Probability in Computing

CS 237

LECTURE 20

Last time

- Discrete Distributions: Binomial, Geometric, Negative Binomial

Today

- Coupon Collector's Problem
- Reservoir Sampling

Reminder

- HW 10 is due Thursday

Top Hat question (Join Code: 413437)

For arbitrary random variables X and Y, by linearity of expectation:

A. $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

B. $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for all $a, b \in \mathbb{R}$.

C. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

D. Both A and B are correct.

E. A,B and C are correct.

Product of independent RVs

- Theorem. For any two **independent** random variables X and Y on the same probability space,

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

- Note. The equality does not hold, in general, for dependent random variables.

  Example. We toss two coins.

  Let X= number of HEADS, Y= number of TAILS.

  Calculate $\mathbb{E}(X)$, $\mathbb{E}(Y)$ and $\mathbb{E}(XY)$.

Top Hat question (Join Code: 413437)

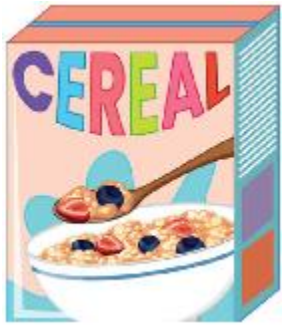
- What is the distribution of the number of rolls of a die until you see a 6?
- What is the expected number of rolls until you see a 6?

A. Geometric(6) and $1/6$
B. Binomial(6,1/6) and 6
C. Geometric(1/6) and 6
D. Geometric(5/6) and 6/5
E. None of the above

Coupon Collector's Problem

- There are n coupons

- Each cereal box has 1 coupon chosen uniformly and independently at random
- What is the expected number of boxes you need to buy to collect all n coupons?

