etherpad.wikimedia.org/p/312 for (anonymous) questions/comments!

Even *More* Applications!

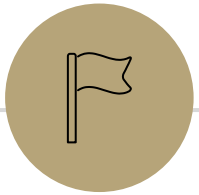
Differential Privacy, How to Lie…

CSE 312 24Su

Lecture 24

Logistics

- Ed post about course grades
- *Optional* Midterm Makeup Opportunity released, closes Friday
  (see Ed post)
- Reminder to fill out course evaluations (+1 extra point), closes tonight
- HW6 due tonight (late due date on Wednesday)
  solutions for HW6 released on Thursday morning
  your answer to Q3 can include a summation
- Review session tomorrow with Lim and Michael at 2pm
- Bring questions to Wednesday's lecture!!

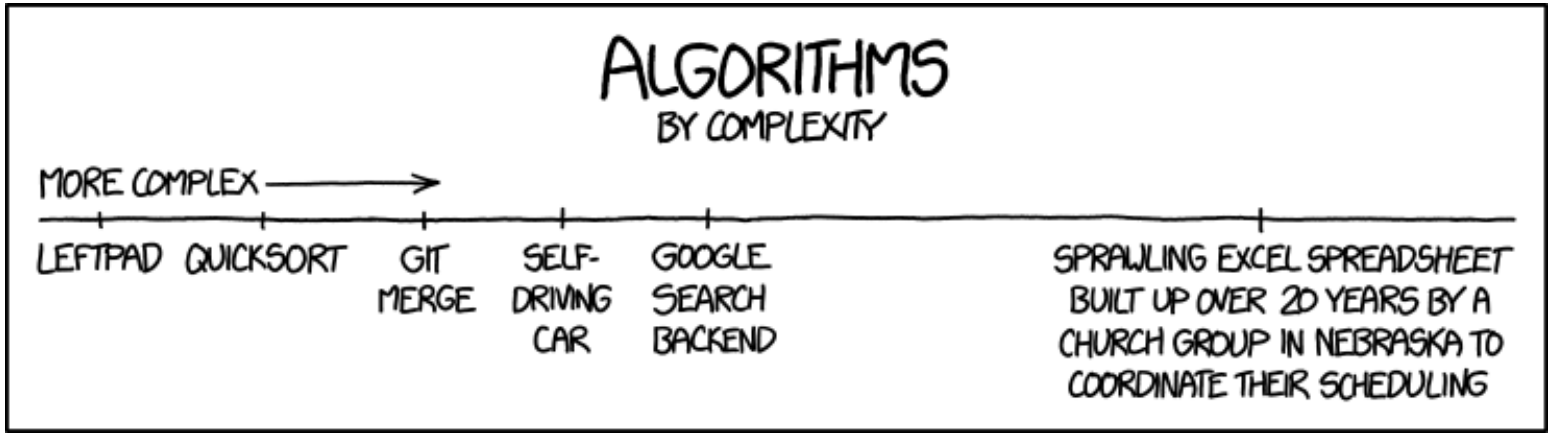


Tail Bounds *In The Wild*

Tail Bounds – Summary

- **Markov's inequality** - $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$
  - Use if X is non-negative and we know the expectation
  - Useful when we don't know much about X
- **Chebyshev's inequality** - $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var(t)}}{t^2}$
  - Use if we know the expectation **and** variance of X
  - Gives better bounds with small variances
- **Chernoff Bound**
  $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{\left(-\frac{\delta^2\mu}{2}\right)}$ *and* $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{\left(-\frac{\delta^2\mu}{3}\right)}$
  - Use if X is a sum of independent Bernoulli random variables
  - Gives a very good bound usually, and is especially helpful when X is binomial and we can't easily computationally compute some summations/probability
- **Union Bound** - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (*technically* not a *tail* bound…)
  - Use if we don't have enough information to find the union (e.g,. ways for at least of __ to occur, for A, or B, or C, or … to occur)

Algorithm Analysis



https://xkcd.com/1667

Randomized Algorithms

Randomized algorithm use *randomness* in the computation

Many algorithms incorporate some level of randomness

We can use the probabilistic techniques we've learned about in this class to analyze these algorithms!

Today…using tail bounds for analysis in randomized algorithms

Two Common Types

Las Vegas Algorithms: We will keep running the algorithm (randomly looking for the solution) till we get a good solution.

What is a bound on the running time for this?

Monte Carlo Algorithms: We will stop at some fixed number of attempts regardless of whether a good solution was found.

What is the probability a correct solution was found?

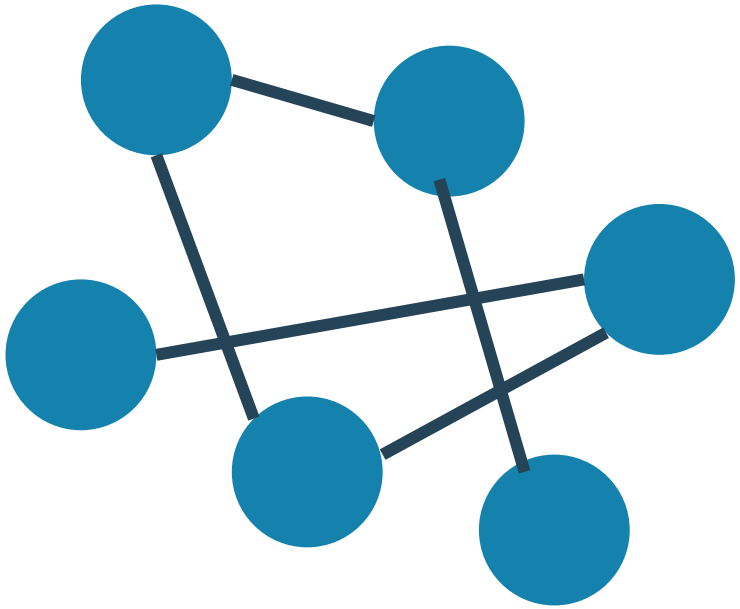
Graphs

A pair of

> Set of **vertices/nodes**

> Set of **edges** between the vertices

- ***Weighted graph***s have weighted edges
- ***Directed graphs*** have edges that either go from A to B, or B to A

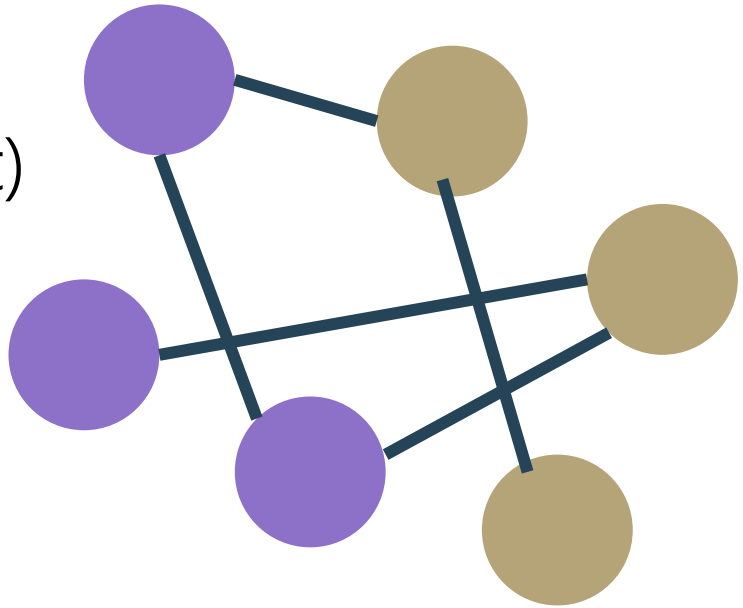


Maximum Cut Problem

The problem: partition the nodes of a graph into two sets A and B such that the number of edges between the sets is maximized

real world examples: binary classification

The *cut* is the set of edges between
the nodes in the two sets
(goal: maximize the number of edges in the cut)

