INDICATOR RANDOM VARIABLES - II

DISCRETE STRUCTURES II

DARRYL HILL

BASED ON THE TEXTBOOK:

DISCRETE STRUCTURES FOR COMPUTER SCIENCE: COUNTING, RECURSION, AND PROBABILITY

BY MICHIEL SMID

Indicator Random Variables are Random Variables with a limited range.

An i.r.v. $X \in \{0,1\}$

Used for counting events.

One side effect:

$$E(X) = \sum_k k \cdot \Pr(X = k)$$

$$= 0 \cdot \Pr(X = 0) + 1 \cdot \Pr(X = 1)$$

$$= \Pr(X = 1)$$

So the expected value of an Indicator Random Variable is the probability that the indicated event will happen.

This makes them easy to compute.

n students $S_1, \ldots, S_n$ have taken a test

How many cheated?

$k =$ number of cheaters.

We want to estimate k.

for $i = 1, \ldots, n$:

I ask: "Hi S_i, did you cheat?“

Let X be the number of students who answer “yes”.

What is $E(X)$?

Remarkably, $E(X) = 0$.

We want a way to poll the students anonymously, so they can answer honestly.

n students $S_1, \dots, S_n$

$k =$ number of cheaters (unknown).

Estimate k without finding out who the cheaters are. Algorithm:

for $i = 1, \dots, n$: "Hi S_i, did you cheat?"

S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer

if TH: S_i replies "I cheated“
if TT: S_i replies "I did not cheat"

(regardless of whether they cheated or not)

I ask, the student says “I cheated”, what are the possibilities?

1. Student flipped HH or HT and are telling the truth (they actually cheated)

2. Student flipped TH (must reply “I cheated”) and they actually did cheat.

3. Student flipped TH (must reply “I cheated”) and they did NOT cheat.

n students $S_1, \ldots, S_n$

$k =$ number of cheaters (unknown).

Estimate k without finding out who the cheaters are.

for $i = 1, \ldots, n$: "Hi S_i, did you cheat?"

S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer

if TH: S_i replies "I cheated“
if TT: S_i replies "I did not cheat"

(regardless of whether they cheated or not)

I ask, the student says “I didn’t cheat”, what are the possibilities?

1. Student flipped HH or HT and are telling the truth (they didn’t cheat)

2. Student flipped HT (must reply “I didn’t cheat”) and they actually did cheat.

3. Student flipped HT (must reply “I didn’t cheat”) and they did not cheat.

n students $S_1, \ldots, S_n$
k = number of cheaters (unknown).
for $i = 1, \ldots, n$: "Hi S_i, did you cheat?"
S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer
if TH: S_i replies "I cheated"
if TT: S_i replies "I did not cheat"

X = # students replying "I cheated“

What is $E(X) = ?$

Define an indicator random variable:

$$X_i = \begin{cases} 1 \text{ if } S_i \text{ says "I cheated"} \\ 0 \text{ if } S_i \text{ says "I didn't cheat"} \end{cases}$$

$$X = X_1 + X_2 + \cdots + X_n$$

Thus

$$E(X) = E(X_1 + X_2 + \cdots + X_n)$$
$$= E(X_1) + E(X_2) + \cdots + E(X_n)$$

$$E(X_i) = 0 \cdot \Pr(X_i = 0) + 1 \cdot \Pr(X_i = 1)$$
$$E(X_i) = \Pr(X_i = 1)$$

We need to determine the $\Pr(X_i = 1)$.

n students $S_1, \dots, S_n$
$k =$ number of cheaters (unknown).
for $i = 1, \dots, n$: "Hi S_i, did you cheat?"
S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer
if TH: S_i replies "I cheated"
if TT: S_i replies "I did not cheat"

$X =$ # students replying "I cheated“

$$X_i = \begin{cases} 1 \text{ if } S_i \text{ says "I cheated"} \\ 0 \text{ if } S_i \text{ says "I didn't cheat"} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1)$$

In this case, $E(X_i)$ will depend on if S_i actually cheated or not.