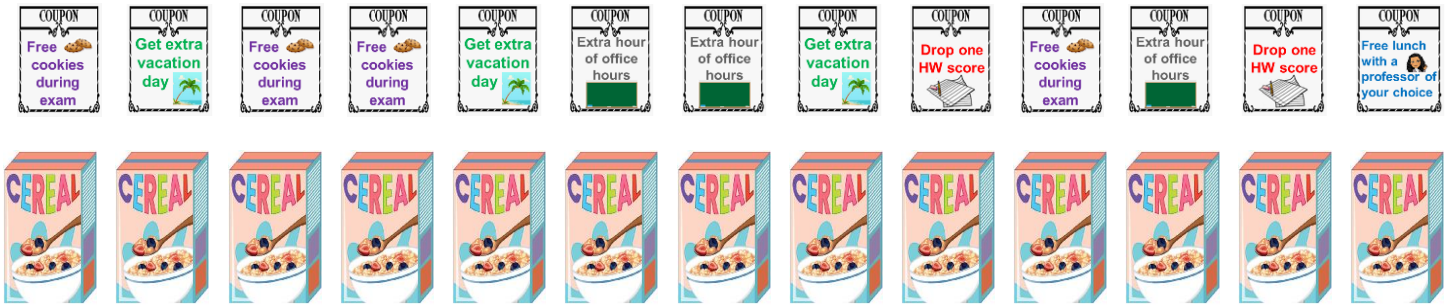


➢ X = the number of boxes bought to collect at least one copy of each coupon

➢ Find $\mathbb{E}(X)$

Example

$n = 5$



$X = 13$

Computing the Expectation

- X = the number of boxes bought to collect all n coupons
- Find $\mathbb{E}(X)$

We could try using the definition of expectation

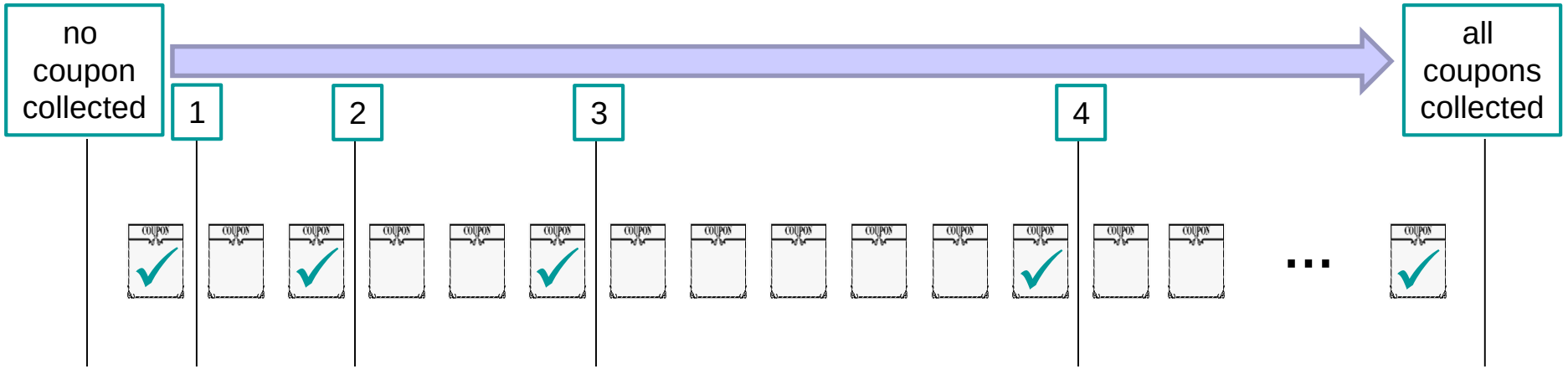
$$\mathbb{E}(X) = \sum_{k=n}^{\infty} k \cdot \underbrace{\Pr(X = k)}_{?}$$

Hard!

Breaking it Down into Phases

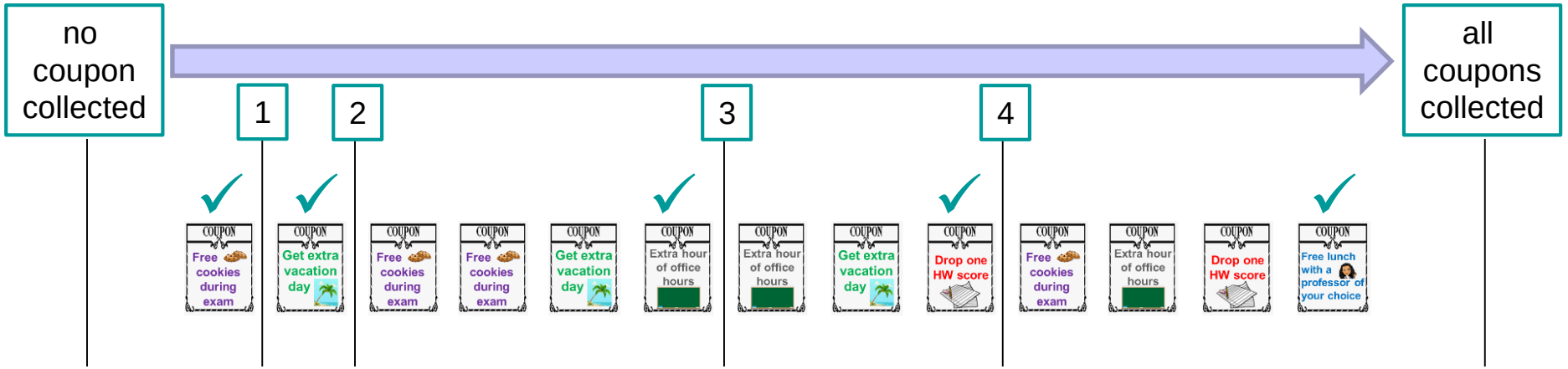
- This transition is not sudden
- Can you figure out how to break it down into phases?