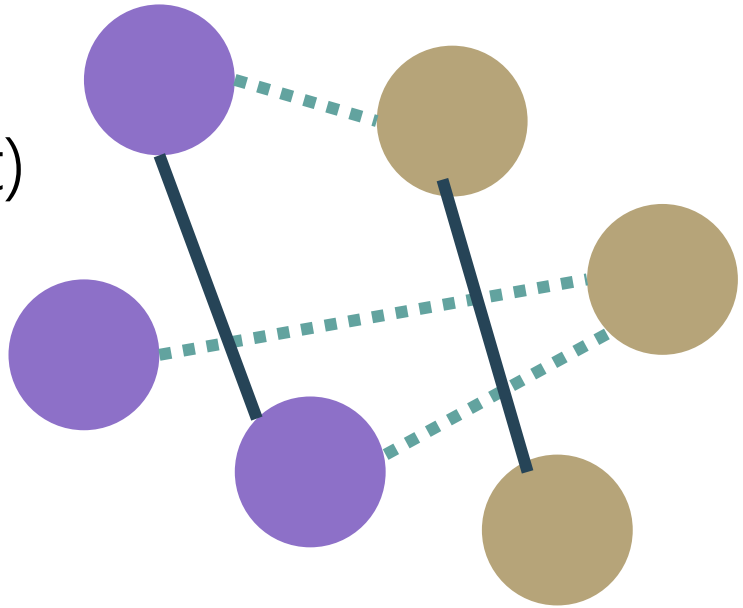


Maximum Cut Problem

The problem: partition the nodes of a graph into two sets A and B such that the number of edges between the sets is maximized

real world examples: binary classification

The *cut* is the set of edges between
the nodes in the two sets
(goal: maximize the number of edges in the cut)

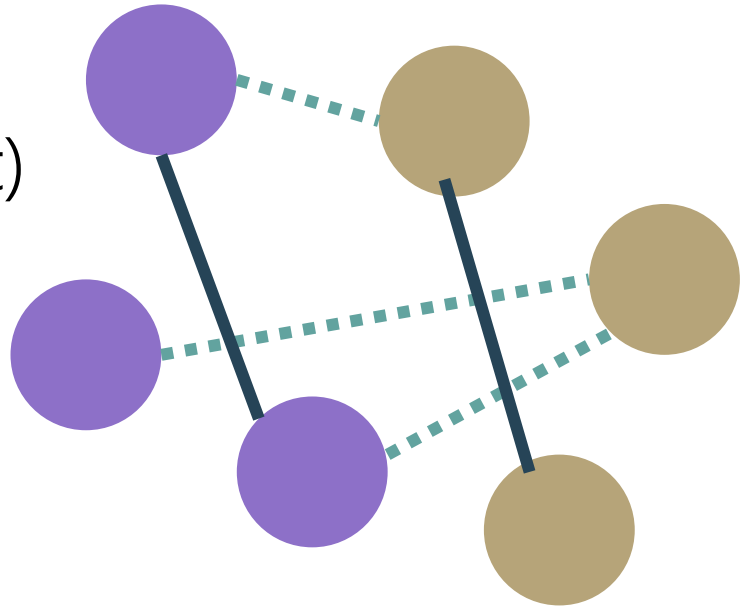


Maximum Cut Problem

The problem: partition the nodes of a graph into two sets A and B such that the number of edges between the sets is maximized

real world examples: binary classification

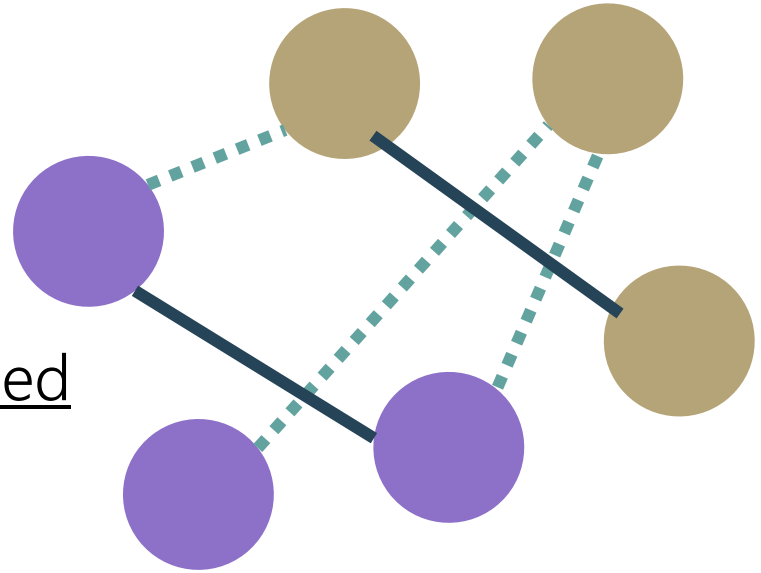
The *cut* is the set of edges between the nodes in the two sets
(goal: maximize the number of edges in the cut)



Simple randomized algorithm:

Each node goes to A or B with probability 1/2

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability 1/2

What's the probability of a "small" cut?

n is number of edges, X is number of edges in cut

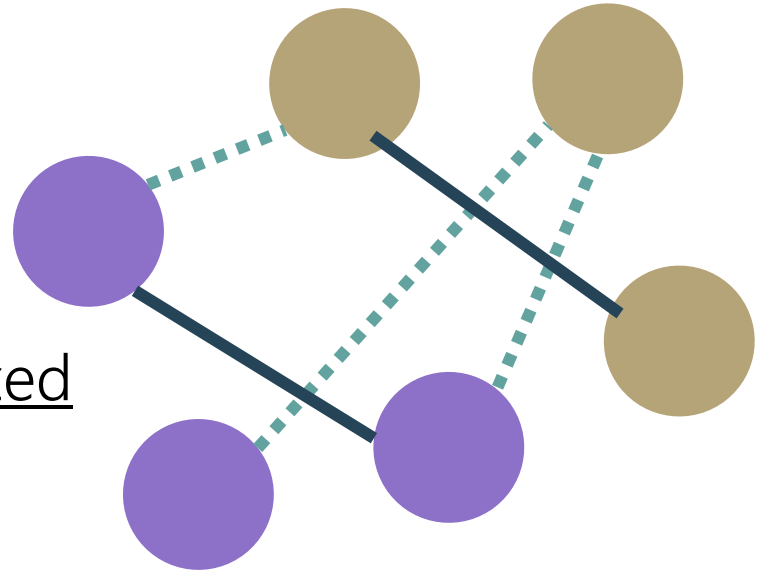
Use Markov's inequality to bound $\mathbb{P}(X \leq n/3)$

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$$

1. Find $\mathbb{E}[X]$

2. Apply Markov's Ineqality.

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability 1/2

What's the probability of a "small" cut?

n is number of edges, X is number of edges in cut

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$$

Use Markov's inequality to bound $\mathbb{P}(X \leq n/3)$

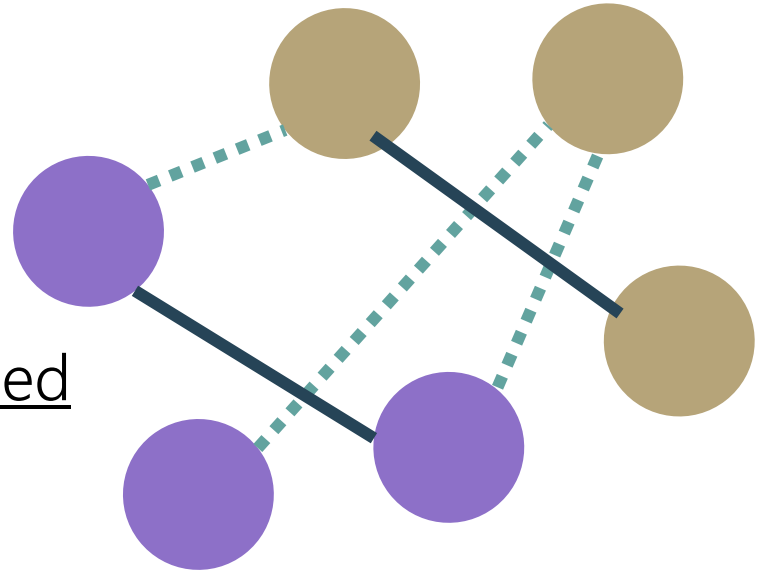
1. Find $\mathbb{E}[X]$. $X_i = 1$ of i'th edge is in the cut. $\mathbb{P}(X_i = 1) = \frac{1}{2}$

   $X = \sum_{i=1}^{n} X_i$ --> $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{2}$

2. Apply Markov's Ineqality. $\mathbb{P}(X \geq n/3) \leq \frac{n/2}{n/3}$ --> *taking complement..*

$\mathbb{P}(X \leq n/3) = 1 - \mathbb{P}\left(X \geq \frac{n}{3}\right) \geq 1 - 3/2 = -0.5$ *a trivial bound*