If S_i cheated:

$$\Pr(X_i = 1)$$

$$= \Pr(HH \text{ or } HT \text{ or } TH)$$

$$= {}^{3}/_{4}$$

If S_i did not cheat:

$$\Pr(X_i = 1)$$

$$= \Pr(TH)$$

$$= {}^{1}/_{4}$$

n students $S_1, \dots, S_n$
k = number of cheaters (unknown).
for $i = 1, \dots, n$: "Hi S_i, did you cheat?"
S_i flips a fair coin twice (doesn't show result)

if HH or HT:	S_i gives an honest answer
if TH:	S_i replies "I cheated"
if TT:	S_i replies "I did not cheat"

X = # students replying "I cheated“

$$X_i = \begin{cases} 1 \text{ if } S_i \text{ says "I cheated"} \\ 0 \text{ if } S_i \text{ says "I didn't cheat"} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1)$$

k = number of cheaters (unknown).

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

For every student S_i who cheated, $E(X_i) = \frac{3}{4}$

For every student S_i who did not, $E(X_i) = \frac{1}{4}$

$$E(X) = \frac{3}{4} \cdot k + \frac{1}{4} \cdot (n - k)$$

$$= \frac{n}{4} + \frac{k}{2}$$

We still don’t know k, but we can solve for k.

n students $S_1, \ldots, S_n$
k = number of cheaters (unknown).
for $i = 1, \ldots, n$: "Hi S_i, did you cheat?"
S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer
if TH: S_i replies "I cheated"
if TT: S_i replies "I did not cheat"

X = # students replying "I cheated“

Define a new random variable:

$$Y = 2X - \frac{n}{2}$$

$$E(X) = \frac{n}{4} + \frac{k}{2}$$

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

Linearity of expectation:

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

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

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

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

$$\mathrm{E(Y)} = 2 \cdot \left(\frac{n}{4} + \frac{k}{2}\right) - \frac{n}{2}$$

$$E(Y) = k$$

n students $S_1, \dots, S_n$
k = number of cheaters (unknown).
for $i = 1, \dots, n$: "Hi S_i, did you cheat?"
S_i flips a fair coin twice (doesn't show result)

if HH or HT: S_i gives an honest answer
if TH: S_i replies "I cheated"
if TT: S_i replies "I did not cheat"

X = # students replying "I cheated“

Define a new random variable:

$$Y = 2X - \frac{n}{2}$$

Thus $E(Y) = k$.

If we run the algorithm, count the number of students who reply “I cheated” (X), then apply

$$Y = 2X - \frac{n}{2}$$

on average we will have

$$Y = \text{the number of cheaters}$$

Also I have no idea who the cheaters are.

```
FindMax(S_1, …, S_n):
  max = −∞;
  for i ∈ (1, …, n):
     if S_i > max:
        max = S_i;   *
  return max;
```

Indicator Random Variables can be used in algorithms with a random component.

How many times is * executed?

How many times does the variable max get a new value?

This depends on the permutation.

Examples:

3,2,4,1,6,5

6,5,4,3,2,1

1,2,3,4,5,6

```
FindMax(S_1, ..., S_n) :
  max = −∞;
  for i ∈ (1, ..., n):
     if S_i > max:
        max = S_i ;    *
  return max ;
```

Indicator Random Variables can be used in algorithms with a random component.

How many times is * executed?

How many times does the variable max get a new value?

This depends on the permutation.

Examples:

$\boxed{3}$,2,$\boxed{4}$,1,$\boxed{6}$,5 -> 3

$\boxed{6}$,5,4,3,2,1 -> 1

1,2,3,4,5,6 -> 6

We will say that $S_1, S_2, \dots S_n$ is a uniformly random permutation of $\{1,2,\dots,n\}$.

That is, each of the $n!$ permutations occurs with probability $1/_{n!}$

```
FindMax(S_1, ..., S_n):
  max = −∞;
  for i ∈ (1, ..., n):
     if S_i > max:
        max = S_i;    *
  return max;
```

$X = \#$ of times $*$ is executed

What is $E(X)$?

We want to use indicator random variables.

For $i = 1, \dots, n$:

$$X_i = \begin{cases} 1 \text{ if } * \text{ is executed in iteration } i \\ 0 \text{ otherwise} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1)$$

1	2	...	i	i+1	...	n-1	n
S_1	S_2		S_i	S_{i+1}		S_{n-1}	S_n

For event $X_i = 1$ to happen, the largest of all values from $1 \dots i$ is at S_i.

Since the first i numbers are in random order, the largest is in locations $1 \dots i$ with equal probability.

$$\Pr(X_i = 1) = \frac{1}{i}$$

That is our educated guess.

```
FindMax(S_1, ..., S_n) :
  max = -∞;
  for i ∈ (1, ..., n):
     if S_i > max:
        max = S_i ;    *
  return max ;
```

$X = \#$ of times $*$ is executed

What is $E(X)$?

