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

Maximum Likelihood Estimators (MLE)

CSE 312 24Su

Lecture 19

Announcements

- HW6 released on Wednesday, due Monday 8/12
  Longer (8 questions), so start early – you have enough to do the first 5 question

Outline

- Finish Chernoff bound example from Monday
- Union bound
- Maximum Likelihood Estimation

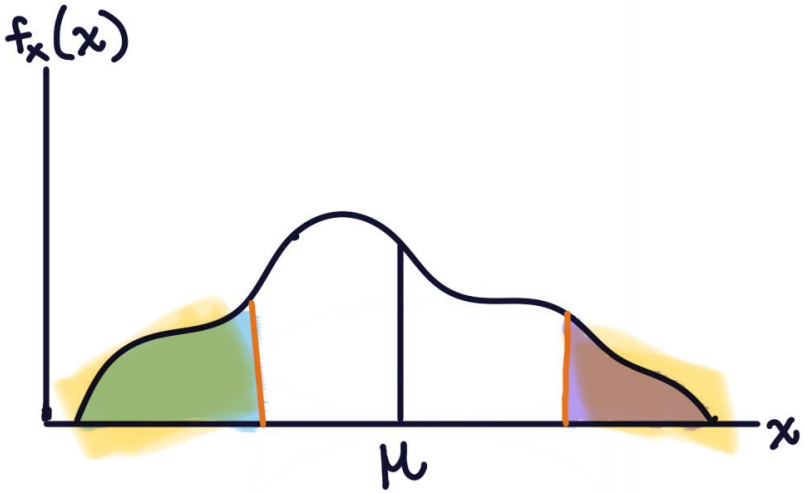
Tail Bounds

What's a Tail Bound?

A **tail bound** (or concentration inequality) bounds the probability in the "tails" of the distribution. e.g., statements like $\mathbb{P}(X \geq 4) \leq 0.8$, $\mathbb{P}(X \geq 4) \leq 0.8$

We've seen this before! We can:

- Compute these probabilities exactly in some cases
- Approximate X as normal using CLT if X is the sum of a bunch of i.i.d random variables



But what if we barely know anything about X and it doesn't fit into the frameworks we've learned about? Can we still make some tail bound guarantees?

Markov's Inequality

Two statements are equivalent.
Left form is often easier to use.
Right form is more intuitive.

Markov's Inequality

Let X be a random variable supported (only) on non-negative numbers. For any $t > 0$

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

Markov's Inequality

Let X be a random variable supported (only) on non-negative numbers. For any $\mathbf{k} > 0$

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Requirements:

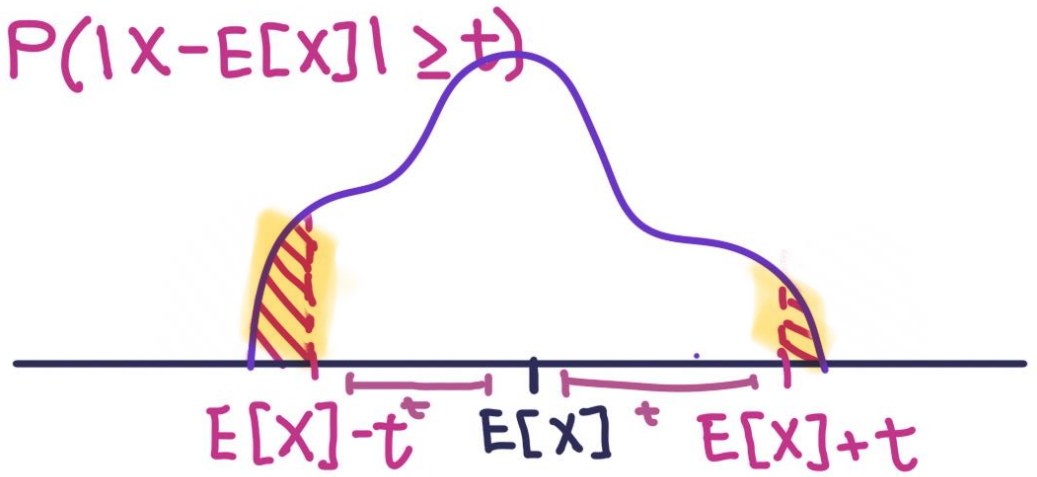
1. X must be non-negative
2. We know the expectation of X

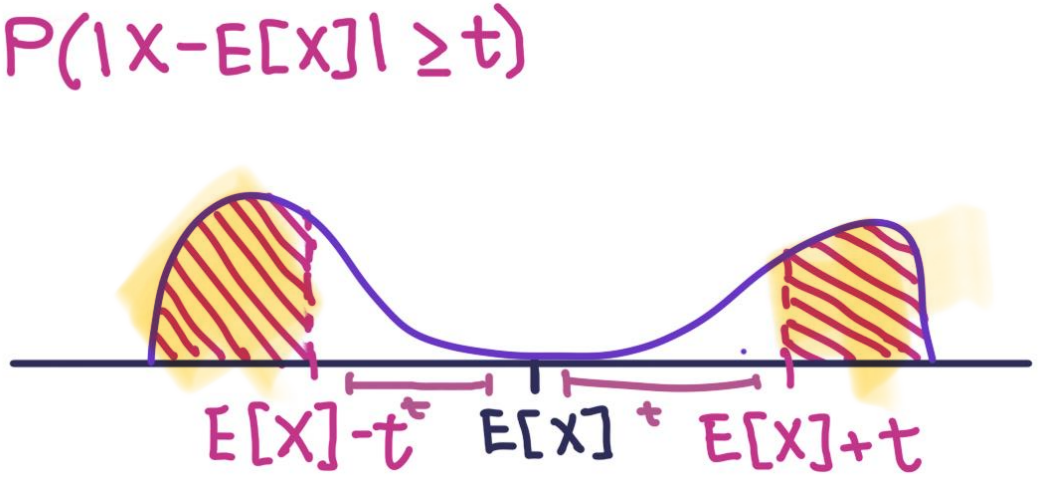
Chebyshev's Inequality

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

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





Requirements:

1. We know the *expectation* of X
2. We know the *variance* of X

Chernoff Bound

(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 $0 \leq \delta \leq 1$

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

LEFT TAIL RIGHT TAIL

Requirements:

1. X is a sum of independent Bernoulli random variables.
2. We know $\mathbb{E}[X]$

Example: Polling (again, but better!)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