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability 1/2

What's the probability of a "small" cut?

n is number of edges, X is number of edges in cut

You don't need to know this trick in this class

Use Markov's inequality to bound $\mathbb{P}(X \leq n/3)$

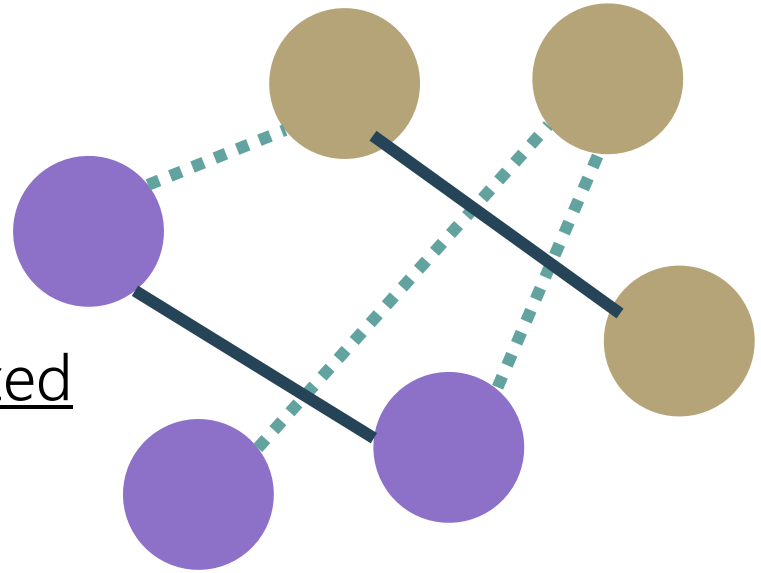
1. Find $\mathbb{E}[X]$. $X_i = 1$ of i'th edge is in the cut. $\mathbb{P}(X_i = 1) = \frac{1}{2}$

$X = \sum_{i=1}^{n} X_i$ --> $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{2}$

2. Apply Markov's Ineqality.

$\mathbb{P}(X \leq n/3) = \mathbb{P}(n - X \geq n - n/3) \leq \frac{\mathbb{E}[n-X]}{n-\frac{n}{3}} = \frac{n-n/2}{n-n/3} = \frac{n/2}{2n/3} = 3/4$ --> *a trick!*

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability 1/2

What's the probability of a "small" cut?

n is number of edges, X is number of edges in cut

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k) \leq \frac{\mathrm{Var}(X)}{k^2}$$

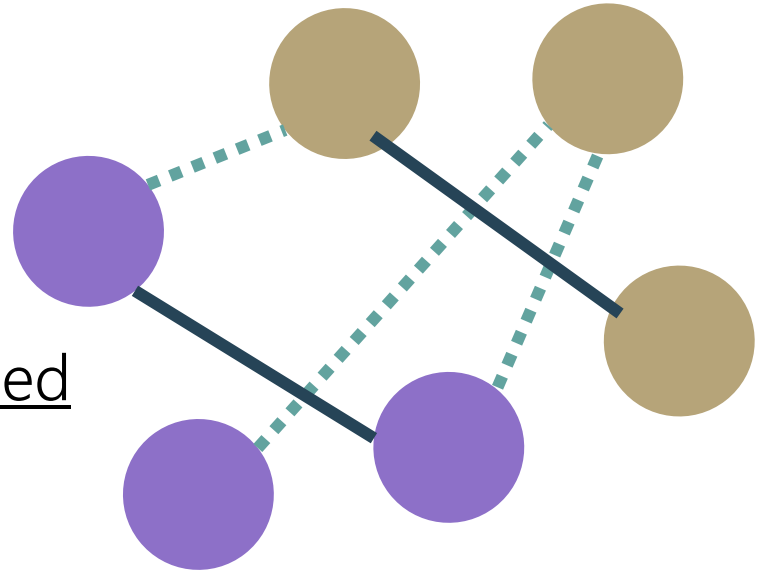
Use Chebyshev's inequality to bound $\mathbb{P}(X \leq n/3)$

1. Find $\mathbb{E}[X]$. $\mathbb{E}[X] = \frac{n}{2}$ 2. Find $\mathrm{Var}(X)$. $\mathrm{Var}(X) = \frac{n}{4}$ (see 3.2 here for explanation)

2. Apply Chebyshev's Ineqality.

$\mathbb{P}(X \leq n/3) =$

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability 1/2

What's the probability of a "small" cut?

n is number of edges, X is number of edges in cut

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k) \leq \frac{\mathrm{Var}(X)}{k^2}$$

Use Chebyshev's inequality to bound $\mathbb{P}(X \leq n/3)$

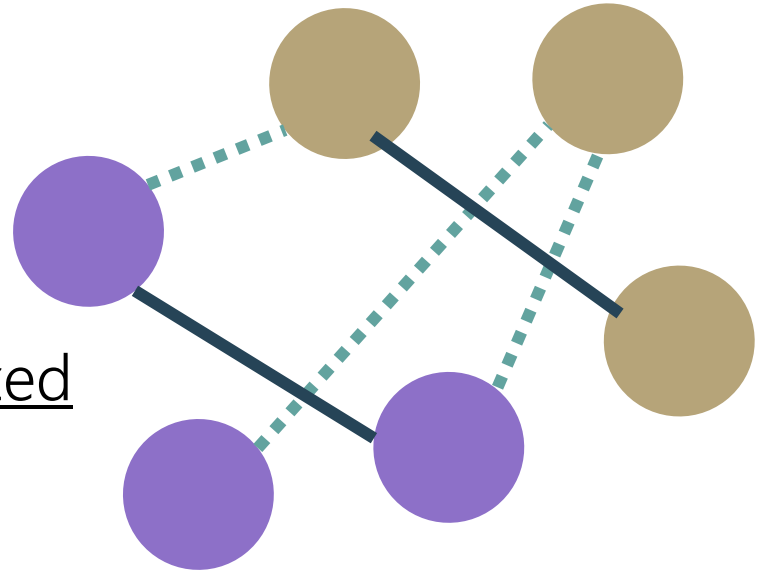
1. Find $\mathbb{E}[X]$. $\mathbb{E}[X] = \frac{n}{2}$ 2. Find $\mathrm{Var}(X)$. $\mathrm{Var}(X) = \frac{n}{4}$

2. Apply Chebyshev's Ineqality.

$$\mathbb{P}(X \leq n/3) = \mathbb{P}\left(X - \frac{n}{2} \leq \frac{n}{3} - \frac{n}{2}\right) \leq \mathbb{P}\left(X - \frac{n}{2} \leq -\frac{n}{6}\right) + \mathbb{P}\left(X - \frac{n}{2} \geq \frac{n}{6}\right)$$

$$\leq \mathbb{P}\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{6}\right) \leq \frac{n/4}{(n/6)^2} = \frac{9}{n}$$

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

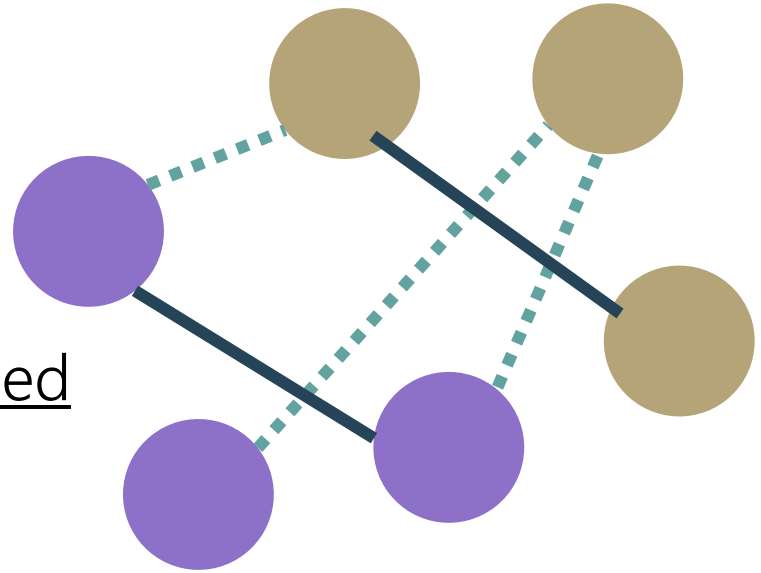
Simple algorithm:
Each node goes to A or B with probability ½

Better, Las Vegas algorithm:

Keep doing this till there is a large cut found (i.e., $X \geq n/3$)

What is the probability that in the first 20 trials, we will have succeeded?

