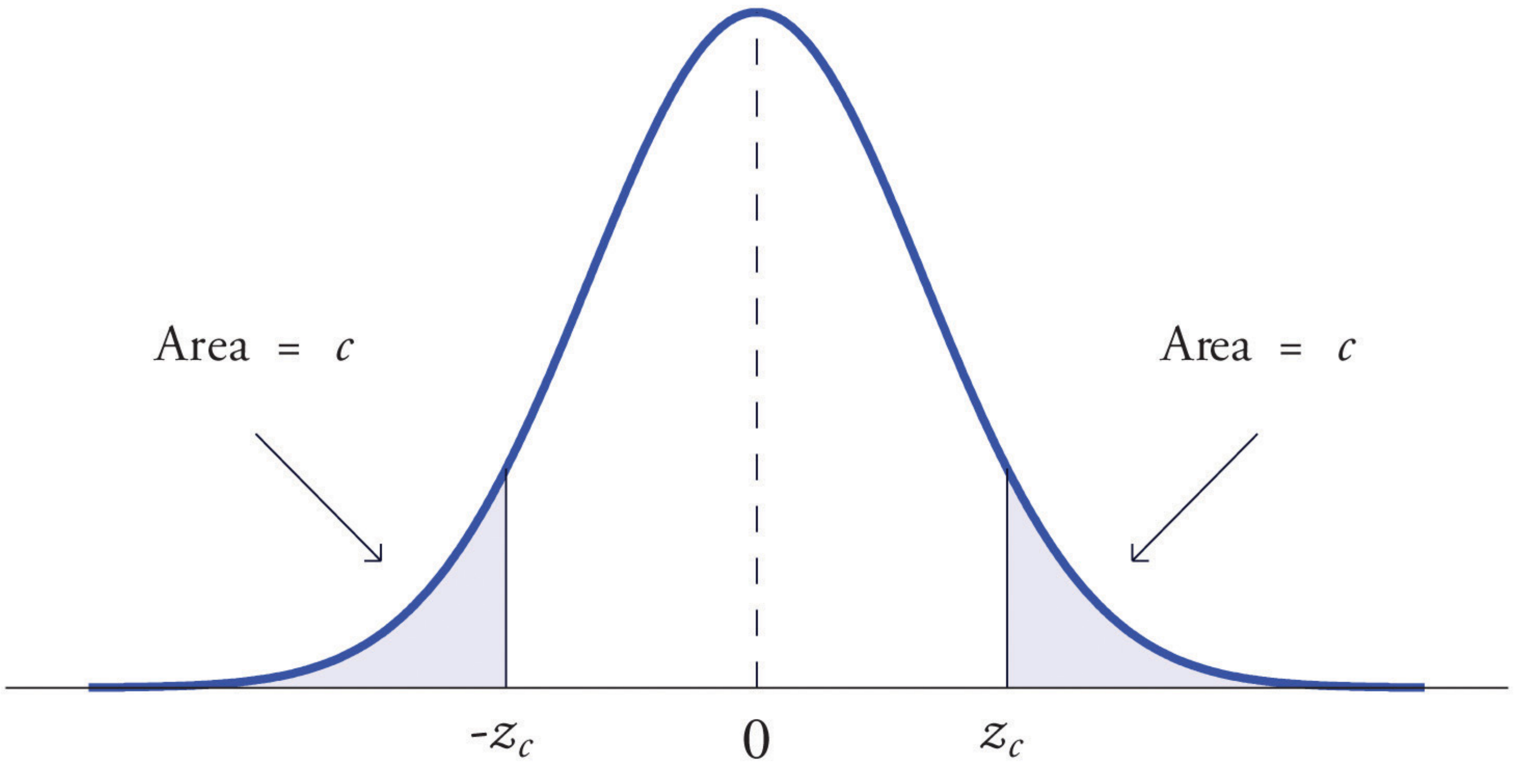
Lecture 19: Concentration Inequalities



Example: Coupon Collector

- In my childhood, there was a brand of instant noodles called little Raccoon (小浣熊).

- If you buy a bag, you get a uniformly random card from n cards.

- How many bag in expectation do you need to buy to collect all cards?