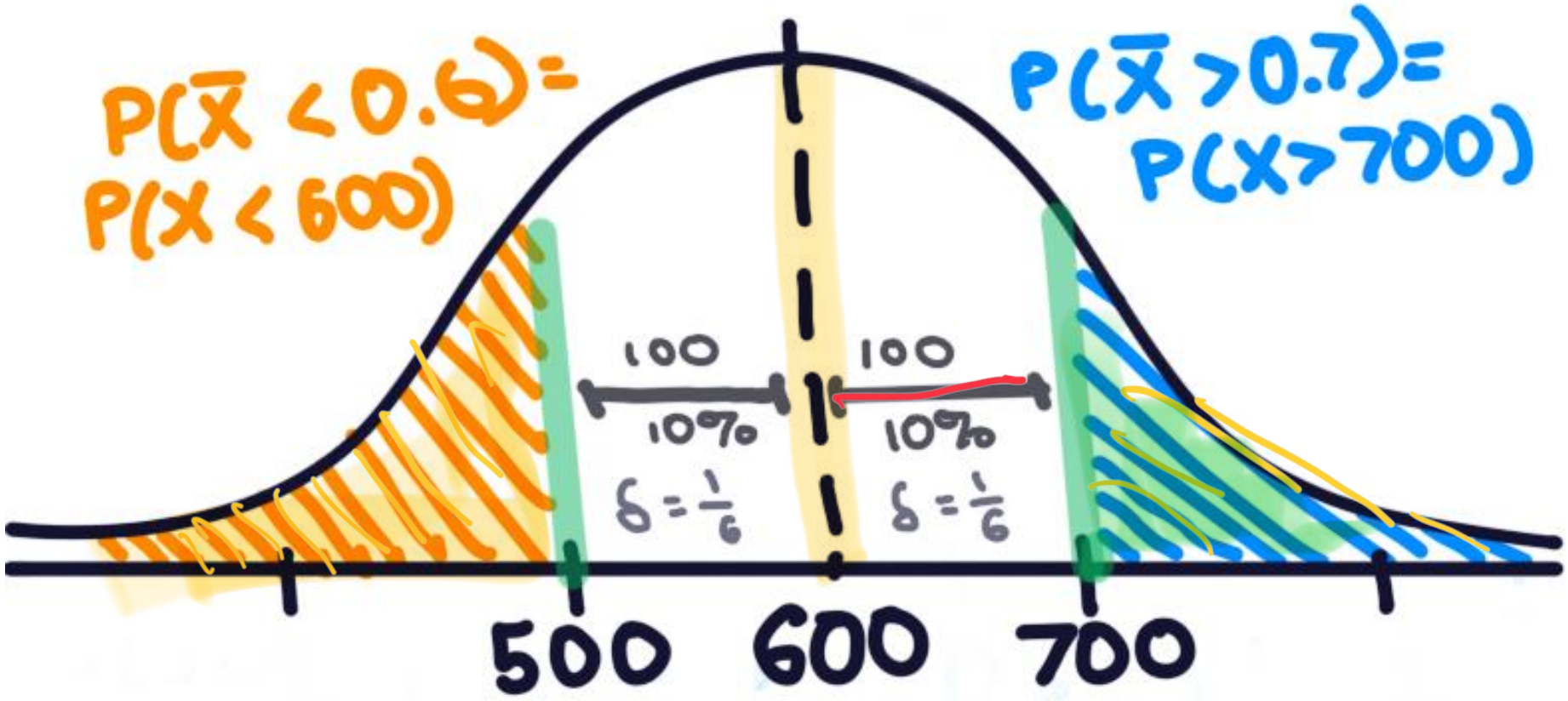
Goal: bound $\mathbb{P}(|\bar{X} - 0.6| \geq 0.1) = \mathbb{P}(\bar{X} \leq 0.5) + \mathbb{P}(\bar{X} \geq 0.7)$

(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 $0 \leq \delta \leq 1$

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

LEFT TAIL RIGHT TAIL



Suppose you run a poll of 1000 *people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?*

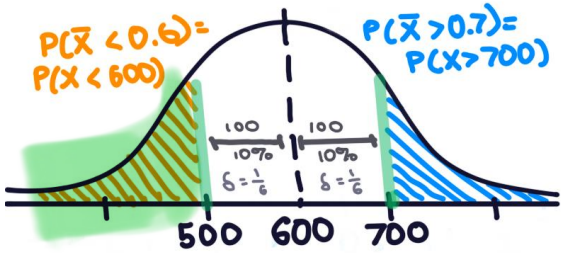
Goal: bound $\mathbb{P}(|\bar{X} - 0.6| \geq 0.1) = \mathbb{P}(\bar{X} \leq 0.5) + \mathbb{P}(\bar{X} \geq 0.7)$

Example: Polling (1. bound the left tail)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$X = \sum X_i$, where $X_i \sim \text{Ber}(0.6)$, $\mu = \mathbb{E}[X] = 1000 \cdot 0.6 = 600$

$$\mathbb{P}\left(\frac{X}{1000} \leq 0.5\right) =$$



Chernoff Bound (left tail)

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 $0 \leq \delta \leq 1$

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

Goal: bound $\mathbb{P}(|\bar{X} - 0.6| \geq 0.1) = \mathbb{P}(\bar{X} \leq 0.5) + \mathbb{P}(\bar{X} \geq 0.7)$

Example: Polling (1. bound the left tail)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$X = \sum X_i$, where $X_i \sim \text{Ber}(0.6)$, $\mu = \mathbb{E}[X] = 1000 \cdot 0.6 = 600$

$$\mathbb{P}\left(\frac{X}{1000} \leq 0.5\right) = \mathbb{P}(X \leq 500)$$

$500 = (1 - \delta)600$ -> $\delta = \frac{1}{6}$

$$\ldots = \mathbb{P}\left(X \leq \left(1 - \frac{1}{6}\right)\mu\right) \leq e^{-\frac{\frac{1}{6^2} \cdot 600}{2}}$$

$$\approx 0.0003$$

Chernoff Bound (left tail)

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 $0 \leq \delta \leq 1$

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

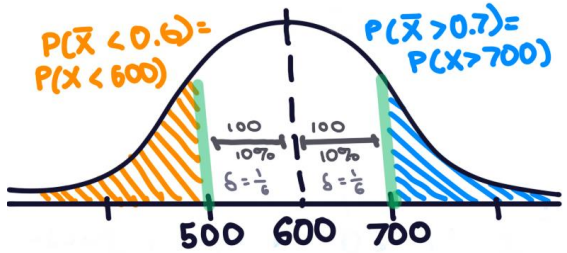
Example: Polling (2. bound the right tail)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$X = \sum X_i$, where $X_i \sim \text{Ber}(0.6)$, $\mu = \mathbb{E}[X] = 1000 \cdot 0.6 = 600$