Maximum Cut Problem



The problem: partition the vertices of a graph into two sets such that the number of edges between the sets is maximized

Simple algorithm:
Each node goes to A or B with probability ½

Better, Las Vegas algorithm:
Keep doing this till there is a large cut found (i.e., $X \geq n/3$)

What is the probability that in the first 20 trials, we will have succeeded?

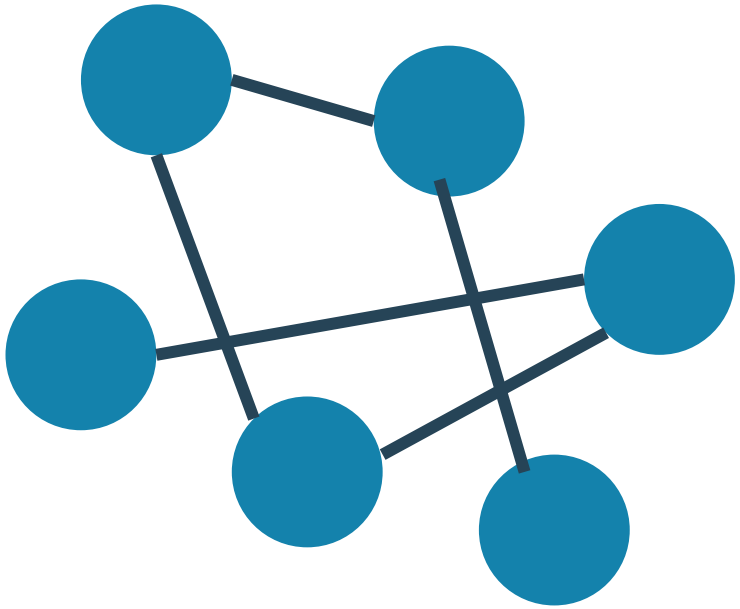
Let X be the number of trials is takes. $X \sim \text{Geo}\left(p \leq \frac{9}{n}\right)$

So, $\mathbb{P}(X \leq 20) \leq 1 - \left(1 - \frac{9}{n}\right)^{20}$

Graph Coloring Problem

The problem: color each node red, blue, or green, BUT minimize nodes with the same color sharing an edge (i.e., max. edges between distinct)

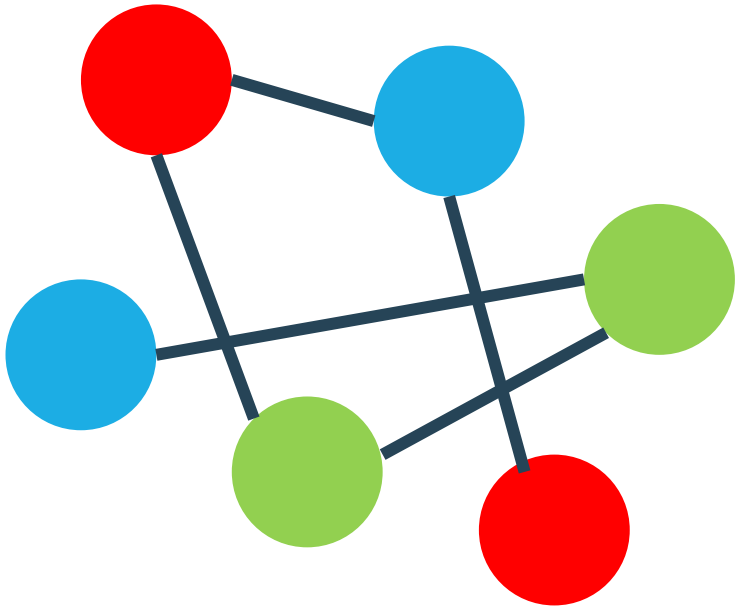
real world examples: scheduling, coloring a map, sudoku solver, CPU allocation



Graph Coloring Problem

The problem: color each node red, blue, or green, BUT minimize nodes with the same color sharing an edge (i.e., max. edges between distinct)

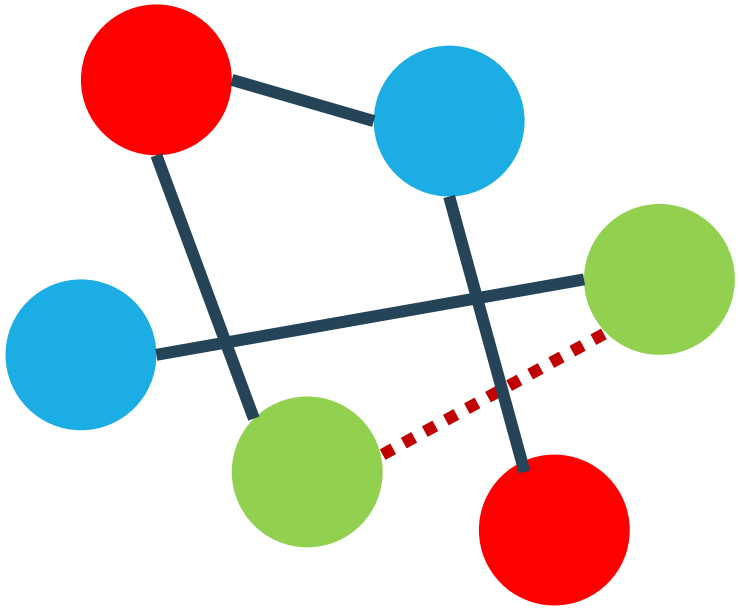
real world examples: scheduling, coloring a map, sudoku solver, CPU allocation



Graph Coloring Problem

The problem: color each node red, blue, or green, BUT minimize nodes with the same color sharing an edge (i.e., max. edges between distinct)

real world examples: scheduling, coloring a map, sudoku solver, CPU allocation



Graph Coloring Problem

The problem: color each node red, blue, or green, BUT minimize nodes with the same color sharing an edge (i.e., max. edges between distinct)

real world examples: scheduling, coloring a map, sudoku solver, CPU allocation

Simple algorithm: Randomly pick a color for each node

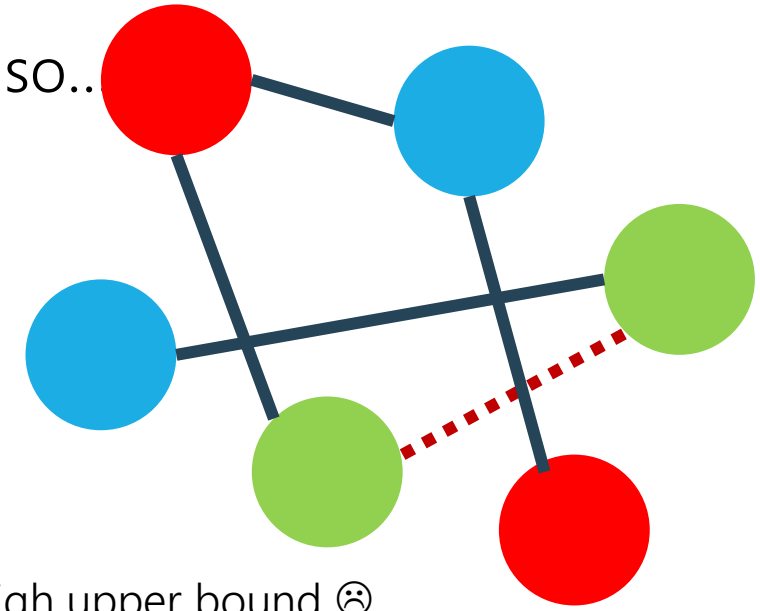
Probability of edge e sharing a color (miscoloring) is $\frac{1}{3}$, so..

