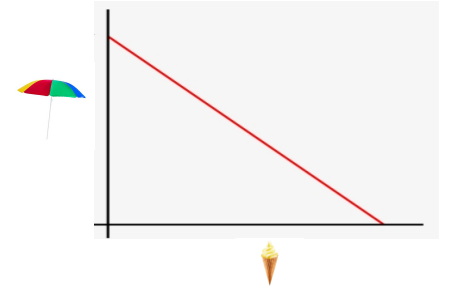
# Correlation $\rightarrow$ Causation?

- “When ice cream sales go up, umbrella sales go down”



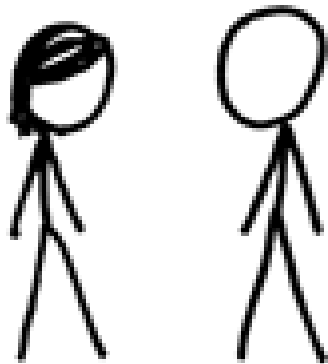
# Correlation → Causation?

- “When ice cream sales go up, umbrella sales go down”
  - Both generally happen in the summer
  - An increase in ice cream sales did not CAUSE umbrella sales to go down.
  - The weather CAUSED both of these things to happen

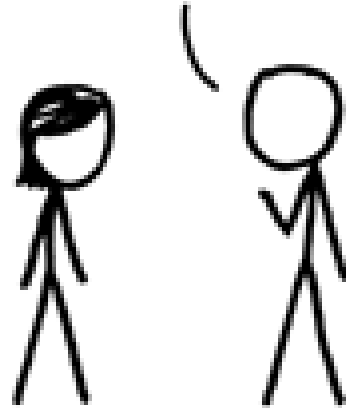


Correlation DOES NOT imply Causation!

I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.

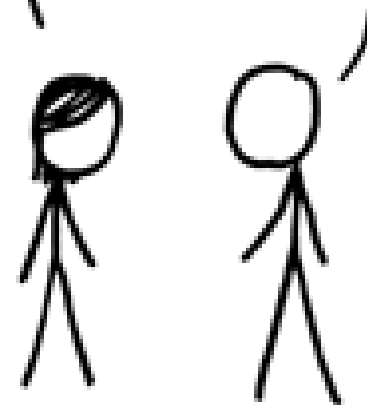


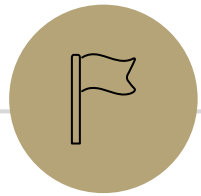
THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.

WELL, MAYBE.

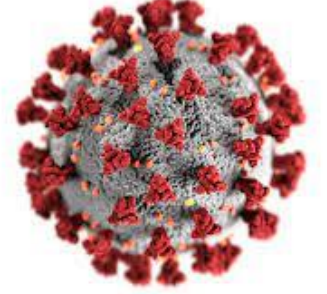




# Conditional Probability

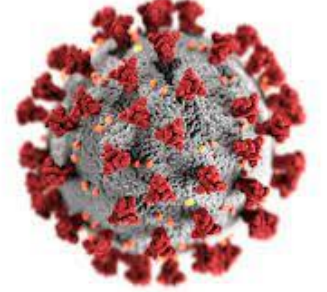


# Medical Tests



Abbott's test for COVID-19 is 99% accurate, and we know that 0.005% of the population has the disease. If you test positive, the probability you have the disease is?

# Medical Tests



Abbott's test for COVID-19 is 99% accurate, and we know that 0.005% of the population has the disease. If you test positive, the probability you have the disease is?

$$\begin{aligned}\mathbb{P}(D|+) &= \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^C)\mathbb{P}(D^C)} \\ &= \frac{0.99 \cdot 0.00005}{0.99 \cdot 0.00005 + 0.01 \cdot 0.9995} \approx \underline{0.49\%}\end{aligned}$$

Much lower than it seems at first glance!



# Biased Carnival?



Suppose there is a carnival game which gives out prizes, and three types of players: children, teenagers, and adults.

Justin thinks the carnival unfairly gives more prizes to children over the other types of players. **Is this true?**

Player Type	% Prizes Won
Child	70%
Teenager	5%
Adult	25%

# Biased Carnival?



Suppose there is a carnival game which gives out prizes, and three types of players: children, teenagers, and adults.