$\mathbb{P}\left(\frac{X}{1000} \geq 0.7\right) = P(X \geq 700)$

$700 = (1+\delta)600 \rightarrow \delta = \frac{1}{6}$



Chernoff Bound (right tail)

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 $0 \leq \delta \leq 1$

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

Example: Polling (2. bound the right tail)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$X = \sum X_i$, where $X_i \sim \text{Ber}(0.6)$, $\mu = \mathbb{E}[X] = 1000 \cdot 0.6 = 600$

$$\mathbb{P}\left(\frac{X}{1000} \geq 0.7\right) = \mathbb{P}(X \geq 700)$$

$700 = (1+\delta)600$ -> $\delta = \frac{1}{6}$

$$\ldots = \mathbb{P}\left(X \geq \left(1 + \frac{1}{6}\right)\mu\right) \leq e^{-\frac{\frac{1}{6^2} \cdot 600}{3}}$$

≈ 0.0039

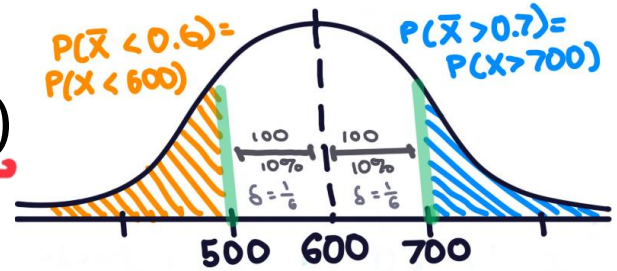
Chernoff Bound (right tail)

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 $0 \leq \delta \leq 1$

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

Example: Polling (3. Putting it all together)

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?



We want $\mathbb{P}(|\bar{X} - 0.6| \geq 0.1) = \mathbb{P}(\bar{X} \leq 0.5) + \mathbb{P}(\bar{X} \geq 0.7)$

We know..

-> $\mathbb{P}(\bar{X} \leq 0.5) = \mathbb{P}(X \leq 500) \leq 0.0003$ (from Chernoff bound, left tail)

-> $\mathbb{P}(\bar{X} \geq 0.7) = \mathbb{P}(X \geq 700) \leq 0.0039$ (from Chernoff bound, right tail)

So, $\mathbb{P}(|\bar{X} - 0.6| \geq 0.1) \leq 0.0003 + 0.0039 = 0.0042$

Less than 1%. That's a better bound than Chebyshev gave!

Wait a Minute

This is just a binomial!

Well if all the X_i have the same probability. It does work if they're independent but have different distributions. But there's bigger reasons to care…

The concentration inequality will let you control n easily, even as a variable. That's not easy with the binomial.

What happens when n gets big?

Evaluating $\binom{20000}{10000}.51^{10000} \cdot .49^{10000}$ is fraught with chances for floating point error and other issues. Chernoff is much better.

Wait a Minute

I asked Wikipedia about the "Chernoff Bound" and I saw something different?

This is the "easiest to use" version of the bound. If you need something more precise, there are other versions.

Why are the tails different??

The strongest/original versions of "Chernoff bounds" are symmetric ($1+\delta$ and $1-\delta$ correspond), but those bounds are ugly and hard to use.

When computer scientists made the "easy to use versions", they needed to use some inequalities. The numerators now have plain old δ's, instead of $1+$ or $1-$. As part of the simplification to this version, there were different inequalities used so you don't get exactly the same expression.

But Wait! There's More

For this class, please limit yourself to:
Markov, Chebyshev, and Chernoff, as stated in these slides…

But for your information. There's more.

> Trying to apply Chebyshev, but only want a "one-sided" bound (and tired of losing that almost-factor-of-two)Try Cantelli's Inequality

> In a position to use Chernoff, but want additive distance to the mean instead of multiplicative? They got one of those.

> Have a sum of independent random variables that aren't indicators, but are bounded, you better believe Wikipedia's got one

> Have a sum of random **matrices** instead of a sum of random numbers. Not only is that a thing you can do, but the eigenvalue of the matrix concentrates

There's a whole book of these!

Tail Bounds – Takeaways

Useful when an experiment is complicated and you just need the probability to be small (you don't need the exact value).

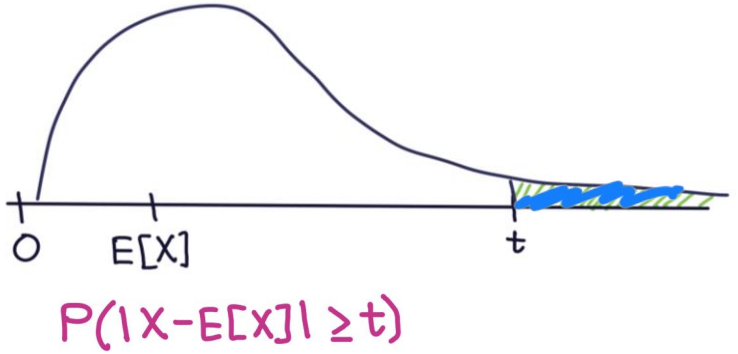
Choosing a minimum n for a poll – don't need exact probability of failure, just to make sure it's small.

Designing probabilistic algorithms – just need a guarantee that they'll be extremely accurate

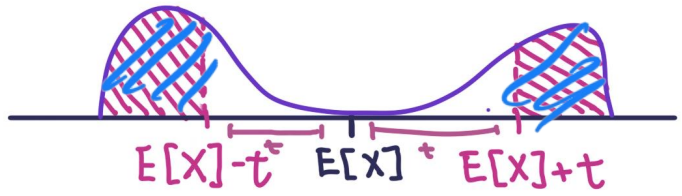
Learning more about the situation (e.g. learning variance instead of just mean, knowing bounds on the support of the starting variables) usually lets you get more accurate bounds.

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



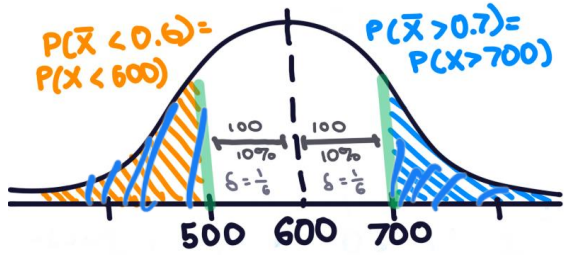
- **Chebyshev's inequality** - $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var(t)}}{t^2}$
  - Use if we know the expectation **and** variance of X
  - Gives better bounds with small variances

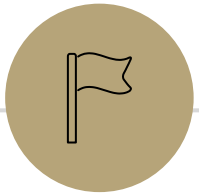


- **Chernoff Bound**

  $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{\left(-\frac{\delta^2\mu}{2}\right)}$ *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





One More Bound – Union Bound

Union Bound (not a *tail* bound, but still a bound)

Union Bound

For any events E, F

$$\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F)$$

Sometimes we don't' have enough information to compute this probability exactly, so we use the union bound to bound that probability

Proof?