$\mathbb{E}[S_e] = \frac{1}{3}$ where S_e is whether edge is miscolored

S (num. of miscolored edges): $S = \sum_i^n S_e$ -->$\mathbb{E}[S] = \frac{n}{3}$

So, by Markov's inequality,

$$\mathbb{P}\left(S \geq 1.1 \cdot \frac{n}{3}\right) \leq \frac{n/3}{1.1 \cdot n/3} = \frac{1}{1.1} \approx 0.91$$



The probability of the algorithm miscoloring more than a third edges has a high upper bound ☹

Graph Coloring Problem

The problem: color each node red, blue, or green, BUT minimize nodes with the same color sharing an edge (i.e., max. edges between distinct)
real world examples: scheduling, coloring a map, sudoku solver, CPU allocation

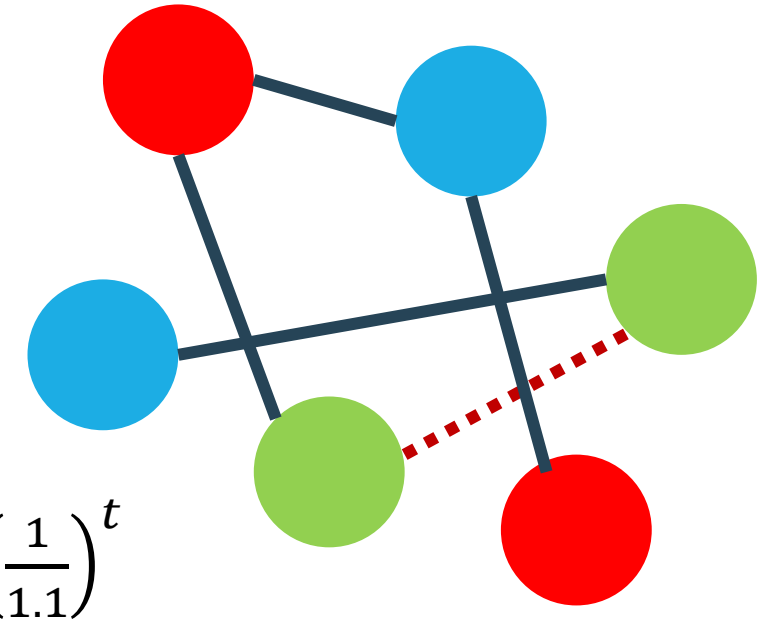
Simple algorithm: Randomly pick a color for each node

S (num. of miscolored edges): $S = \sum_i^n S_e$ --> $\mathbb{E}[S] = \frac{n}{3}$

So, by Markov's inequality,

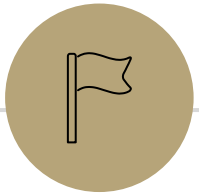
$$\mathbb{P}\left(S \geq 1.1 \cdot \frac{n}{3}\right) \leq \frac{n/3}{1.1 \cdot n/3} = \frac{1}{1.1} \approx 0.91$$

We can use a Monte Carlo algorithm!
Keep repeating the algorithm t times.
Probability you **fail to find a good coloring** is at most $\left(\frac{1}{1.1}\right)^t$
probability is *very* low with high values of t



If you like this kind of stuff…

- CSE 421 covers algorithms (like min cut, graph color, and more!)
- CSE 431 covers the theory behind this algorithms (which includes analysis of randomize algorithms!)



Differential Privacy

Privacy Preservation

A real-world example (adapted from The Ethical Algorithm by Kearns and Roth; based on protocol by Warner [1965]).

And gives a sense of how randomness is actually used to protect privacy.

Privacy Preservation with Randomness

You're working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

> Use CLT or a tail-bound to estimate the needed number n get a guaranteed good estimate, right?

> Do a poll, call up a random sample of adults and ask them "have you ever cheated on your romantic partner?"

You do that, and somehow, no one says they cheated. I wonder why…

What's the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

You don't want to hold this data, and the people you're calling don't want you to hold this data.

Doing It Better!

You don’t need to know **who** was cheating. Just how many people were.

Here’s an idea:

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me ‘heads’
If the coin is tails and you have not ever cheated, please tell me ‘tails’

We have two concerns with this:

> Will it now be private?

> Will we be able to make accurate estimates using this data?

Will it be private?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

If you are someone who has cheated, and you report heads, can that be used against you? Not substantially – just blame it on the coin!

You discover your partner said heads (H), what's the probability that they cheated (C)?

Will it be private?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

If you are someone who has cheated, and you report heads, can that be used against you? Not substantially – just blame it on the coin!

You discover your partner said heads (H), what's the probability that they cheated (C)?

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

Will it be private?

Please flip a coin.

If the coin is heads, or you have ever cheated, please tell me 'heads'

If the coin is tails and you have not ever cheated, please tell me 'tails'

If you are someone who has cheated, and you report heads, can that be used against you? Not substantially – just blame it on the coin!

You discover your partner said heads (H), what's the probability that they cheated (C)?

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~26%). But that isn't too damaging.

But will it be accurate?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

We poll n people.

X is number of people who said "heads", Y is number who cheated in sample,
p is true probability of cheating in the population.

Can we make an estimate for Y? We would use MLE but intuitively…
$\hat{Y} =$

But will it be accurate?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

We poll n people.

X is number of people who said "heads", Y is number who cheated in sample,
p is true probability of cheating in the population.

Can we make an estimate for Y? We would use MLE but intuitively…
$\hat{Y} = 2\left(X - \frac{n}{2}\right)$, this turns out to be an unbiased estimator

But will it be accurate?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

We poll n people.

X is number of people who said "heads", Y is number who cheated in sample, p is true probability of cheating in the population.

What about the variance of the estimates?

If we did not do the anonymous poll, and just asked directly…

$$A \sim \mathrm{Bin}(n, p) \rightarrow \mathrm{Var}(A) = np(1-p) \leq \frac{n}{4}$$

With this "anonymized" poll, the variance of our estimate is…

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}\left(2\left(X - \frac{n}{2}\right)\right) = 4\mathrm{Var}\left(X - \frac{n}{2}\right) = 4\mathrm{Var}(X) \leq n$$

But will it be accurate?

Details for computing $\mathrm{Var}(X)$:

$\mathbb{P}(X_i = 1) = \frac{1}{2} + \frac{1}{2} \cdot p$, $X = \sum_{i=1}^{n} X_i$, $\mathbb{E}[X] = \frac{1}{2} + \frac{np}{2} = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$

$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$

$Y = 2\left(X - \frac{n}{2}\right)$

$\mathrm{Var}(X) = \mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$

$\mathrm{Var}(X_i)$? It's an indicator with parameter $p + (1-p) \cdot \frac{1}{2} = \frac{1}{2} + \frac{p}{2}$

So $\mathrm{Var}(X_i) = \left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right)$

$\mathrm{Var}(Y) = 4\mathrm{Var}(X) = 4n\mathrm{Var}(X_i) = 4n\left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right) \leq \frac{4n}{4} = n$

The variance is 4 times as much as it would have been for a non-anonymous poll.

But will it be accurate?

Please flip a coin.
If the coin is heads, or you have ever cheated, please tell me 'heads'
If the coin is tails and you have not ever cheated, please tell me 'tails'

We poll n people.

X is number of people who said "heads", Y is number who cheated in sample, p is true probability of cheating in the population.

What about the variance of the estimates?

If we did not do the anonymous poll, and just asked directly…

$$A \sim \mathrm{Bin}(n, p) \rightarrow \mathrm{Var}(A) = np(1-p) \leq \frac{n}{4}$$

With this "anonymized" poll, the variance of our estimate is…

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}\left(2\left(X - \frac{n}{2}\right)\right) = 4\mathrm{Var}\left(X - \frac{n}{2}\right) = 4\mathrm{Var}(X) \leq n$$

Can we use Chernoff?

Goal: Use a bound to understand how far this estimate tends to be from true value

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables.
Let $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) \text{ and } \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

What happens with $\mathbf{n = 1000}$ people?

What range will we be within at least 95% of the time?

☹ Can't bound δ without bounding p

The right tail is the looser bound, so ensuring the right tail is less than 2.5% gives us the needed guarantee.

$$\mathbb{P}(Y \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) = \exp\left(-\frac{\delta^2 1000p}{3}\right) \leq .025$$

$$-\frac{\delta^2 1000p}{3} \leq \ln(.025)$$

$$-\delta^2 \leq \frac{3\cdot\ln(.025)}{1000p}$$

$$\delta \geq \sqrt{\frac{-3\ln(.025)}{1000p}}$$

As $p \to 0$, $\delta \to \infty$ – we're not actually making a claim anymore.

A different inequality

If we try to use Chernoff, we'll hit a frustrating block.

Since μ depends on p, p appears in the formula for δ. And we wouldn't get an absolute guarantee unless we could plug in a p.

And it'll turn out that as $p \to 0$ that $\delta \to \infty$ so we don't say anything then.