We want to use indicator random variables.

For $i = 1, \ldots, n$:

$$X_i = \begin{cases} 1 \text{ if } * \text{ is executed in iteration } i \\ 0 \text{ otherwise} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1)$$

1	2	…	i	i+1	…	n-1	n
S_1	S_2		S_i	S_{i+1}		S_{n-1}	S_n

How many permutations of $\{1 \ldots n\}$ have the largest of the first i numbers at position i?

Choose the first i values.
$\binom{n}{i}$ ways to do that.
Put the largest at position i – 1 way.
Put the rest in positions $1 \ldots i - 1$
$(i-1)!$ ways
Place the remaining $n - i$ values
$(n-i)!$ ways.

```
FindMax(S_1, ..., S_n) :
  max = -∞;
  for i ∈ (1, ..., n):
      if S_i > max:
          max = S_i ;    *
  return max ;
```

$X = \#$ of times $*$ is executed

What is $E(X)$?

We want to use indicator random variables.

For $i = 1, \dots, n$:

$$X_i = \begin{cases} 1 \text{ if } * \text{ is executed in iteration } i \\ 0 \text{ otherwise} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1) = \frac{1}{i}$$

1	2	…	i	i+1	…	n-1	n
S_1	S_2		S_i	S_{i+1}		S_{n-1}	S_n

How many permutations of $\{1 \dots n\}$ have the largest of the first i numbers at position i?

Let $A =$ the largest so far is at i

$$|A| = \binom{n}{i} \cdot 1 \cdot (i-1)! \cdot (n-i)!$$

$$= \frac{n!}{i! \cdot (n-i)!} \cdot (i-1)! \cdot (n-i)!$$

$$\Pr(A) = \frac{|A|}{|S|} = \frac{1}{n!} \cdot \frac{n!}{i! \cdot (n-i)!} \cdot (i-1)! \cdot (n-i)!$$

$$= \frac{1}{i}$$

```
FindMax(S_1, ..., S_n):
  max = -∞;
  for i ∈ (1, ..., n):
     if S_i > max:
        max = S_i;    *
  return max;
```

$X = \#$ of times $*$ is executed

What is $E(X)$?

For $i = 1, \dots, n$:

$$X_i = \begin{cases} 1 \text{ if } * \text{ is executed in iteration } i \\ 0 \text{ otherwise} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1) = \frac{1}{i}$$

$$X = X_1 + X_2 + \ \dots + X_n$$
$$E(X) = E(X_1 + X_2 + \ \dots + X_n)$$
$$E(X) = E(X_1) + E(X_2) + \dots + E(X_n)$$
$$E(X) = \Pr(X_1) + \Pr(X_2) + \dots + \Pr(X_n)$$
$$E(X) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{n}$$

There is no closed form for this, so it was given a name:

“Harmonic number n“ is H_n (the n th harmonic number).

This is about $\ln n$ (natural log of n).