By *inclusion-exclusion*, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

And $\mathbb{P}(E \cap F) \geq 0$.

Concentration Applications

A common pattern:

"What's the probability something goes wrong?"

> Figure out "what could possibly go wrong" – often these are dependent. Use a tail bound for each of the things that could go wrong.

> Union bound over everything that could go wrong.

Example: Frogs 🐸

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability at least one square ends up with at least 36 frogs.

Example: Frogs 🐸

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability at least one square ends up with at least 36 frogs.

A_i is the event the i'th square has at least 36 frogs

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_{25})$$
$$\leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots + \mathbb{P}(A_{25}) \quad \textit{by the union bound}$$

How do we find $\mathbb{P}(A_i)$? Use another bound!

These events are dependent – adjacent squares affect each other!

Example: Frogs 🐸

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability at least one square ends up with at least 36 frogs.

A_i is the event the i'th square has at least 36 frogs

How do we find $\mathbb{P}(A_i)$? Use another bound!

Let Y be the number frogs in i'th square

$P(Y \geq 36)$

$20 \cdot 5 = 100$

Y_j ~ jth frogs ends

$Y_j \sim \text{Ber}(\frac{1}{5})$ $Y = \sum_{j=1}^{100} Y_j$

Example: Frogs 🐸

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability at least one square ends up with at least 36 frogs.

A_i is the event the i'th square has at least 36 frogs

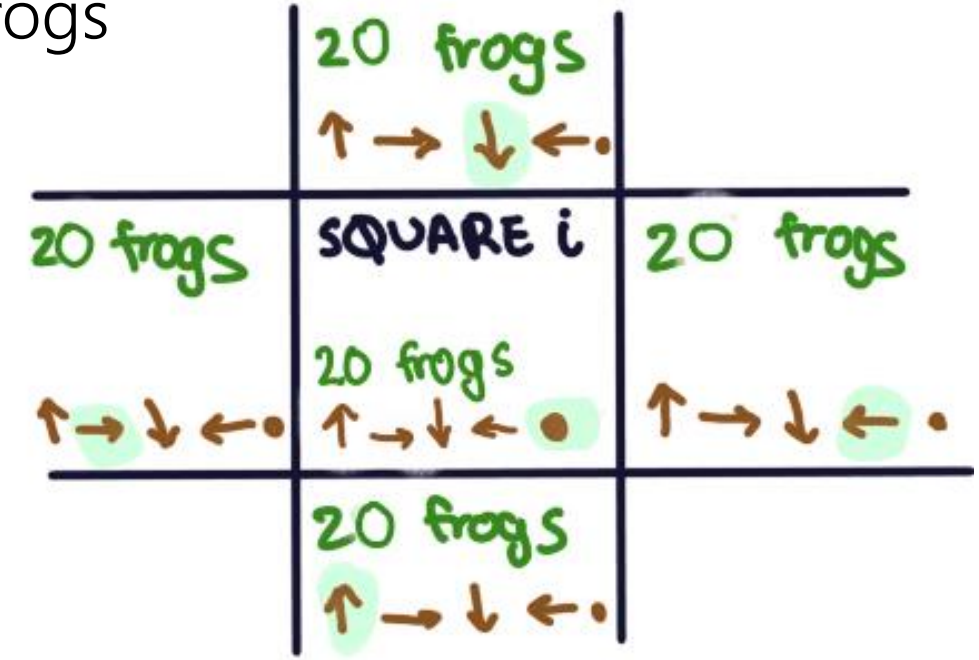
How do we find $\mathbb{P}(A_i)$? Use another bound!

Let Y be the number frogs in i'th square

$Y = \sum_{j=1}^{100} X_j$, $X_j \sim \text{Ber}(1/5)$, $E[Y] = \frac{100}{5} = 20$

$\mathbb{P}(A_i) = \mathbb{P}(Y \geq 36) = \mathbb{P}\left(Y \geq \left(1 + \frac{4}{5}\right) 20\right)$

$\leq e^{\left(-\frac{\left(\frac{4}{5}\right)^2 \cdot 20}{3}\right)} \leq 0.015$ *by the* Chernoff bound



Example: Frogs 🐸

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability at least one square ends up with at least 36 frogs.

A_i is the event the i'th square has at least 36 frogs

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_{25})$$
$$\leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots + \mathbb{P}(A_{25}) \quad \textit{by the union bound}$$
$$\leq 0.015 + 0.015 + 0.015 + \cdots + 0.015 = 25 \cdot 0.015 \quad \textit{by the Chernoff bound}$$
$$= 0.375$$

Example: Frogs 🐸

For an arbitrary location:

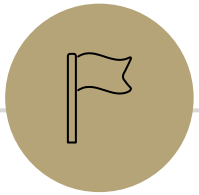
There are 100 frogs who could end up there (those above, below, left, right, and at that location). Each with probability **.2.** Let X be the number that land at the location we're interested in.

$$\mathbb{P}(X \geq 36) = \mathbb{P}(X \geq (1+\delta)20) \leq \exp\left(-\frac{\left(\frac{4}{5}\right)^2 \cdot 20}{3}\right) \leq 0.015$$

There are 25 locations. Since all locations are symmetric, by the union bound the probability of at least one location having 36 or more frogs is at most $25 \cdot 0.015 \leq 0.375$.

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** - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
  - 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)



Maximum Likelihood Estimation

Up till now…

So far, the probability questions we've asked have followed a pattern:

You're given a model with the probabilities you need to make predictions.

> $X \sim \text{Bin}(n, p)$, compute some probabilities about X, compute $E[X]$

> We have a distribution that takes on these outcomes with these probabilities. Compute the probability of some event or set of outcomes.

> *Before we run the entire experiment, let's make some predictions*

Up till now…

So far, the probability questions we've asked have followed a pattern:

You're given a model with the probabilities you need to make predictions.

> $X \sim \text{Bin}(n, p)$, compute some probabilities about X, compute $E[X]$

> We have a distribution that takes on these outcomes with these probabilities. Compute the probability of some event or set of outcomes.

> *Before we run the entire experiment, let's make some predictions*

In real world, we usually don't know all the rules of a random experiment
hence tail bounds, CLT, etc. to estimate probabilities in these situations
But, can we estimate those missing rules/parameters to a distribution?