Breaking it Down into Phases



Phases for Our Example

$n = 5$



Computing the Expectation

- X = the number of boxes bought to collect all n coupons
- Find $\mathbb{E}(X)$

- Represent X as a sum:

$$X = \sum_{i=1}^{n} X_i$$

where X_i = the number of boxes bought

from the moment $i - 1$ distinct coupons were collected

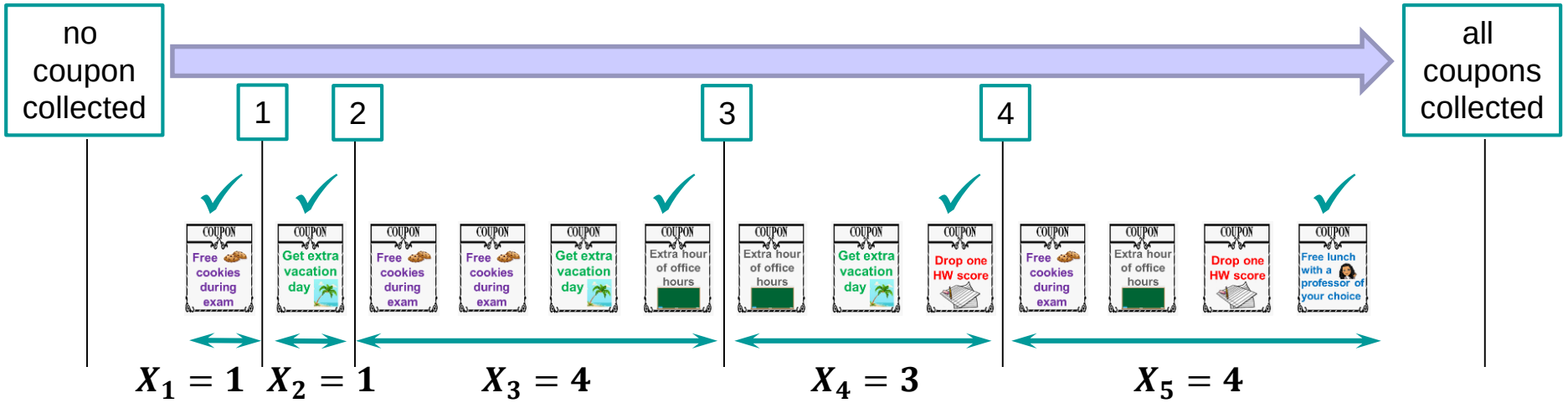
until i distinct coupons were collected

- By linearity of expectation,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i)$$