Luckily, there's always another bound…

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i/n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$. Why?

$$Y = 2\left(X - \frac{n}{2}\right) \text{ or } X = \frac{Y+n}{2}$$

$$|X - \mathbb{E}[X]|$$
$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y+n}{2}\right]\right|$$
$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y}{2}\right] - \frac{n}{2}\right|$$
$$= \left|\frac{Y}{2} - \mathbb{E}\left[\frac{Y}{2}\right]\right|$$
$$= \frac{1}{2}|Y - \mathbb{E}[Y]|$$

So $|X - \mathbb{E}[X]| \geq t$ if and only if $\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$ iff $|Y - \mathbb{E}[Y]| \geq 2t$.

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i/n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

How close will we be with *n=1000 with probability at least* .95*?*

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$.

Margin of Error

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp(-2nt^2) \leq .05$

For $n = 1000$, we get:

$2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05 \Rightarrow -\frac{2000t^2}{4} \leq \ln(.025) \Rightarrow t \leq .086.$

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq .086) \leq .05$

So our margin of error is about 8.6%.

To get a margin-of-error of 5% need $2\exp\left(-2n\left(\frac{.05}{2}\right)^2\right) \leq .05$

$n \geq 2952$

How much do we lose?

We lose a factor of two in the length of the margin (equivalently, we'd need to talk to 4 times as many people to have the same confidence.

You can also control this tradeoff.

Want more accuracy? Make it roll a die: report 1 if cheated (truth o/w)

Want more security? Make it Bernoulli with probability $p \gg \frac{1}{2}$ or cheated have the same report (e.g. report "die roll 1 [and didn't cheat]" or "die roll 2-6 [or did cheat]"

In The Real World

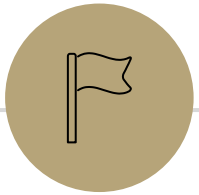
Injecting randomness to preserve privacy is a real thing.

Instead of having everyone flip a coin, "random noise" can be inserted after all the data has been collected.

Differential privacy is being used to protect the 2020 Census data.

The overall count of people in each state is exact (well, exactly the data they collected). But the data per block or per city will be randomized to protect against .

This video nicely explains what's involved. Notice that the accuracy guarantees come in the same "inside-margin-of-error-with-probability" guarantees we've been giving for our randomness (just much stronger).



How to *Lie* With Statistics

Acknowledgements

"How to Lie with Statistics" by Darrell Huff (1954)

Previous versions of this lecture by Stefano Tessaro, Anna Karlin and Alex Tsun

Jevin West "Naked Statistics: Stripping the Dread from Data" by Charles Wheelan

And more!

"The book"

Published in 1954, over 500,000 copies sold

- "A great introduction to the use of statistics, and a great refresher for anyone who's already well versed in it" - Bill Gates.
- 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 all the time, e.g., by politicians, and marketers.
  - Often make decisions based on these lies: "4 out of 5 dentists recommend…."

Some of the ways we can get fooled…

- Sampling
- Faking Distributions
- Hypothesis testing + p-hacking
- There's so much more!!
  Simpson's paradox, base rate fallacy, gambler's fallacy, spurious correlations, etc.

Statistical Inference

Making an estimate or prediction about a population based on a sample.

Statistical Inference *gone wrong*

Making an estimate or prediction about a population based on a sample.

For example…

"The Literary Digest" Magazine poll to predict 1936 election.

Alfred Landon vs. Franklin D Roosevelt

10 million surveys, 2.4 million responses subscribers

owners of cars and telephones on a List

Electoral Votes	Prediction	Actual
Landon	370	8
Roosevelt	161	523

Statistical Inference *gone wrong*

Making an estimate or prediction about a population based on a sample.

For example…

"The Literary Digest" Magazine poll to predict 1936 election.

Alfred Landon vs. Franklin D Roosevelt

10 million surveys, 2.4 million responses subscribers

owners of cars and telephones on a List

Electoral Votes	Prediction	Actual
Landon	370	8
Roosevelt	161	523

Statistical Inference *gone wrong*

Making an estimate or prediction about a population based on a sample.

For example…

"The Literary Digest" Magazine poll to predict 1936 election.

Alfred Landon vs. Franklin D Roosevelt

10 million surveys, 2.4 million responses subscribers

owners of cars and telephones on a List

Electoral Votes	Prediction	Actual
Landon	370	8
Roosevelt	161	523

Problem:
The survey was not representative!

Sampling

We've often said "Let $X_1, X_2, \ldots, X_n$ be i.i.d samples from a distribution"...

- Not **representative**
  - Voluntary response bias (24% responded)
  - Not the right population (more money/education /info than average American)
- Not **random**
  - Convenience sampling

⚠️ More sampling doesn't fix bad sampling techniques.

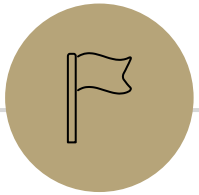
This i.i.d. assumption is often a great place to start with doing analysis!
But be careful with the conclusions drawn from the analysis.

Sampling

We've often said "Let $X_1, X_2,\ldots,X_n$ be i.i.d samples from a distribution"…

This i.i.d assumption is often a great place to start with doing analysis!
But be careful with the conclusions drawn from the analysis.

- e.g., MLEs – we use some samples from a distribution to estimate a parameter to a distribution, and use this to make predictions for the future
  - selection bias in data --> biased model --> biased and incorrect conclusions



Faking a Distribution

Faking a Probability Distribution

Faking a probability distribution…

> Wrong shape

> Too close to expected value (especially replicated)

> Too far from expected value

> Replications too good to be true

e.g., Gegor Mendel's Sweet Peas

Postulated that self fertilization of hybrid yellow-seeded sweet peas would yield offspring with :

> 0.75 chance yellow-seeded

> 0.25 chance green seeded.

In 1865, reported results of 8023 experiments:

> 0.7505 yellow-seeded

> 0.2495 green-seeded.

Using the probability techniques we've learned, probability of observations as close to expected value as he reported is minute…

https://www.benchling.com/blog/gregor-mendel-villain-or-victim

In this case, his conclusions were correct

But in other cases, they aren't always!

> Fabricated paper linking vaccines to autism

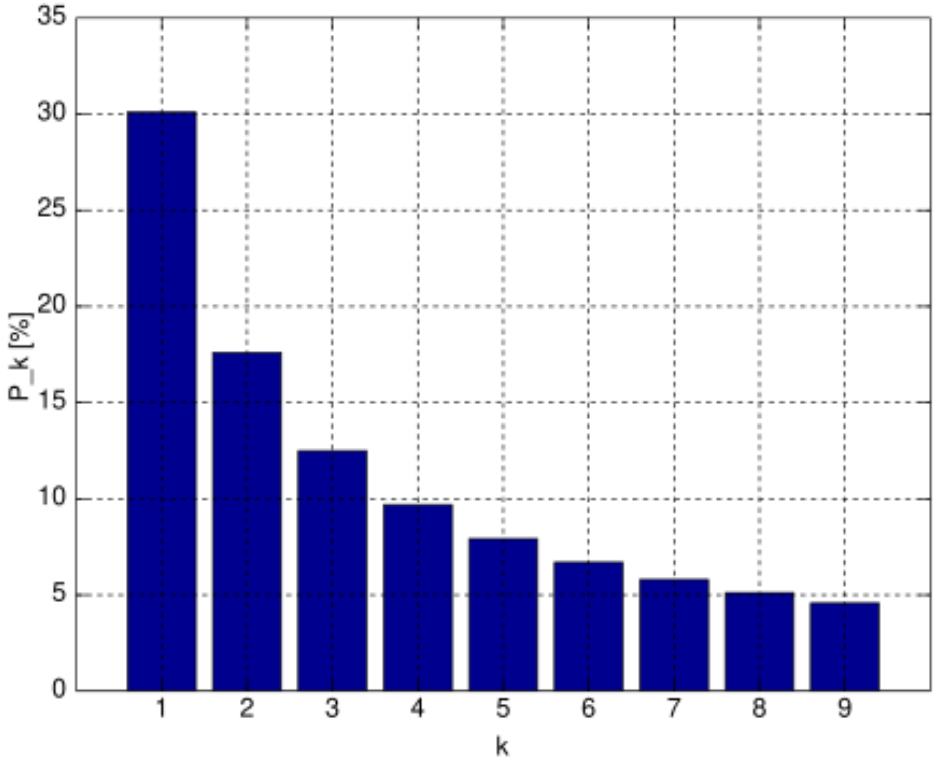
> Fabricated data in a paper about honesty…

https://www.benchling.com/blog/gregor-mendel-villain-or-victim

Benford's Law

In many real-world datasets, the leading digit is likely to be small.