What could those "missing rules/parameters" be?

We're going to call the unknown parameter(s) to a distribution θ

All distributions from the zoo we've are a ***distribution*** + ***parameter(s)*** $\boldsymbol{\theta}$*:*

> $\text{Ber}(p)$ --> $\theta = p$

> $\text{Poi}(\lambda)$ --> $\theta = \lambda$

> $\text{Unif}(a, b)$ --> $\theta = (a, b)$

Some ***probability distributions are in terms of some unknown parameter(s)*** $\boldsymbol{\theta}$

> *e.g.,* X follows the probability distribution $p_X(k) = \begin{cases} \theta & k = 1 \\ 2\theta & k = 2 \\ 1 - 3\theta & k = 3 \\ 0 & \text{otherwise} \end{cases}$

Our goal is to estimate the value of θ

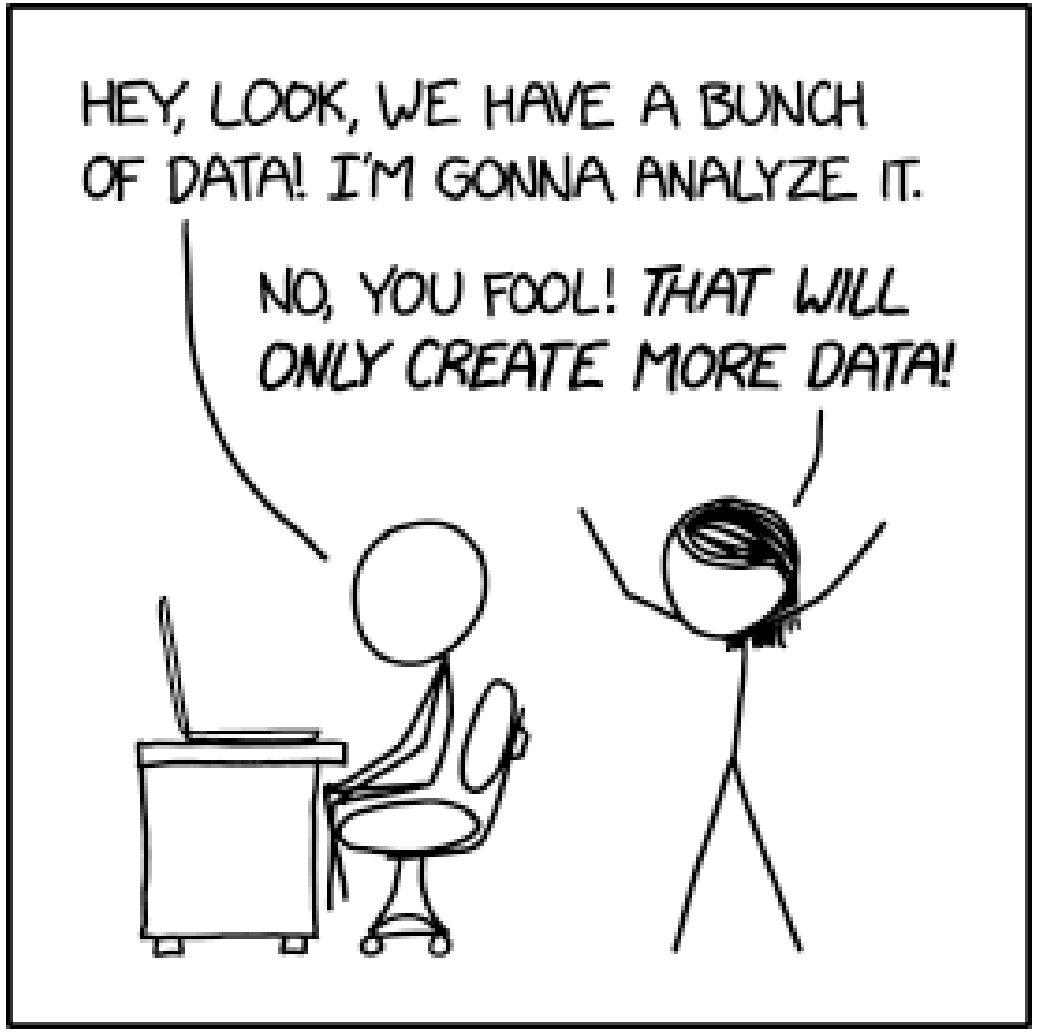
The Remix – going *backwards*

Let's say you have a coin. You don't know if it's fair or not.

It's reasonable to think that a coin flip follows a Bernoulli distribution.

But what's the parameter?

In the real world, you're not given parameters :(



https://xkcd.com/2582/

The Remix – going *backwards*

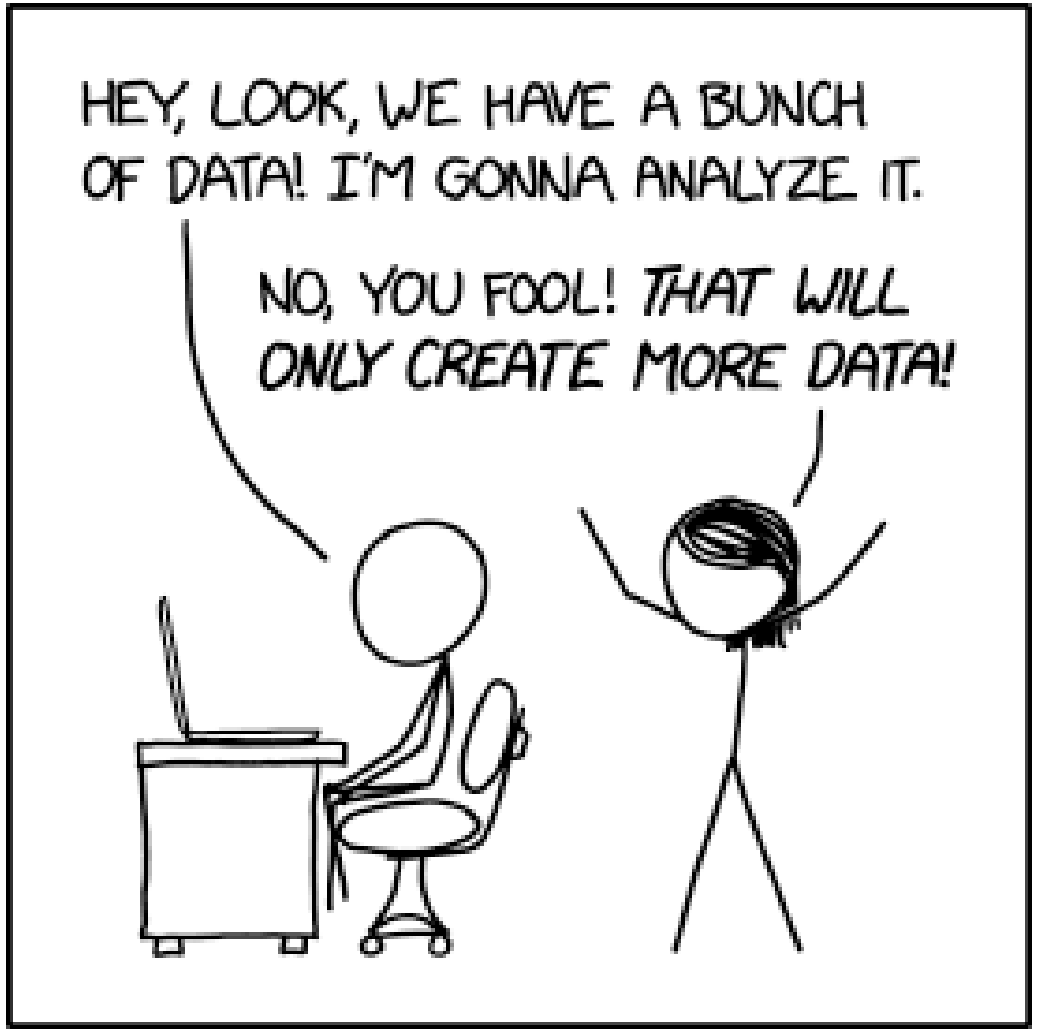
Let's say you have a coin. You don't know if it's fair or not.