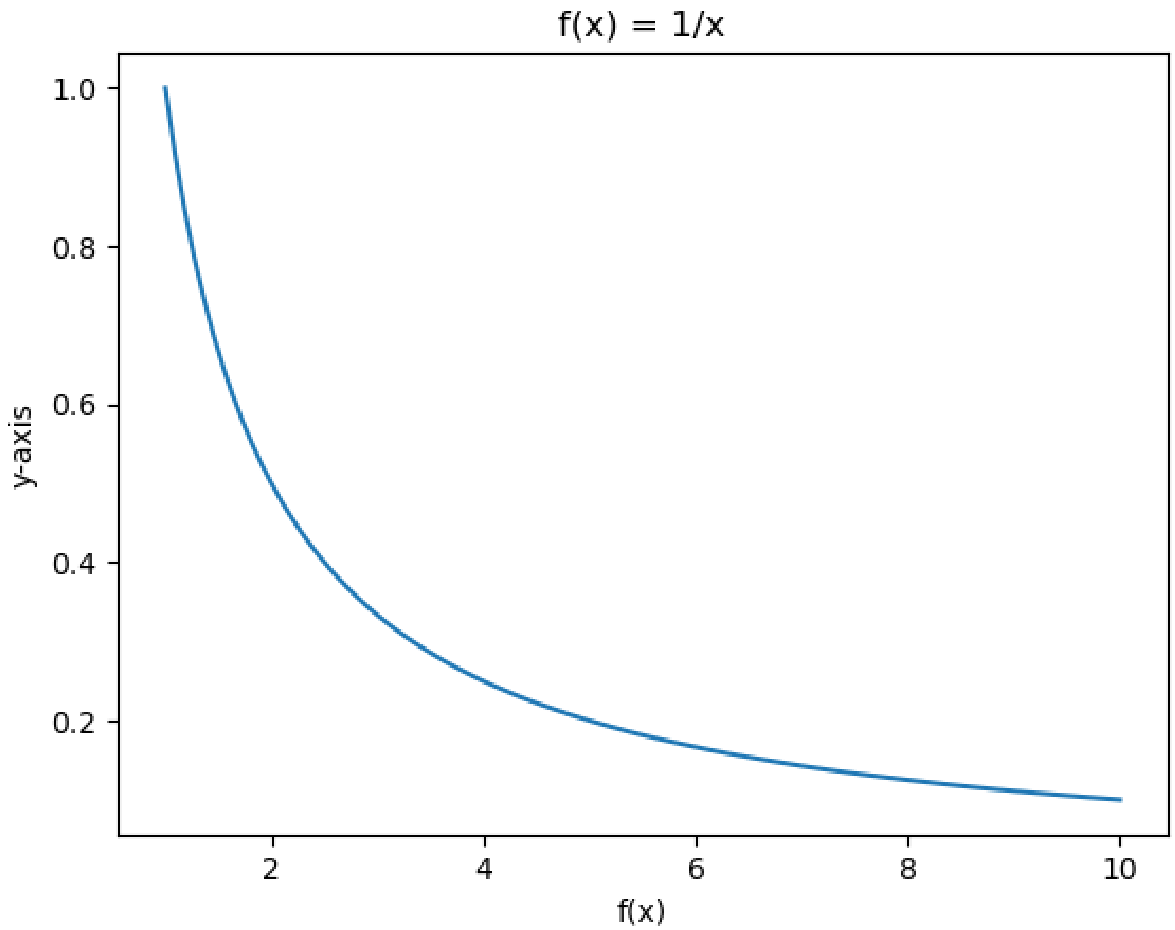
f is a decreasing positive function

e.g. $f(x) = \frac{1}{x}, x > 0$

We want to estimate a function with discrete input:

$$f(1) + f(2) + f(3) + \cdots + f(n)$$

We will estimate it using the plot of the continuous function (e.g., 1/x)



$$f(1) + \underbrace{f(2) + f(3) + \cdots + f(n)}$$

$=$ total area of the rectangles

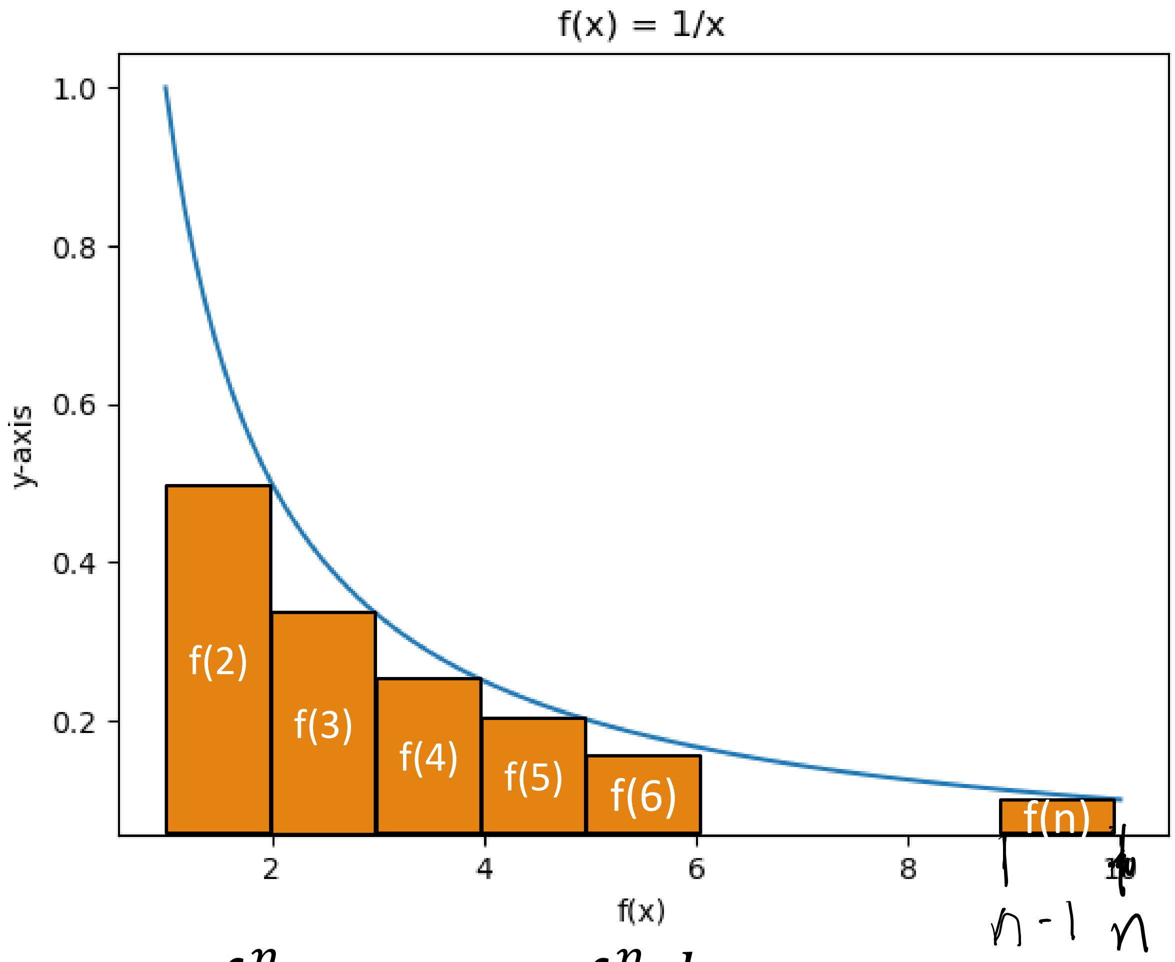
What can we say about the area of the rectangles?