Example: Coupon Collector - Expectation

The trick:

Linearity of expectation

Let X be the number of bags until you collect all cards. We want $\mathbb{E}[X]$.

Let X_1 be the number of bags until you collect the first card ($X_1 = 1$ always)

Let X_2 be the number of additional bags until you collect the second card

……..

$X = X_1 + X_2 + \cdots + X_n$ is true in any outcome.

Example: Coupon Collector - Expectation

The trick:

Linearity of expectation

Let X be the number of bags until you collect all cards. We want $\mathbb{E}[X]$.

Let X_1 be the number of bags until you collect the first card ($X_1 = 1$ always)

Let X_2 be the number of additional bags until you collect the second card

....

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

Example: Coupon Collector - Expectation

The trick:

Linearity of expectation

Let X be the number of bags until you collect all cards. We want $\mathbb{E}[X]$.

$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$.

What is the distribution of X_i?

You have collected $i - 1$ cards. Every bag you buy, there is a $\frac{i-1}{n}$ chance you get a old card.

There is a $p = \frac{n-(i-1)}{n}$ chance you success and get a new card.

Example: Coupon Collector - Expectation

The trick:

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

You have collected $i-1$ cards. Every bag you buy, there is a $\frac{i-1}{n}$ chance you get a old card.

There is a $p_i = \frac{n-(i-1)}{n}$ chance you success and get a new card.

This is a Bernoulli process. $X_i \sim \text{Geometric}(p_i)$.

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-(i-1)}.$$

Example: Coupon Collector - Expectation

The trick:

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

You have collected $i-1$ cards. Every bag you buy, there is a $\frac{i-1}{n}$ chance you get a old card.

There is a $p_i = \frac{n-(i-1)}{n}$ chance you success and get a new card.

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-(i-1)}.$$

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \approx n \log n$$

Example: Coupon Collector - Variance

The trick:

$X_1, X_2, X_3 \cdots\cdots$ are independent geometric random variables.

Last lecture: Geometric(p_i) has variance $\frac{1-p_i}{{p_i}^2}$.

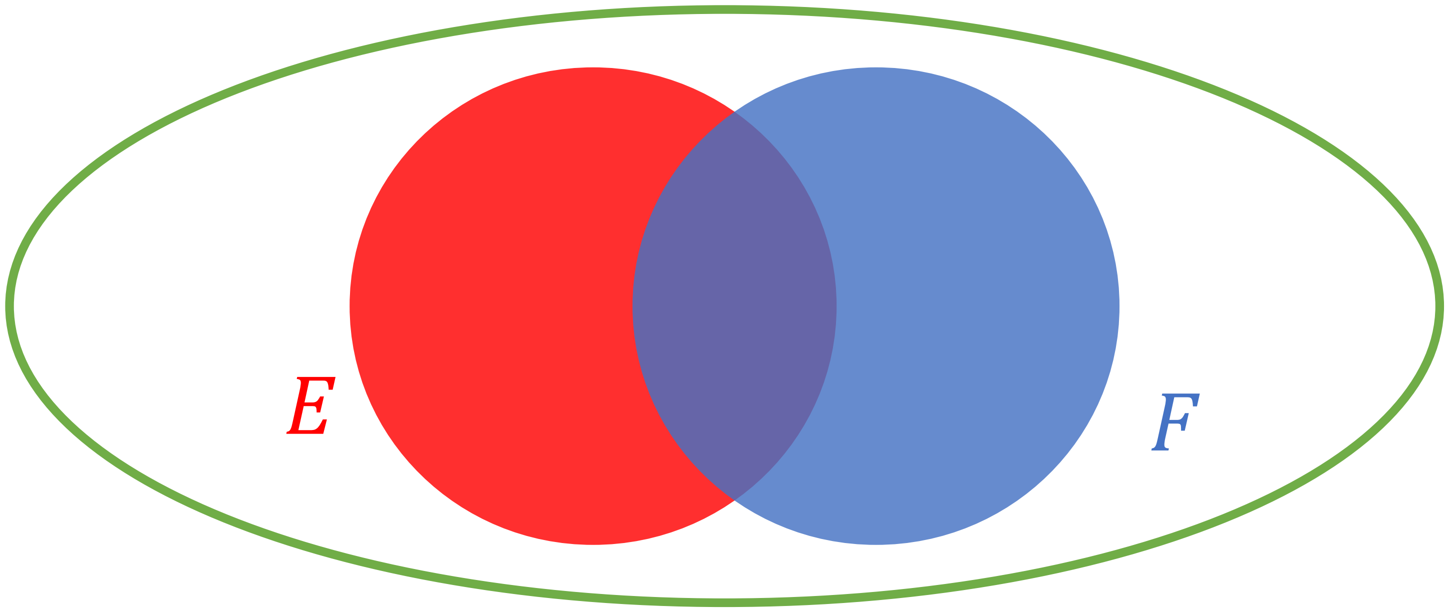
$$\text{Var}[X] = \text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_n] = \sum_{i=1}^{n} \frac{1-\left(\frac{n-i+1}{n}\right)}{\left(\frac{n-i+1}{n}\right)^2}$$