It's reasonable to think that a coin flip follows a Bernoulli distribution.

But what's the parameter?

In the real world, you're not given parameters :(

But we do have... data!



https://xkcd.com/2582/

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results - HTTTH*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results – HTTTH…*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

(6 heads, 4 tails)

What is your estimate of p, the probability the coin comes up heads?

0.6

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *5 times and write down the results - HTTTH*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

(6 heads, 4 tails)

What is your estimate of p, the probability the coin comes up heads?

Maybe $\mathrm{p} = \frac{6}{10}$ ($X \sim \mathrm{Ber}(\frac{6}{10})$. *But how to argue "objectively" this is the best estimate?*

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results – HTTTH…*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results – HTTTH…*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

How do we do this? Well, we got some data - High probability events happen more often than low probability events. So, ***guess the rules that maximize the probability of the events we saw*** (relative to other choices of the rules).

e.g., what is the value of p *that makes the probability of seeing HTTTH… the highest?*

Maximum Likelihood Estimation

We derive an estimate $\hat{\theta}$ for the parameter θ based on observed data

1. We're going to run the random experiment a bunch of times (i.e., collect a bunch of samples from the distribution) -> this gives *data*

e.g., we flip a coin that follows $\mathrm{Ber}(p)$ *10 times and write down the results – HTTTH…*

2. Estimate the missing rules (unknown parameter(s)) *based on the data*

How do we do this? Well, we got some data - High probability events happen more often than low probability events. So, ***guess the rules that maximize the probability of the events we saw*** (relative to other choices of the rules).

e.g., what is the value of p *that makes the probability of seeing HTTTH… the highest?*

To do this, we will define a function that will tell us the probability of seeing particular data (a particular set of samples from the distribution) based on a particular value of the unknown parameter(s) θ

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E; \theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\theta) =$

"Probability of observing HTTTHHTHHH if θ is probability of heads on a single flip"

prob heads

6H, 4T

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

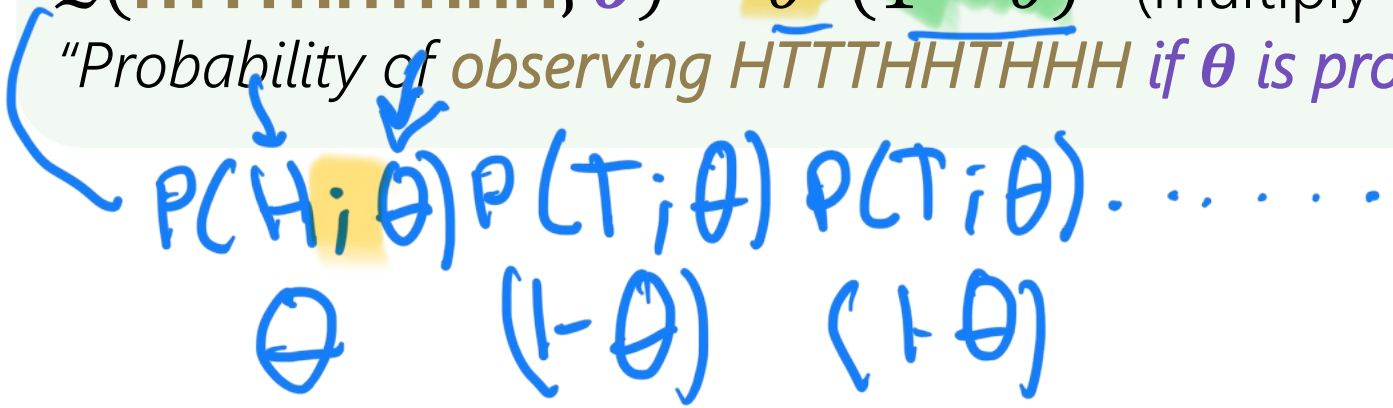
We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\theta) = \theta^6(1-\theta)^4$ (multiply because independent)

"Probability of observing HTTTHHTHHH if θ is probability of heads on a single flip"



Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\mathbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

"Probability of observing HTTTHHTHHH if $\boldsymbol{\theta}$ is probability of heads on a single flip"

Likelihood Function: Likelihood of n observations (from discrete distribution)

$$\mathcal{L}(x_1, x_2, \ldots, x_n; \theta) = \Pi_i^n \mathbb{P}(x_i; \theta)$$

PL

Notation comparison

$\mathbb{P}(X|Y)$ probability of X, conditioned on the **event** Y having happened (Y is a subset of the sample space).

$\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ. Since θ isn't an event (*it's not a subset of the sample space*), this is not conditioning.

We have a <u>fixed model</u>, want to find the probability of seeing some data

$\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ. *Likelihoods don't have all the properties we associate with probabilities (e.g. they don't all sum up to 1) and this isn't conditioning on an event (θ is a parameter/rule of how the event could be generated).*

We have some <u>fixed data</u>, we want to look at the probability of a particular model

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\theta) = \theta^6(1-\theta)^4$ (multiply because independent)

"Probability of observing HTTTHHTHHH if θ is probability of heads on a single flip"

Likelihood function

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

"what is probability of seeing the event E (in our case, the set of data), if the experiment is run with the parameter θ?"