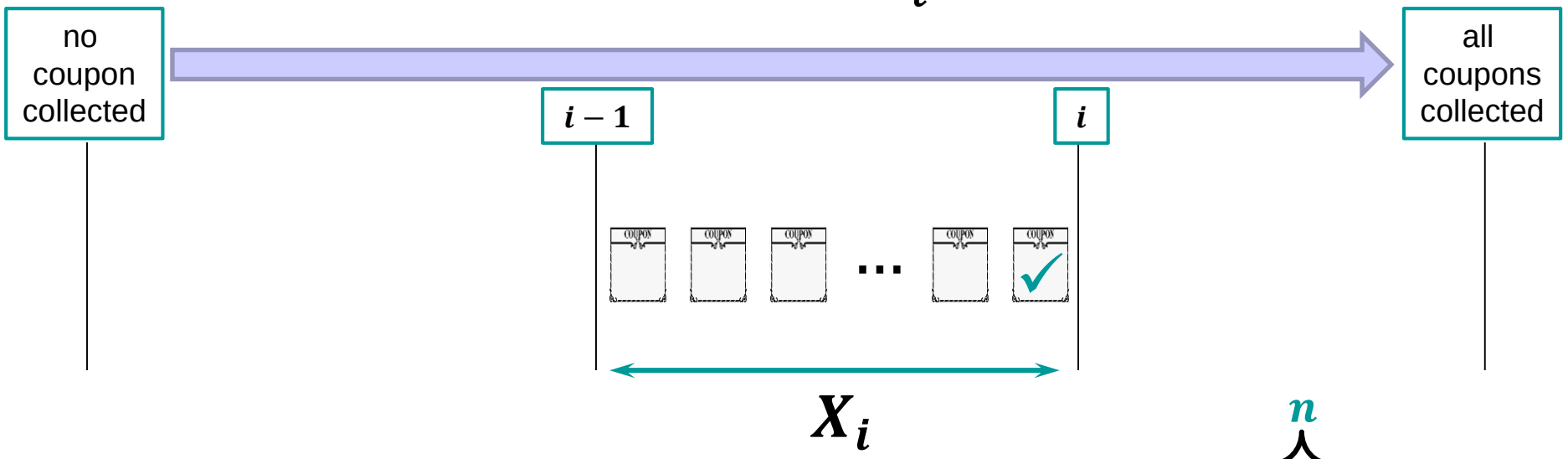
Random Variables X_i for Our Example

$n = 5$

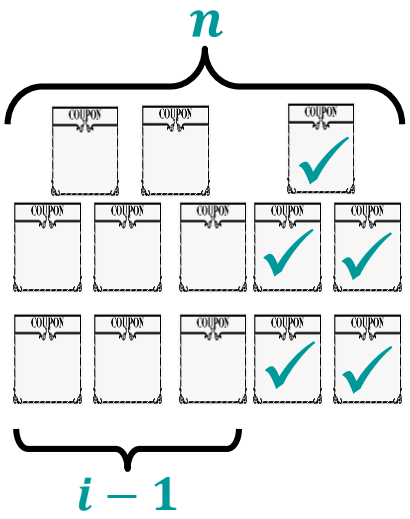


Calculating the Expectation of X_i

- What is the distribution of X_i?



- $X_i \sim Geom(p)$, where $p = \dfrac{n-i+1}{n}$
- $\mathbb{E}(X_i) =$



Computing the Expectation

- X = the number of boxes bought to collect all n coupons
- Find $\mathbb{E}(X)$
- By linearity of expectation,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1}$$

$$= n\underbrace{\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + 1\right)}_{\text{The } n^{th} \text{ harmonic number } H_n} = \boldsymbol{nH_n}$$

Coupon Collector's: Expectation

- X = the number of boxes bought to collect all n coupons
- Find $\mathbb{E}(X)$

$\mathbb{E}(X) = n\,H_n$, where $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is the n^{th} Harmonic number

- Lemma. $\ln n \leq H_n \leq \ln n + 1$
- Therefore, $n \ln n \leq \mathbb{E}(X) \leq n \ln n + n$
- Examples:

$n = 10$: $\mathbb{E}(X) \approx 29.2987$ $\quad n \ln n \approx 23.0258$

$n = 100$: $\mathbb{E}(X) \approx 518.738$ $\quad n \ln n \approx 460.517$

Application of Coupon Collector's

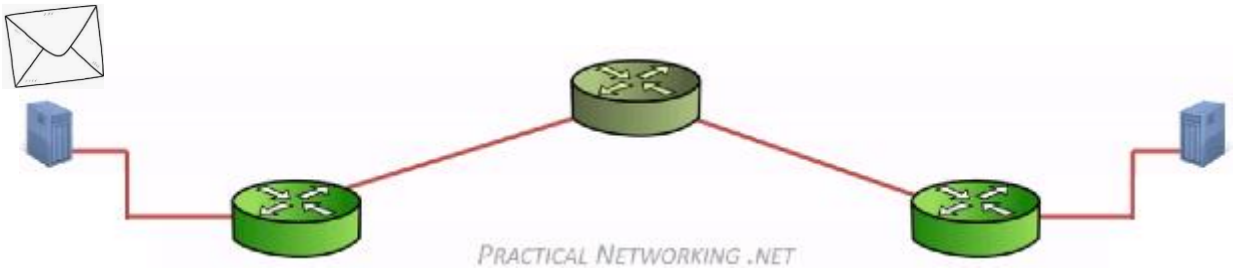
Coupon Collector's Problem has many applications in CS

Example:

- Packets passing along a fixed path of n routers.
- Designation host wants to collect IDs of the routers, but each packet has only space for one ID.