Recap: Inclusion-Exclusion

Inclusion Exclusion

Let E, F be two (not necessarily independent) events. We have

$$\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F]$$

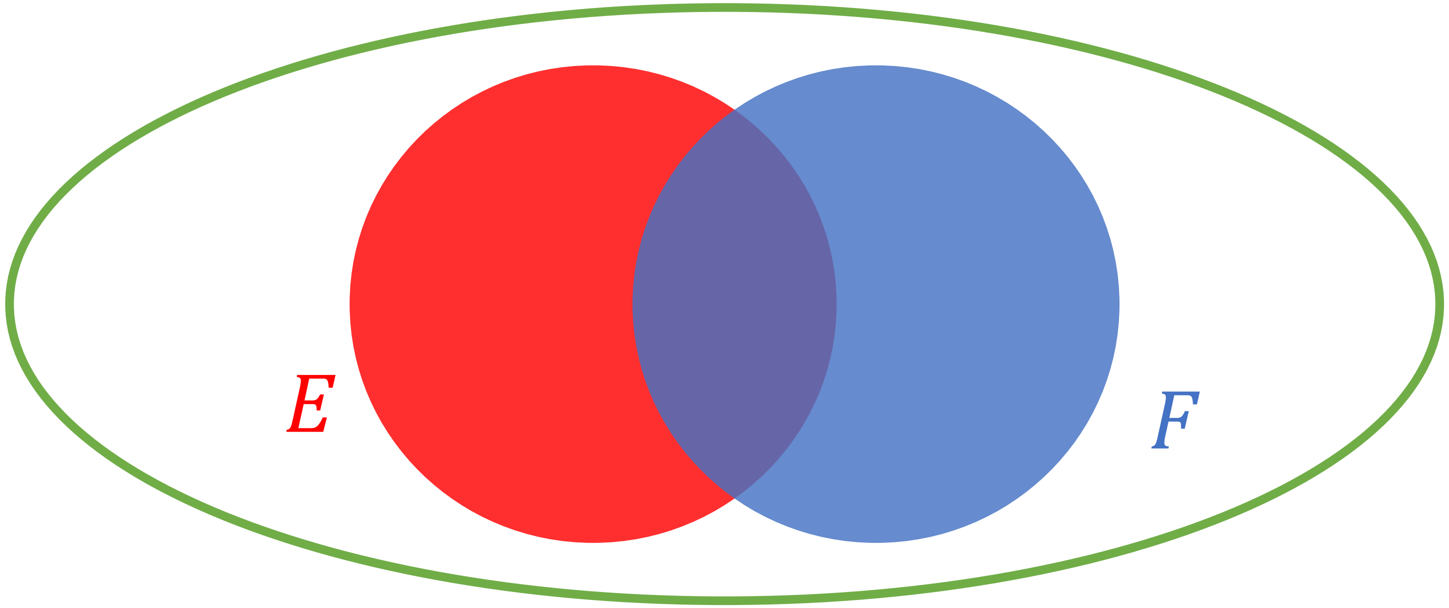


Union bound

Union bound

Let E, F be two (not necessarily independent) events. We have

$$\mathbb{P}[E \cup F] \leq \mathbb{P}[E] + \mathbb{P}[F] \sout{- \mathbb{P}[E \cap F]}$$



Union bound

Example.