We can't use probability notation because likelihood doesn't follow the same rules

Coin example

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\theta) = \theta^6(1-\theta)^4$ (multiply because independent)

"Probability of observing HTTTHHTHHH if θ is probability of heads on a single flip"

We want to pick the value of θ that make this likelihood function the largest. So...

Maximum Likelihood Estimation

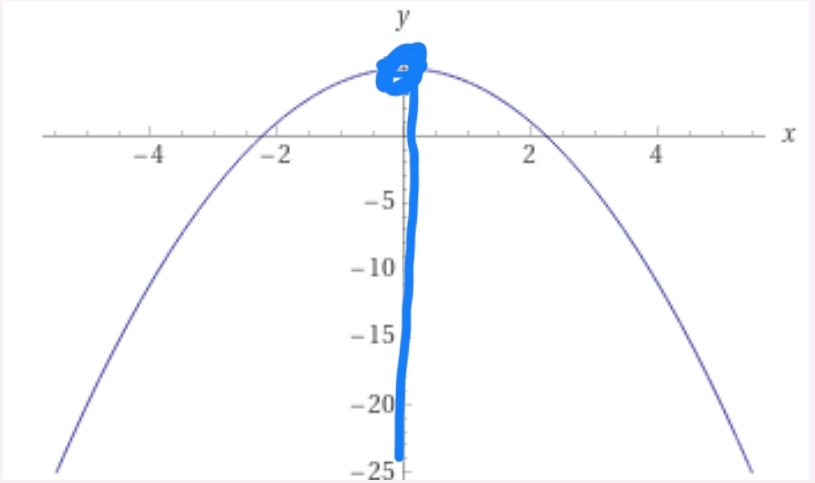
We will choose the estimator $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

"the value of θ that makes the likelihood of seeing the observed data the highest"

What is argmax?

For example $\text{argmax}_x \ (5 - x^2) = 0$

$\max_x(5 - x^2) = 5$, the input (argument) which produces 5 is 0, so the argmax is 0



Remember, our goal:

1. Collect some data/samples from the distribution
2. Find an estimate for θ

Maximum Likelihood Estimation

We will choose the estimator $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

"the value of θ that makes the likelihood of seeing the observed data the highest"

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$$

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

Maximum Likelihood Estimation

We will choose the estimator $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E; \theta)$

"the value of θ that makes the likelihood of seeing the observed data the highest"

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E; \theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

Use $\hat{\theta}_{\text{MLE}}$ if we want to emphasize how we found the estimator.

Remember, our goal:
1. Collect some data/samples from the distribution
2. Find an estimate for θ

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is: $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

Coin example (goal: estimate $\boldsymbol{\theta} = \boldsymbol{p}$, the probability of heads on a flip)

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

Now, find the value of θ that maximizes the likelihood...How do we find a max?

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is: $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

Coin example (goal: estimate $\boldsymbol{\theta} = \boldsymbol{p}$, the probability of heads on a flip)

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

Now, find the value of θ that maximizes the likelihood...How do we find a max?

Calculus!! 🥳 Take the derivative of $\mathcal{L}(\mathrm{E};\theta)$, set to 0, and solve for $\hat{\theta}$

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is: $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

Coin example (goal: estimate $\boldsymbol{\theta} = \boldsymbol{p}$, the probability of heads on a flip)

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

Now, find the value of θ that maximizes the likelihood…How do we find a max?

Calculus!! 🥳 Take the derivative of $\mathcal{L}(E;\theta)$, set to 0, and solve for $\hat{\theta}$

Take the derivative: $\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$

Set to 0 and solve: (now, we're solving for the *maximum* likelihood estimator, $\hat{\boldsymbol{\theta}}$)

$6\hat{\theta}^5(1-\hat{\theta})^4 - 4\hat{\theta}^6(1-\hat{\theta})^3 = 0 \Rightarrow 6(1-\hat{\theta}) - 4\hat{\theta} = 0 \Rightarrow -10\hat{\theta} = -6 \Rightarrow \hat{\theta} = \frac{3}{5} = 0.6$

The MLE $\hat{\theta}$ estimating the true $\boldsymbol{\theta} = \boldsymbol{p}$ is **3/5** just like we expected!

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is: $\hat{\theta} = \text{argmax}_\theta \ \mathcal{L}(E;\theta)$

Coin example (goal: estimate $\boldsymbol{\theta} = \boldsymbol{p}$, the probability of heads on a flip)

We ran the experiment 10 times independently. The result was **HTTTHHTHHH**

$\mathcal{L}(\textbf{HTTTHHTHHH};\boldsymbol{\theta}) = \theta^6(1-\theta)^4$ (multiply because independent)

Now, find the value of θ that maximizes the likelihood.

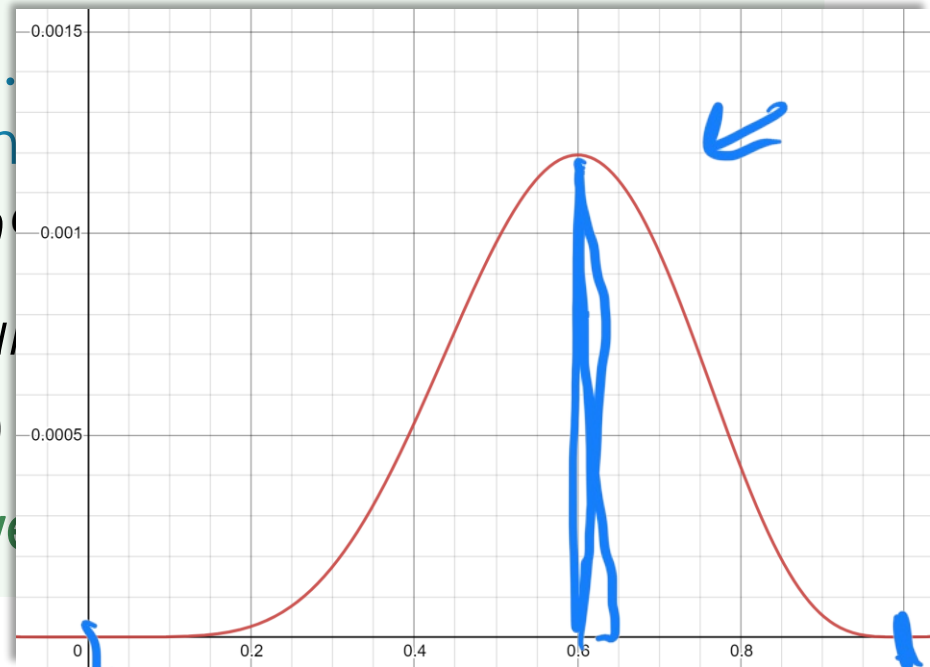
Calculus!! 🥳 Take the derivative of $\mathcal{L}(\mathrm{E};\theta)$, set to 0, an

Take the derivative: $\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^$

Set to 0 and solve: (now, we're solving for the *maximu*

$6\hat{\theta}^5(1-\hat{\theta})^4 - 4\hat{\theta}^6(1-\hat{\theta})^3 = 0 \Rightarrow 6(1-\hat{\theta}) - 4\hat{\theta} = 0$

The MLE $\hat{\theta}$ estimating the true $\boldsymbol{\theta} = \boldsymbol{p}$ is $\mathbf{3/5}$ just like we



Is that really the maximum?