$\leq$ area under the function starting from 1 (under the blue line).

The area under the line (from 1 to n) is given by:

$$= \int_1^n f(x)dx$$

For $f(x) = {}^1\!/_x$ we get:



$$\int_1^n f(x)dx = \int_1^n \frac{dx}{x} = \ln n - \underbrace{\ln 1}_{0}$$

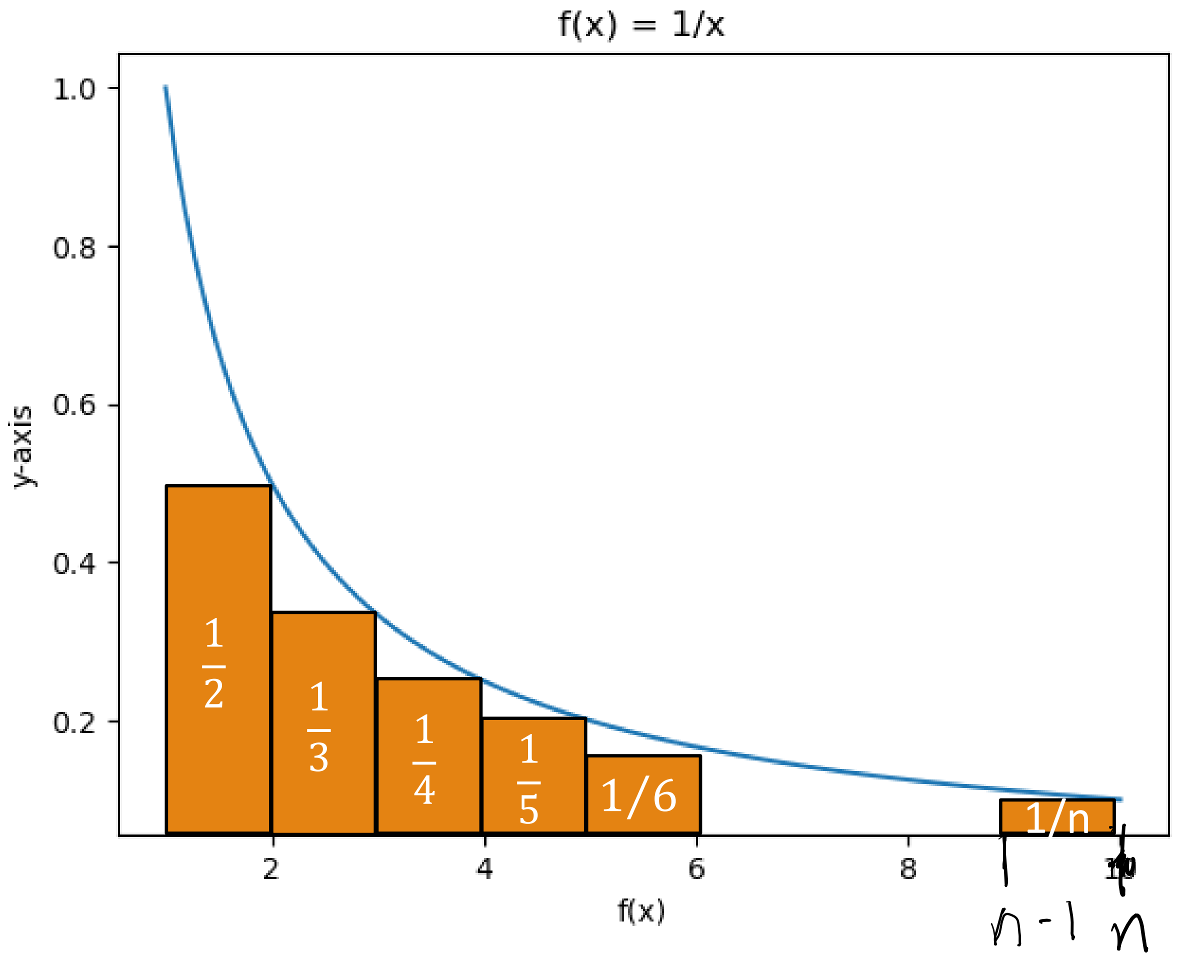