Elon Musk's spaceship has 3,000,000 parts. Suppose each part has 10^{-20} probability of failure during the mission (not necessarily independent), and one failure could destroy the entire mission.

Can you give an estimate of the success probability of the mission?

Solution.

Apply union bound on $E_1, E_2, \dots, E_{3,000,000}$.

$$\mathbb{P}\left[E_1 \cup E_2 \cup \cdots \cup E_{3,000,000}\right] \le \sum_{i=1}^{3,000,000} \mathbb{P}[E_i] \le 3 * 10^6 * 10^{-20}$$

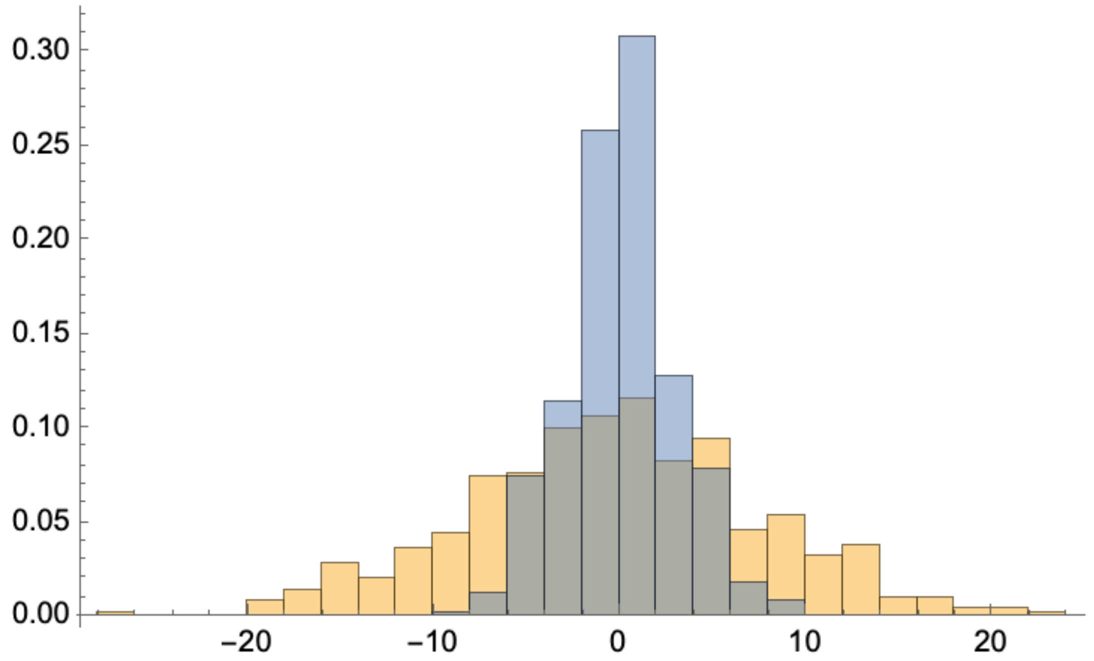
Concentration and tail bound

Intuition.

Previously, we saw two distributions.

The Blue distribution is more concentrated than the Orange one.

The Orange one is more uniform / spread out / uncertain….