To be precise, a set of numbers is said to fulfill Benford's law if the leading digit $d \in \{1,2,\dots,9\}$ occurs with probability $\mathbb{P}(d) = \log_{10}(d+1) - \log_{10}(d)$



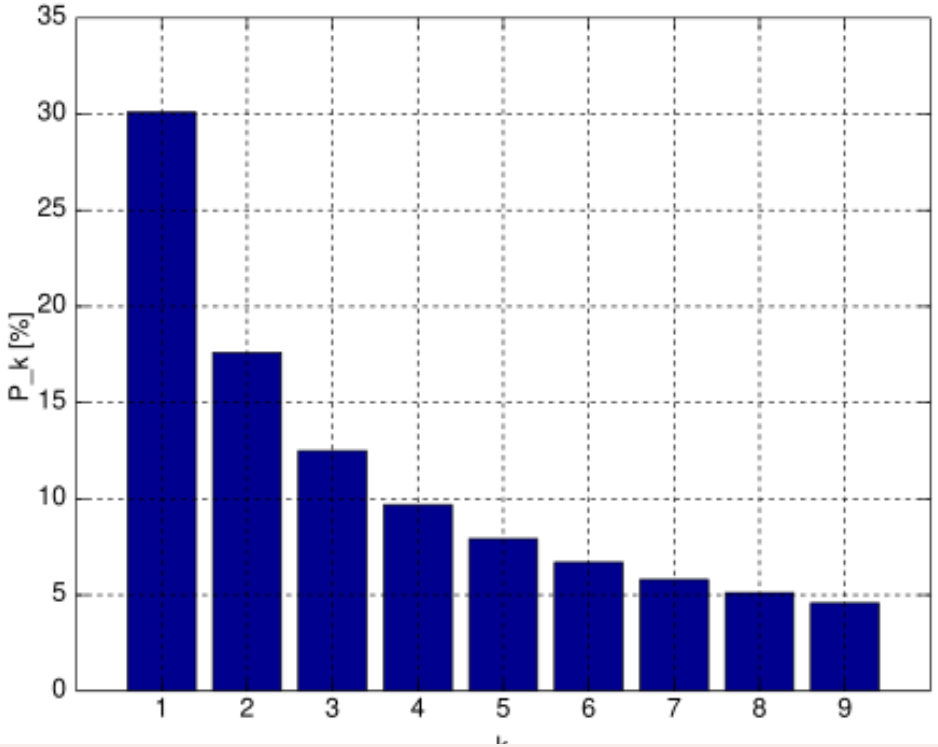
This can be used to detect fraudulent datasets

https://commons.wikimedia.org/w/index.php?curid=4509760

Benford's Law

In many real-world datasets, the leading digit is likely to be small.

To be precise, a set of numbers is said to fulfill Benford's law if the leading digit $d \in \{1,2,\dots,9\}$ occurs with probability $\mathbb{P}(d) = \log_{10}(d+1) - \log_{10}(d)$



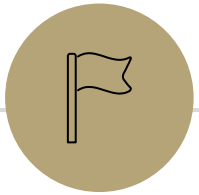
> Detecting fraud in financial documents (see Wesley Rhodes)
> Bogus retweets, bot networks of followers
> Greece manipulating macroeconomic data to join the Eurozone
> Vote-rigging in Iran's 2009 presidential election

Takeaways

~~*If you want to make fake data, now you know what to look out for….*~~

> Don’t make up data or statistics. (Obviously, but also because it’s pretty easy to catch you).

> More importantly, know that you can look at data and figure out whether to look more closely at it or not



p-Hacking

Hypothesis Testing

We are researching jelly beans and whether they cause acne in teenagers.

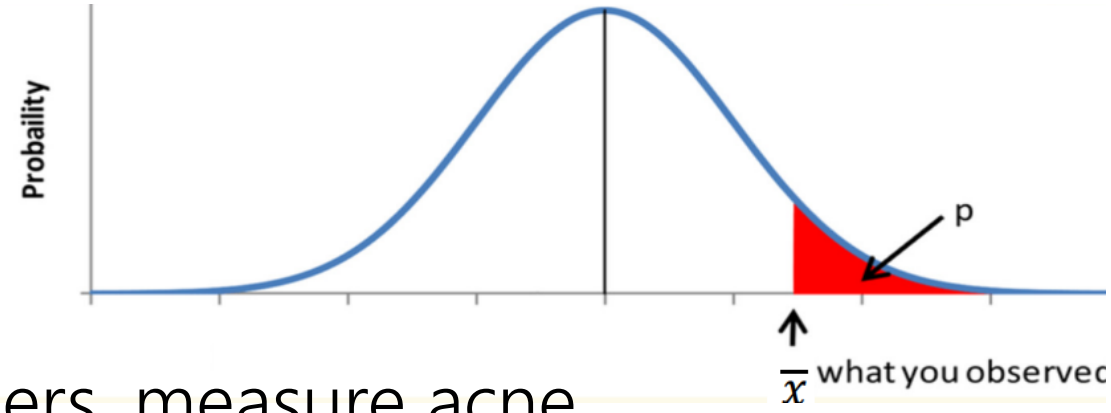
The average teen has amount of acne with mean μ and variance σ^2

H_0 (null hypothesis): Jelly beans have no effect on acne
(i.e., mean acne for someone who eats jelly beans $= \mu$)

H_1 (alternative hypothesis): Jelly beans increase acne
(i.e., mean acne for someone who eats jelly beans $> \mu$)

Goal: provide enough evidence that we can reject the null hypothesis

Hypothesis Testing



> Choose a significance level, say $\mathbf{0.05}$

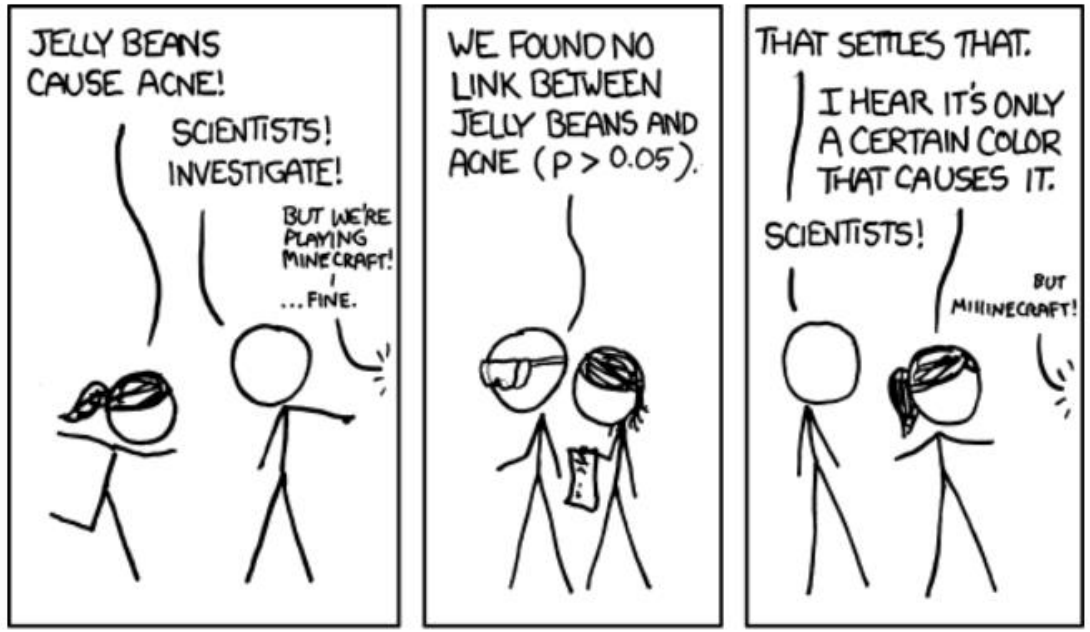
> Observe 100 jelly bean eating teenagers, measure acne.
Sample mean observed: $\overline{x}$