$$= \ln n$$

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

$$\leq 1 + \boxed{\int_1^n \frac{dx}{x}}$$

$$= 1 + \boxed{\ln n}$$

$$H_n \leq 1 + \ln n$$

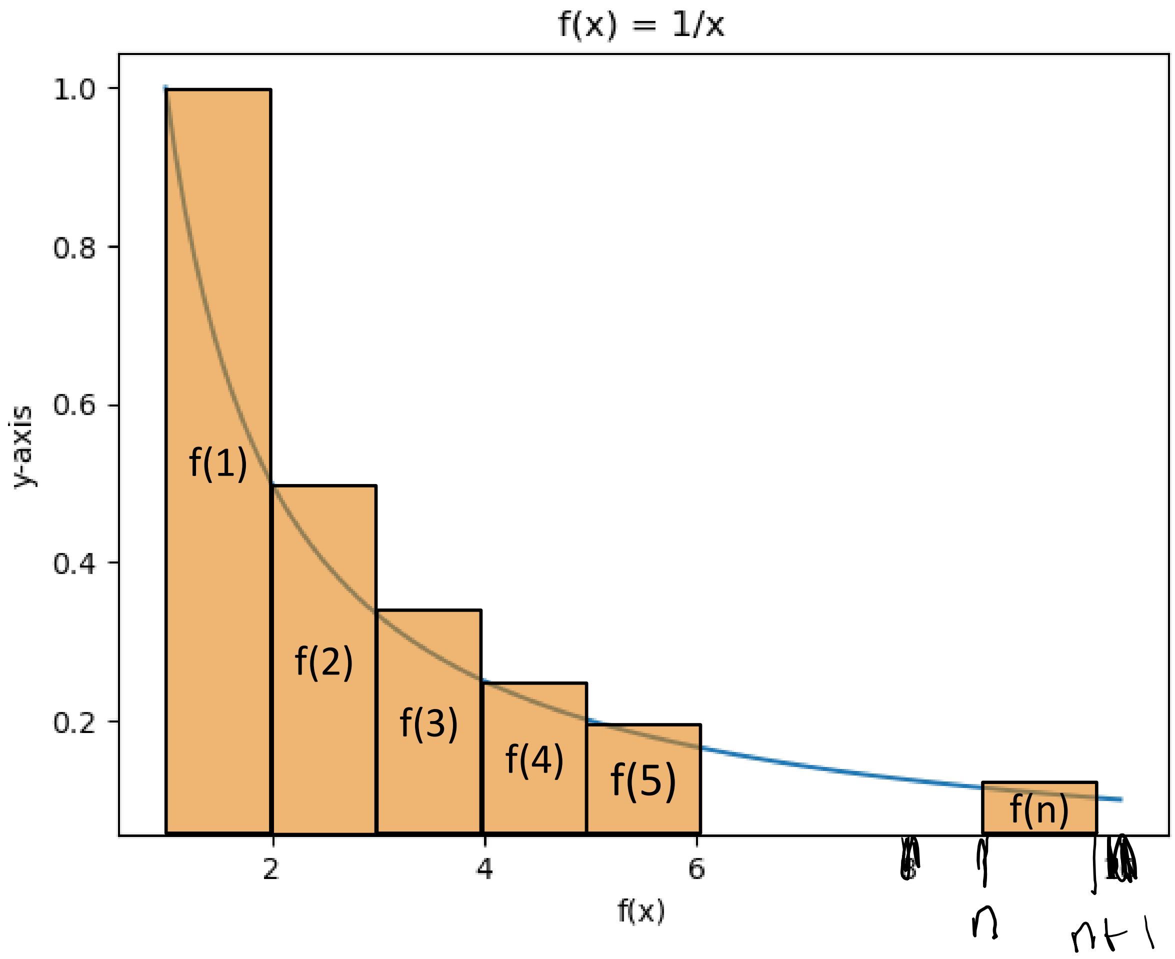


$$\underbrace{f(1)+f(2)+f(3)+\cdots+f(n)}$$

$=$ total area of the rectangles