What we really did was find the critical point (which could either be the maximum **or** the minimum), so ideally do **second derivative test** to check

1. Take the *second* derivative (the derivative of the derivative)
2. If negative everywhere around the critical point, it is the maximum

In this class, we won't ask you to do the second derivative test, you can assume the solution you find is a maximum ☺

> to sanity check your answer, at least make sure that the estimator you find is valid for what you are trying to estimate

Half a step backwards…

Since the likelihood function is a product of probabilities of seeing each of the samples, we're going to be taking the derivative of products a lot

The product rule is not fun!! There has to be a better way!

Half a step backwards…

Since the likelihood function is a product of probabilities of seeing each of the samples, we're going to be taking the derivative of products a lot

The product rule is not fun!! There has to be a better way!

Take the log of the likelihood function before taking the derivative!

Recall: $\ln(a \cdot b) = \ln(a) + \ln \text{product}(b)$

And, we don't need the product rule if our expression is a sum!

Can we still take the max? Yes! $\ln()$ is an increasing function, so

$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$

"the log of the likelihood will increase as the likelihood increases and vice verse, so, the value of θ that maximizes the log likelihood also maximized the likelihood"

Half a step backwards…

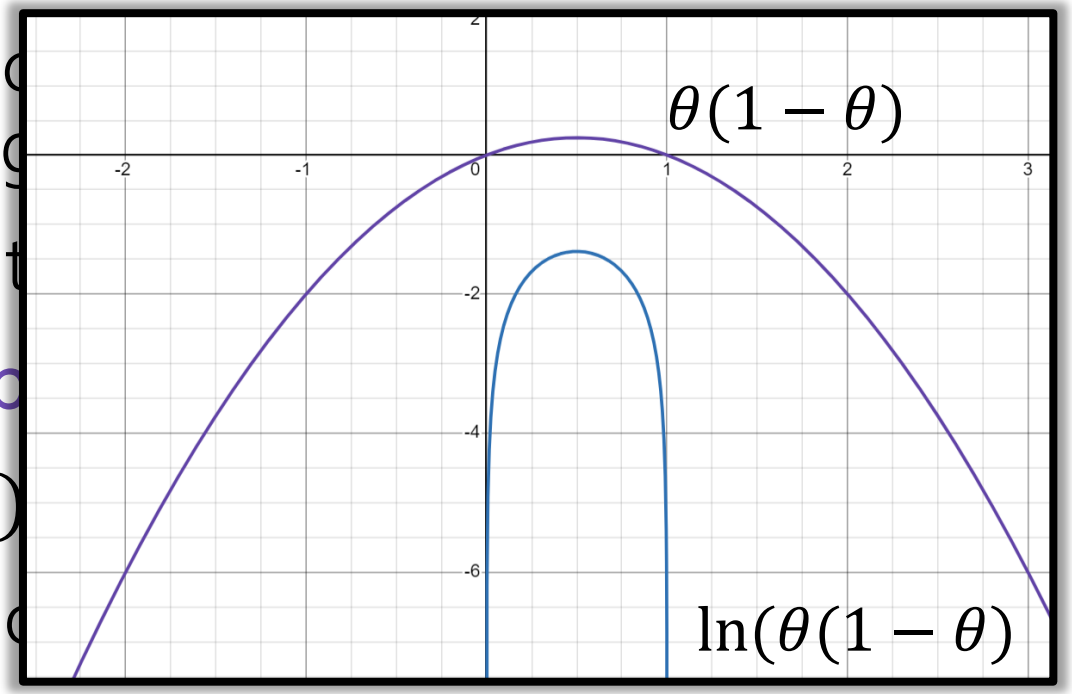
Since the likelihood function is a produc
of the samples, we're going to be taking

The product rule is not fun!! There has t

Take the log of the likelihood function b

Recall: $\ln(a \cdot b) = \ln(a) + \ln \text{product}(b)$

And, we don't need the product rule if



Can we still take the max? Yes! $\ln()$ is an increasing function, so

$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$

"the log of the likelihood will increase as the likelihood increases and vice verse, so, the value of θ that maximizes the log likelihood also maximized the likelihood"

Coin flips is easier

1. Likelihood function: $\mathcal{L}(\text{HTTTHHHTHHH}; \theta) = \theta^6(1-\theta)^4$
2. Take the log: $\ln(\mathcal{L}(\text{HTTTHHHTHHH}; \theta) = 6\ln(\theta) + 4\ln(1-\theta)$
3. Take the derivative: $\frac{d}{d\theta}\ln(\mathcal{L}(\cdot)) = \frac{6}{\theta} - \frac{4}{1-\theta}$

Derivative is much easier!!

4. Set to $\mathbf{0}$ and solve:

$$\frac{6}{\hat{\theta}} - \frac{4}{1-\hat{\theta}} = 0 \Rightarrow \frac{6}{\hat{\theta}} = \frac{4}{1-\hat{\theta}} \Rightarrow 6 - 6\hat{\theta} = 4\hat{\theta} \Rightarrow \hat{\theta} = \frac{3}{5}$$

5. Check it's a maximum (can skip in 312)

$\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ everywhere, so any critical point is a maximum.

Solving MLE (the process)

We're given that there's a distribution with some unknown parameter(s) θ. There are independent observations $x_1, x_2, \ldots, x_n$ from this distribution.

To find the MLE $\hat{\theta}$ for this unknown parameter(s) θ....

1. Write the likelihood function
   multiply (not add) probabilities of seeing each of the observations based on θ
2. Take the log $ln(..)$ of the likelihood function (makes the math easier)
3. Take the derivative of the log-likelihood function
4. Set the derivative to 0, and solve for the MLE $\hat{\theta}$
   remember to switch from θ to $\hat{\theta}$ in this step because we're now solving for the MLE
5. Verify it is a maximum with second derivative test (not required for 312)