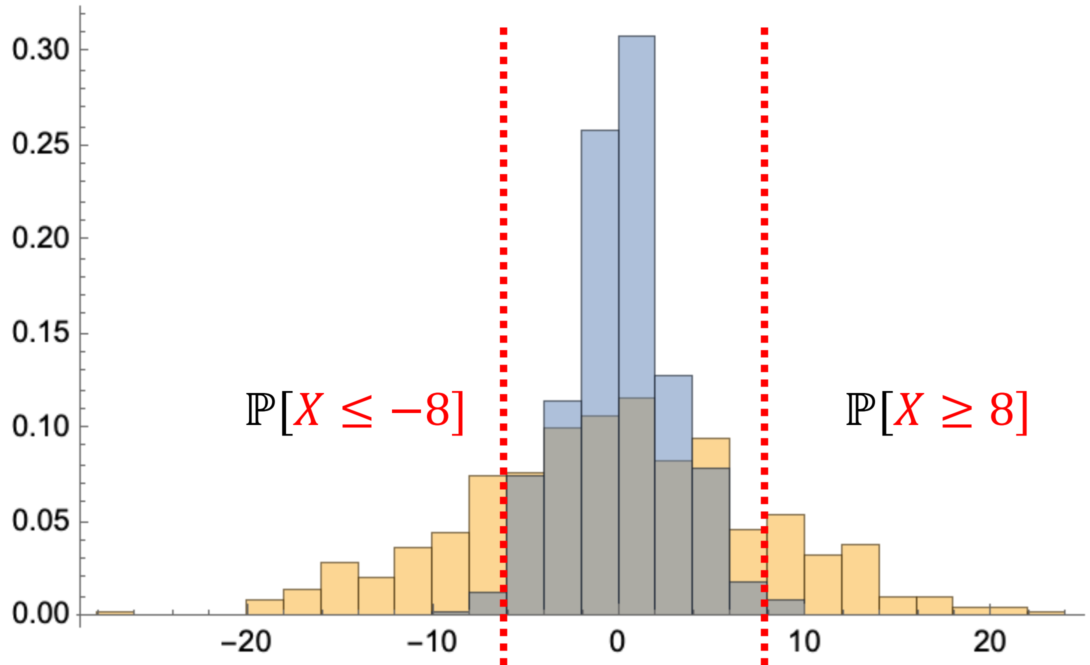


Concentration and tail bound

Intuition.

One way to compare them is by looking at variance.

The is another way: Looking at tail probabilities.



Markov's Inequality

Theorem.

Let X be a positive random variable.

We have

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$$

Proof.

$$\mathbb{E}[X] = \mathbb{P}[X \geq c] \cdot \mathbb{E}[X \mid X \geq c] + \mathbb{P}[X < c] \cdot \mathbb{E}[X \mid X < c]$$

(law of total expectation)

$$\geq \mathbb{P}[X \geq c] \cdot c + 0$$

(positivity)

Markov's Inequality

Theorem.

Let X be a positive random variable.

We have

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$$

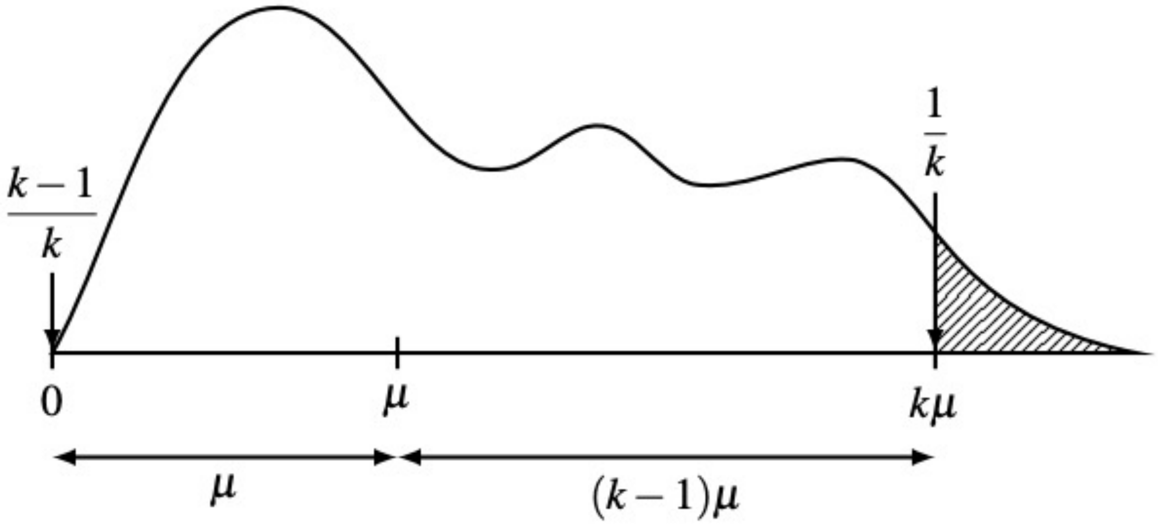


Figure 1: Markov's inequality interpreted as balancing a seesaw.

Markov's Inequality

Example (Lottery).

Say Hongxun bought a lottery and wins $X \geq 0$ dollars.