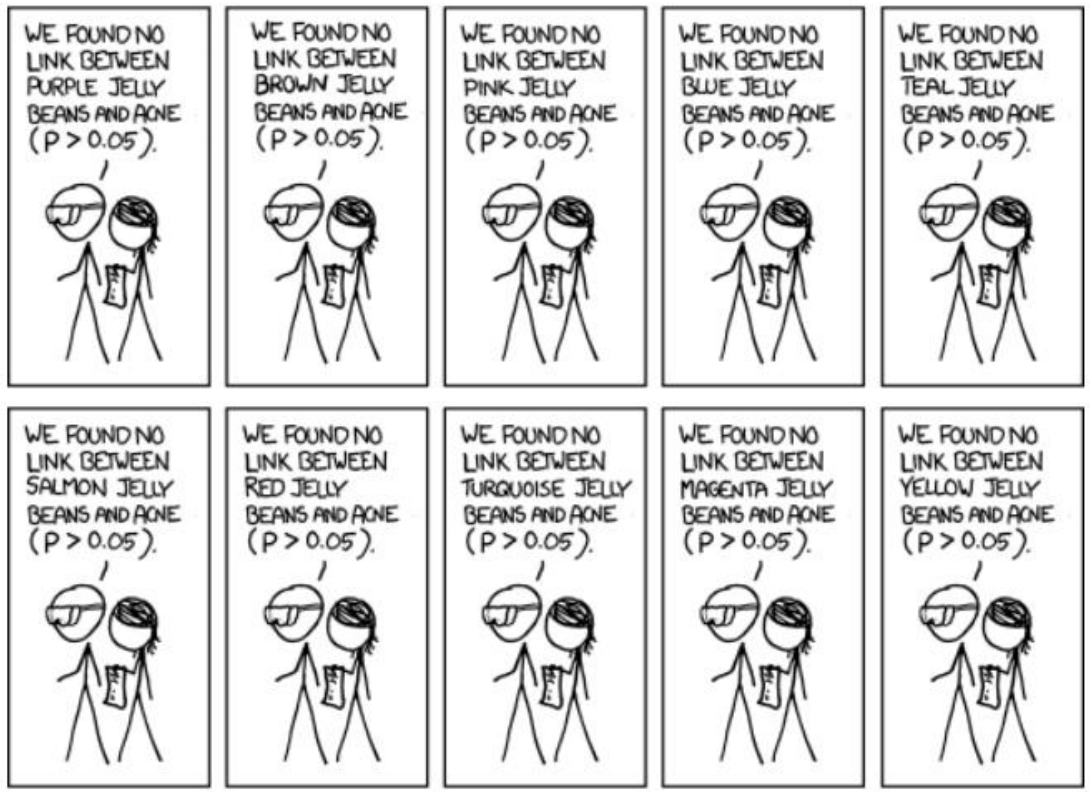
> **p-value**: *If null hypothesis is true*, what is the probability of observing the amount of acne we saw? (i.e., the red region): $p = \mathbb{P}(\overline{X} \geq \overline{x}) = 0.0162$

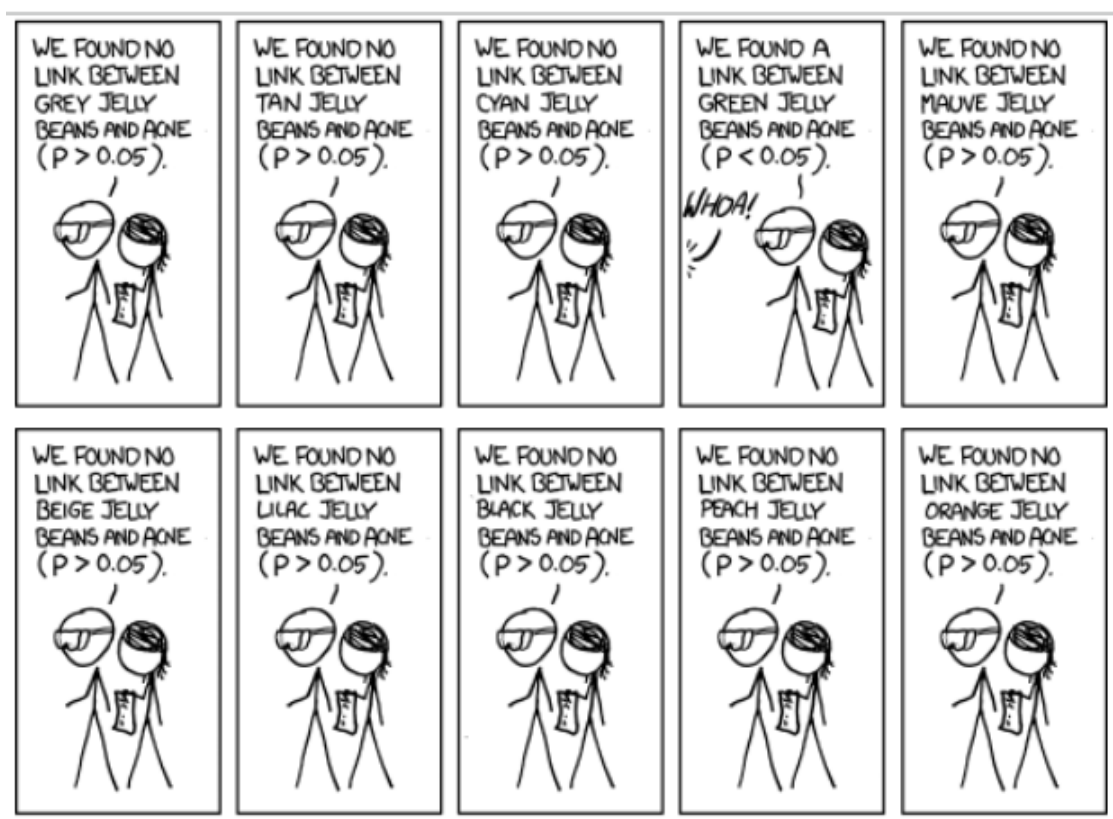
If $p < \mathbf{0.05}$, reject H_0 at the $\mathbf{0.05}$ significance level
(i.e. strong statistical evidence jelly beans cause an increase in acne).

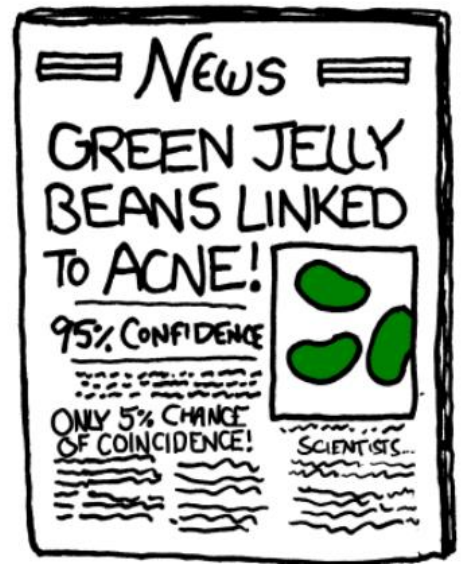
If $p > \mathbf{0.05}$, accept H_0 at the $\mathbf{0.05}$ significance level
(i.e. it's possible this observation was just by chance, not enough info to reject).

Goal: provide enough evidence that we can reject the null hypothesis


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

p-Hacking

$p < 0.05$: 5% chance of seeing this much acne if jelly beans don't cause acne

But what if I repeat similar experiments 20 times?

The probability at least one of those trials will have p value < 0.05 is

$$1 - \mathbb{P}(\text{all trials have } p > 0.05) = 1 - (0.95^{20}) \approx 0.65$$

In other words 64% of the time one of these tests will be significant. So, this result of having value of p > 0.05 for the green jelly beans was most likely random chance!

p-Hacking

P-hacking: Performing a hypothesis test multiple times to get at least one statistically significant result…

And…(the particularly evil thing)…not reporting the 19 insignificant tests!

p-Hacking

> Don’t p-hack.

> Know that p-hacking is very prevalent in industry.

> Leads to better headlines. Publication bias.

Science is human-motivated, sometimes. Maybe most of the time.

How (not to) lie with statistics

Often, the problem is not that a statistic is incorrect.

> Context matters.

> Units matter.

> Relative numbers matter.

> Research: choose what to study and how to study it

Truths can compete: there may be multiple interpretations of facts.

How (not to) lie with statistics

1. Determine if the samples are random and representative.
2. Ask for a confidence interval.
3. Be dubious. Be extremely dubious.
4. Don’t make up statistics. You’ll get caught.
5. Be wary of p-hacking (and don’t do it yourself)!
6. Be careful about seeing patterns where there are none.
7. Correlation does not imply causation.
8. Be careful with interpreting conditional probabilities. Intuition sometimes doesn’t work here!
9. Be wary of assuming things are independent that aren’t independent.