$\geq$ area under the function starting from 1 (under the blue line):

$$= \int_1^{n+1} f(x)dx$$



$$\underbrace{f(1) + f(2) + f(3) + \cdots + f(n)}$$

$=$ total area of the rectangles

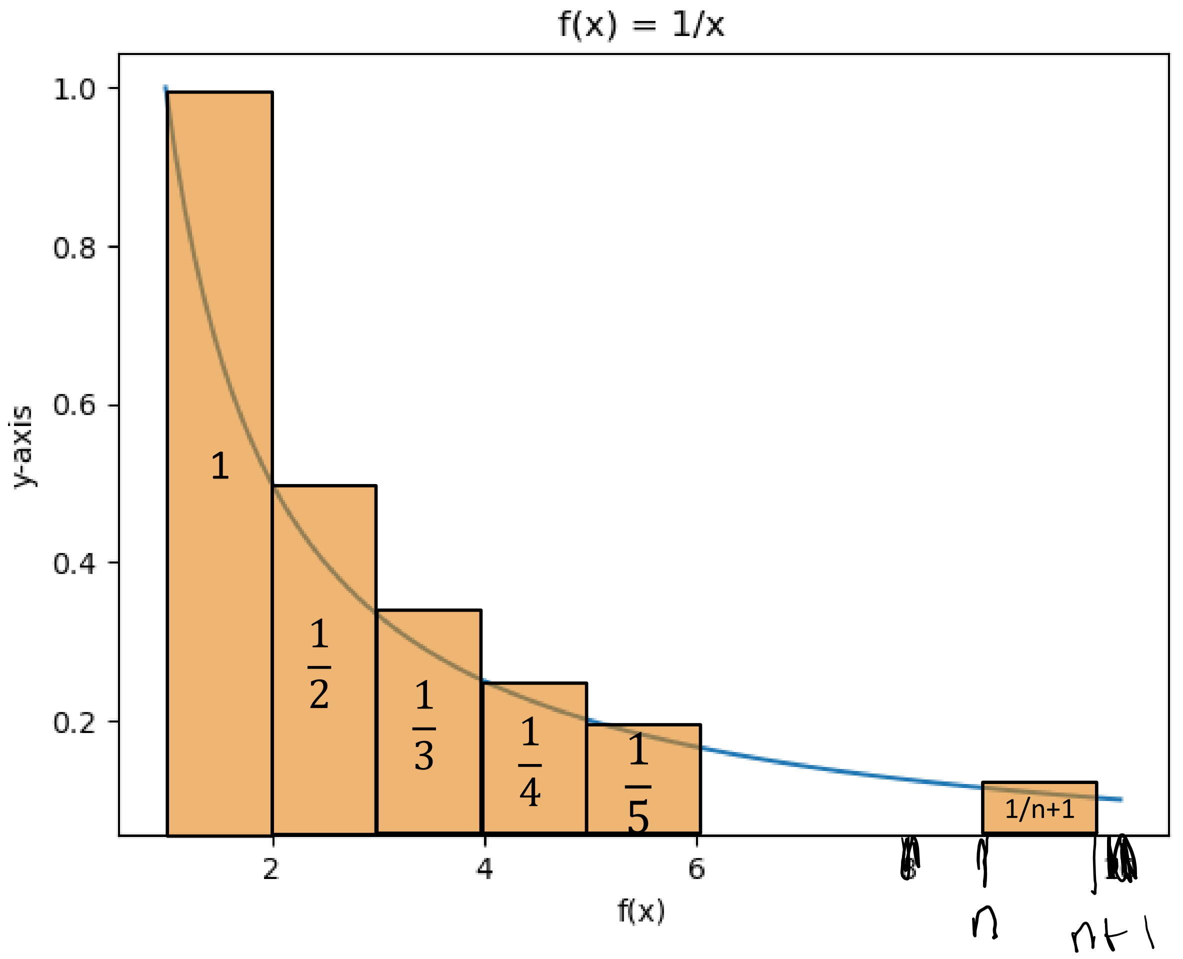
$\geq$ area under the function starting from 1 (under the blue line):