The lottery is worth $\mathbb{E}[X] = 10$ dollars.

What is the probability that Hongxun wins $X \geq 1{,}000{,}000$ dollars?

$$\mathbb{P}[X \geq 1{,}000{,}000] \leq \frac{10}{1{,}000{,}000} \leq 10^{-5} \text{ probability.}$$

Chebyshev's Inequality

Motivation.

Expectation $\mathbb{E}[X]$ is "first-moment" information.

Variance $\mathrm{Var}[X]$ is "second-moment" information.

With more information, can we give tighter (&two-sided) tail bound?

Theorem.

Let X be a random variable.

We have

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

Chebyshev's Inequality

Theorem.

Let X be a random variable.

We have

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

Proof.

$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

By Markov inequality,

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

Learn from Samples

Setup.

Say there is a coin with head probability p (fixed but unknown).

We can flip the coin and get independent samples

$$X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$$

How do we estimate p? How good is our estimation?

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Learn from Samples

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

How good is it?

Expectation: $\mathbb{E}[\hat{p}] = \frac{\mathbb{E}[X_1+X_2+\cdots+X_n]}{n} = \frac{n \cdot \mathbb{E}[X_1]}{n} = p$

Unbiased.

Learn from Samples

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

How good is it?

Variance: $\mathrm{Var}[\hat{p}] = \frac{\mathrm{Var}[X_1+X_2+\cdots+X_n]}{n^2} = \frac{n \cdot \mathrm{Var}[X_1]}{n^2}$

$$\mathrm{Var}[X_1] = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = p - p^2 = p(1-p)$$

$$\mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n}.$$

Learn from Samples

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

How good is it?

Chebyshev's Inequality:

$$\mathbb{P}[|\hat{p} - p| \geq c] \leq \frac{\mathrm{Var}[\hat{p}]}{c^2} = \frac{p(1-p)}{n \cdot c^2}$$

With n samples, to make this probability < 0.1, $c \leq \frac{1}{\sqrt{n}}$.

To get accuracy c, we need $n \approx \frac{1}{c^2}$ samples.

Learn from Samples

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

How good is it?

Chebyshev's Inequality:

$$\mathbb{P}[|\hat{p} - p| \geq c] \leq \frac{\mathrm{Var}[\hat{p}]}{c^2} = \frac{p(1-p)}{n \cdot c^2} = \frac{\text{constant}}{n \cdot c^2} \leq 0.1$$

$$\rightarrow. \; c^2 \geq \frac{\text{constant}}{n \cdot 0.1} = \frac{\text{constant}}{n}$$

With n samples, to make this probability < 0.1, $c \geq \frac{1}{\sqrt{n}}$.

To get accuracy c, we need $n \approx \frac{1}{c^2}$ samples.

Learn from Samples

Estimator.

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

How good is it?

Chebyshev's Inequality:

$$\mathbb{P}[|\hat{p} - p| \geq c] \leq \frac{\mathrm{Var}[\hat{p}]}{c^2} = \frac{p(1-p)}{n \cdot c^2}$$

Fix c. As $n \to \infty$, the probability $\mathbb{P}[|\hat{p} - p| \geq c] \to 0$.

Law of large numbers.

Law of large numbers

Theorem.

Let $X_1, X_2, \ldots, X_n$ be I.I.D. (independent & identically distributed) random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}[X_i] = \sigma^2$.

For every $\epsilon > 0$, as $n \to \infty$, we have

$$\mathbb{P}\left[\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right] \to 0$$

This justifies the foundation of the scientific paradigm of repeating experiments and taking their average.

Estimating Variance?

Biased Estimator.