Idea: Sample the ID of a uniformly random router.

- What is the expected number of packets the designation host need to see in order to collect the IDs of all routers?



CS 237

Reservoir Sampling

How to sample a random item from a stream of unknown length

- Steam of items: $ID_1, ID_2, \ldots$
- The packet is seeing one item (ID) at a time
- The number of items is not known in advance

Figuring out how to do it: Thought Experiment

- Stream ends after the **first** item

Reservoir Sampling

How to sample a random item from a stream of unknown length

- Stream of items: $ID_1, ID_2, \ldots$
- The packet is seeing one item (ID) at a time
- The number of items is not known in advance

Figuring out how to do it: Thought Experiment

- Stream ends after the **second** item

- Stream ends after the n-th item

Reservoir Sampling: Correctness

Next we show that after each step, x is uniform item from the items we have seen so far

- Let $ID_1, \ldots, ID_n$ be the stream items
- Let X_i be the stored item x after processing $ID_1, \ldots, ID_i$
- We will prove by induction on i that $X_i \sim Uniform(ID_1, \ldots, ID_i)$

Reservoir Sampling: Correctness

Claim. For each $i \in \{1,2,\dots,n\}$,

$$X_i \sim Uniform(ID_1, \dots, ID_i)$$

Proof by induction:

Base case ($i = 1$):

Inductive step ($i > 1$):

Reservoir Sampling

https://blog.cloudera.com/blog/2013/04/hadoop-stratified-randosampling-algorithm/

Algorithms Every Data Scientist Should Know: Reservoir Sampling

April 23, 2013 | By Josh Wills (@josh_wills) (@josh_wills) | 3 Comments

Categories: Data Science How-to

Data scientists, that peculiar mix of software engineer and statistician, are notoriously difficult to interview. One approach that I've used over the years is to pose a problem that requires some mixture of algorithm design and probability theory in order to come up with an answer. Here's an example of this type of question that has been popular in Silicon Valley for a number of years:

Say you have a stream of items of large and unknown length that we can only iterate over once. Create an algorithm that randomly chooses an item from this stream such that each item is equally likely to be selected.

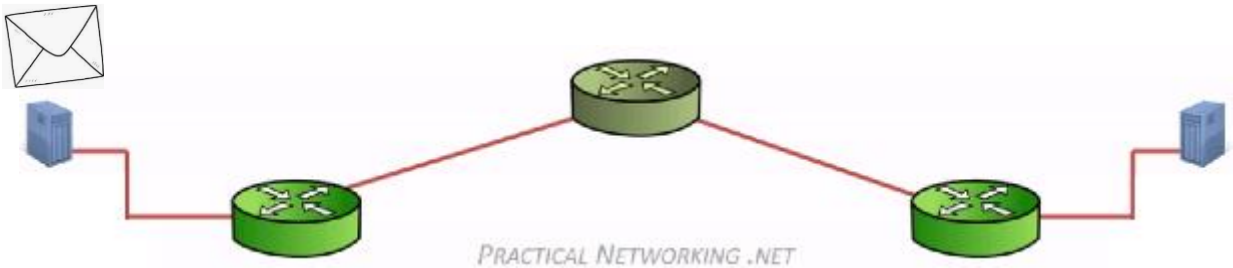
Application of Coupon Collector's

Coupon Collector's Problem has many applications in CS

Example:

- Packets passing along a fixed path of n routers.
- Designation host wants to collect IDs of the routers, but each packet has only space for one ID and a counter.

  Idea: Sample the ID of a uniformly random router.
- What is the expected number of packets the designation host need to see in order to collect the names of all routers?



Duration of a Random Experiment

Let random variable X denote the duration of a random experiment.

Approach to calculate $\mathbb{E}(X)$

- Carefully **partition** the experiment into phases.
- Calculate the expected duration of each phase.
- Use linearity of expectation to calculate $\mathbb{E}(X)$.

Sofya Raskhodnikova, Wayne Snyder; based on slides by Surender Baswana