$$= \int_1^{n+1} f(x)dx$$

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

$$\geq \int_1^{n+1} \frac{dx}{x} = \ln(n+1) - \ln 1$$

$$\geq \ln n$$



Thus:

$$\ln n \leq Hn \leq \ln n + 1$$

And $\ln n$ is a decent approximation of H_n.

```
FindMax(S_1, …, S_n) :
  max = −∞;
  for i ∈ (1, …, n):
      if S_i > max:
          max = S_i ;    *
  return max ;
```

$X = \#$ of times $*$ is executed

What is $E(X)$?

For $i = 1, \ldots, n$:

$$X_i = \begin{cases} 1 \text{ if } * \text{ is executed in iteration } i \\ 0 \text{ otherwise} \end{cases}$$

$$E(X_i) = \Pr(X_i = 1) = \frac{1}{i}$$

$$E(X) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

$\mathrm{E}(X) = H_n \approx \ln n$ (natural log)

This was done using indicator random variables in a sort of obvious way – we either execute * or we don't.

Exercise 6.60

Professor M. S. has a grad student.

Every day this TA scrubs all of M. S.'s used beer mugs. In return the professor gives the TA one of his n different homebrewed IPA's, uniformly at random.

This TA decides to remain M. S.'s student at least until she tries all n of M. S.'s IPA's, then she will finish her thesis and graduate.

How many days until she has tried all n of M. S.'s homebrew IPAs?

$X =$ number of days until TA has tried all n IPAs.

What is $E(X)$?

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.
$E(X) = ?$

Define random variables $X_1, X_2, X_3, \ldots, X_n$, where X_i = # of days after new beer $i-1$ until new beer i.

$$X = X_1 + X_2 + \cdots + X_n$$

$$E(X) = E(X_1 + X_2 + \cdots + X_n)$$

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

by the linearity of expectation.

We must determine:

$$E(X_i), 1 \le i \le n$$

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.

$X_i = \#$ of days after new beer $i-1$ until new beer i.

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

What possible values can X_i take?

$$X_i \in \{1 \ldots \infty\}$$

What events do these values correspond to?

$$O = \text{received old beer}$$
$$N = \text{received new beer}$$

$X_i = 4$ is the event: $\{OOON\}$

Each of these events are independent.

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.

X_i = # of days after new beer $i - 1$ until new beer i.

This is the problem of (possibly infinite) independent trials until success.
What is $\Pr(\text{Success}) = \Pr(\text{receiving new beer i})$?

We have seen $i - 1$ beer so far. There are n beer total, so there are $n - i + 1$ beer that we have not seen.

$$\Pr(\text{Success}) = \frac{n - i + 1}{n}$$

Thus

$$E(X_i) = \frac{1}{\Pr(\text{Success})} = \frac{n}{n - i + 1}$$

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.

X_i = # of days after new beer $i-1$ until new beer i.

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

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

$$= \sum_{i=1}^{n} \frac{1}{\Pr(\text{Success})}$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

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

We can add these in reverse:

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.

X_i = # of days after new beer $i-1$ until new beer i.

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

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

Added in reverse:

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

$$= \sum_{i=1}^{n} \frac{n}{i}$$

$$= n \cdot \sum_{i=1}^{n} \frac{1}{i}$$

n different IPA's. Each day, one IPA is chosen uniformly at random.

X = number of days until TA has tried all n IPAs.

X_i = # of days after new beer $i-1$ until new beer i.

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

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

$$E(X) = n \cdot \sum_{i=1}^{n} \frac{1}{i}$$

$$E(X) = n \cdot H_n$$

$$E(X) \approx n \ln n$$

The last few take the longest time.

M.S. has n different IPA's. Each day, one IPA is chosen uniformly at random.

Student stays for m days, then graduates.

$X =$ number of new IPA's the TA tries

Solve this using indicator random variables.