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2$$

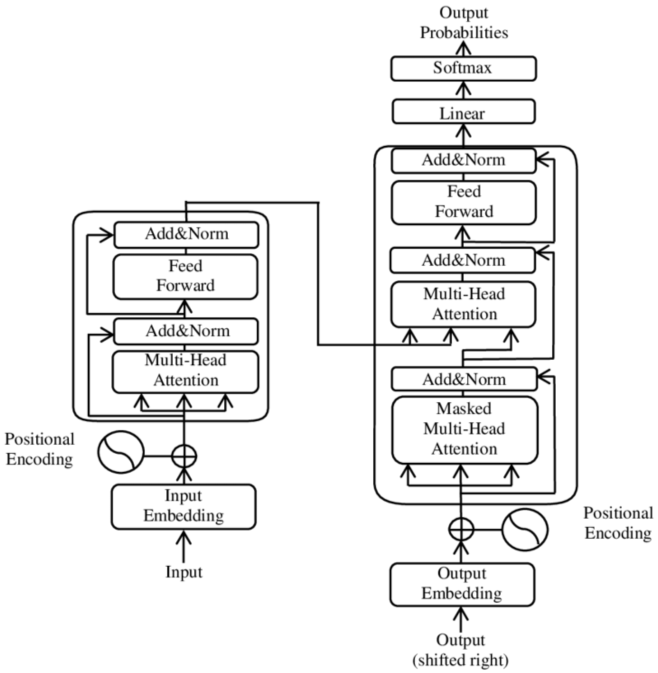
It does NOT satisfy $\mathbb{E}[\hat{V}] = \sigma^2$. In fact, it under-estimates σ^2.

Why? See discussion session.

Statistical learning & AI systems

The biggest thing these days.

Chatgpt



Language models

Statistical learning & AI systems

One view of the world:

The world can be viewed as a joint distribution:

Let $s = w_1 w_2 \cdots w_n$ be an English sentence.

$\mathbb{P}(\mathrm{s}) = \mathbb{P}[\mathrm{s}$ appears as a random sensible English sentence$]$

Example:

$\mathbb{P}(\text{one plus one equals two}) = 1\times10^{-7}$

$\mathbb{P}(\text{one plus one equals three}) = 1\times10^{-30}$

Statistical learning & AI systems

One view of the world:

The world can be viewed as a joint distribution:

Let $s = w_1 w_2 \cdots w_n$ be an English sentence.

$\mathbb{P}(\mathrm{s}) = \mathbb{P}[\mathrm{s}\text{ appears as a random sensible English sentence}]$

Inference:

select w_i that maximizes $\mathbb{P}(w_i | w_1, w_2, \ldots, w_{i-1})$

$\mathbb{P}(\text{two} \mid \text{one plus one equals }) = 0.9$

$\mathbb{P}(\text{three} \mid \text{one plus one equals}) = 0.01$

Statistical learning & AI systems

One view of the world:

The world can be viewed as a joint distribution:

Let $s = w_1 w_2 \cdots w_n$ be an English sentence.

$\mathbb{P}(\mathrm{s}) = \mathbb{P}[\mathrm{s}$ appears as a random sensible English sentence$]$

Inference:

select w_i that maximizes $\mathbb{P}(w_i | w_1, w_2, \ldots, w_{i-1})$

$\mathbb{P}(\text{sad} \mid \text{Hearing your loss, I am really}) = 0.4$

$\mathbb{P}(\text{sorry} \mid \text{Hearing your loss, I am really}) = 0.4$

$\mathbb{P}(\text{laughing} \mid \text{Hearing your loss, I am really}) = 10^{-7}$

Statistical learning & AI systems

One view of the world:

The world can be viewed as a joint distribution:

Let $s = w_1 w_2 \cdots w_n$ be an English sentence.

$\mathbb{P}(\mathrm{s}) = \mathbb{P}[\mathrm{s}$ appears as a random sensible English sentence$]$

The beauty of this view:

We want model output to be true facts / be grammatically correct / have emotion / ……

$\approx$ find s that maximizes this probability.

What used to be an issue: Curse of dimensionality

The issue in the past:

Suppose there are (n = 100) words

$$\mathbb{P}(s = w_1 w_2 \cdots w_n) = \mathbb{P}[s \text{ appears as a random sensible English sentence}].$$

With even just 100 most frequent words,

there are 100^{100} many probabilities to estimate.

Entire Internet size in 2023: 1.2×10^{17} MB.

The modern approach:

Instead assume that $\mathbb{P}(s = w_1 w_2 \cdots w_n) = \mathrm{f}_\theta(s)$ by a function f_θ with fewer parameters θ.

f_θ is your neural network (nowadays more specifically, your transformers.)

Magically one can learn θ from data and magically it works.