Justin thinks the carnival unfairly gives more prizes to children over the other types of players. **Is this true?**

Player Type	% Prizes Won
Child	70%
Teenager	5%
Adult	25%

# Biased Carnival?



Suppose there is a carnival game which gives out prizes, and three types of players: children, teenagers, and adults.

Justin thinks the carnival unfairly gives more prizes to children over the other types of players.

Player Type	% Prizes Won	% Global Population
Child	70%	25%
Teenager	5%	15%
Adult	25%	60%

How about now?

# Biased Carnival?



Suppose there is a carnival game which gives out prizes, and three types of players: children, teenagers, and adults.

Justin thinks the carnival unfairly gives more prizes to children over the other types of players.

Player Type	% Prizes Won	<del>% Global Population</del>	% Carnival Population
Child	70%	<del>25%</del>	71%
Teenager	5%	<del>15%</del>	4.5%
Adult	25%	<del>60%</del>	24.5%

This looks very fair now!

# Biased Carnival?



Player Type	% Prizes Won	<del>% Global Population</del>	% Carnival Population
Child	70%	<del>25%</del>	71%
Teenager	5%	<del>15%</del>	4.5%
Adult	25%	<del>60%</del>	24.5%

This looks very fair now!

Player Type and Prize won are (almost independent)

$$\mathbb{P}(\textit{child} \mid \textit{prize won}) = \underline{0.7}$$

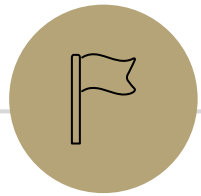
$$\mathbb{P}(\textit{teenager} \mid \textit{prize won}) = \underline{0.05}$$

$$\mathbb{P}(\textit{adult} \mid \textit{prize won}) = \underline{0.25}$$

$$\mathbb{P}(\textit{child}) = \underline{0.71}$$

$$\mathbb{P}(\textit{teenager}) = \underline{0.045}$$

$$\mathbb{P}(\textit{adult}) = \underline{0.245}$$



# Simpson's Paradox



# Simpson's Paradox

An analysis of the admission rates for the UC Berkeley grad school in 1973 is a great example of Simpson's Paradox.

	Applicants	Admitted
Men	8442	<u>44%</u>
Women	4321	<u>35%</u>
Total	12763	<u>41%</u>

Was the office of admissions unfair?

# Simpson's Paradox

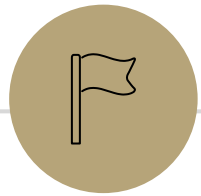
Department	Men		Women		Total	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%	933	64%
B	560	63%	25	68%	585	63%
C	325	37%	593	34%	918	35%
D	417	33%	375	35%	792	34%
E	191	28%	393	24%	584	25%
F	373	6%	341	7%	714	6%

How about now?



# Simpson's Paradox

Simpson's paradox is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.



# Gambler's Fallacy



# Gambler's Fallacy



- “Play another round of blackjack – you have to win soon! You have been losing too much!”
  - Each game is **independent**, and so even if you already lost 10 times, the probability of you winning the next game is the same as any other
  - Remember “Memorylessness” property for Geometric RV!
  - $\mathbb{P}(\text{win} \mid 1000 \text{ losses}) = \mathbb{P}(\text{win} \mid 10 \text{ losses}) = \mathbb{P}(\text{win})$

# How to better understand Statistics?

1. Who says so?
2. How do they know this is true?
3. What's missing?
4. Did somebody change the subject?
5. Does it make sense?

# Conclusions

1. Determine if the samples are random and representative.
2. Ask if the statistic represents the mean, median, or mode.
3. Inquire about the size of the sample relative to the population, and/or ask for a confidence interval.
4. Correlation does not imply causation.
5. Check the distribution of the samples (are they uniform, or not)?
6. Interpret conditional probabilities properly. Intuition sometimes doesn't work here!
7. Does the data give you the full picture? If there are subcategories, enquire into them!
8. Independent events! Don't gamble, ever.

“95.73% of all statistics are made up!”  
- Kushal Jhunjhunwala

