CSE 312

Foundations of Computing II

20: Counting Distinct Elements

www.slido.com/2226110

my office hours today at 2pm in CSE ~~1~~ $\boxed{203}$

Conditional Expectation

Definition. Let X be a discrete random variable then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)$$

Note:

- Linearity of expectation still applies here

$$\mathbb{E}[aX + bY + c \mid A] = a\,\mathbb{E}[X \mid A] + b\,\mathbb{E}[Y \mid A] + c$$

Law of Total Expectation

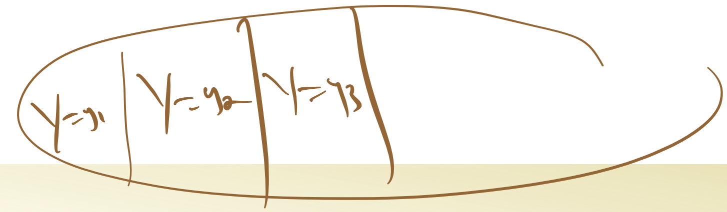
Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \dots, A_n$ partition the sample space. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)$$

Law of Total Expectation (random variable version). Let X be a random variable and Y be a discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$$

A_i's $= \{Y = y\}$



Law of total probability for continuous random variables.

Y discrete

$$P(A) = \sum_{y \in \Omega_Y} P(A \mid Y=y)\, P(Y=y)$$

Definition. Let A be an event and Y a continuous random variable. Then

$$P[A] = \int_{-\infty}^{\infty} P(A|Y = y) f_Y(y) \mathrm{d}y$$

Data mining – Stream Model

- In many data mining situations, data often not known ahead of time.
  - Examples: Google queries, Twitter or Facebook status updates, YouTube video views
- Think of the data as an <u>infinite stream</u>
- Input elements (e.g. Google queries) enter/arrive one at a time.
  - We cannot possibly store the stream.

Question: How do we make critical calculations about the data stream using a limited amount of memory?

x_N ... $x_3 x_2 x_1$

Stream Model – Problem Setup

Input: sequence (aka. "stream") of N elements $x_1, x_2, \ldots, x_N$ from a known universe U (e.g., 8-byte integers).

Goal: perform a computation on the input, in a single left to right pass, where:

- Elements processed in real time
- Can't store the full data $\Rightarrow$ use minimal amount of storage while maintaining working "summary"

What can we compute?

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

Some functions are easy:

- Min
- Max
- Sum
- Average

$x_1 \ldots$ x_N

32
12
14
7
4

Today: Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

Application

You are the content manager at YouTube, and you are trying to figure out the **distinct** view count for a video. How do we do that?

Note: A person can view their favorite videos several times, but they only count as 1 **distinct** view!

Other applications

- IP packet streams: How many distinct IP addresses or IP flows (source+destination IP, port, protocol)
  - Anomaly detection, traffic monitoring
- Search: How many distinct search queries on Google on a certain topic yesterday
- Web services: how many distinct users (cookies) searched/browsed a certain term/item
  - Advertising, marketing trends, etc.

Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

N = # of IDs in the stream = 11, m = # of distinct IDs in the stream = 5

Want to compute number of **distinct** IDs in the stream.

- *Naïve solution: As the data stream comes in, store all distinct IDs in a hash table.*
- *Space requirement:* $\Omega(m)$

YouTube Scenario: m is huge!

Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

N = # of IDs in the stream = 11, m = # of distinct IDs in the stream = 5

Want to compute number of **distinct** IDs in the stream.

How to do this <u>without</u> storing all the elements?

Detour – I.I.D. Uniforms

$$E\Big[\underbrace{\min(Y_1, Y_2, .. Y_n)}_{Y}\Big]$$

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

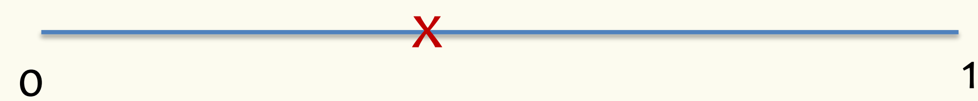
$m = 1$

$E(Y) = \frac{1}{2}$

Detour – I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

$m = 1$



$m = 2$

$E(\min(Y_1, Y_2)) = \frac{1}{3}$

Detour – I.I.D. Uniforms

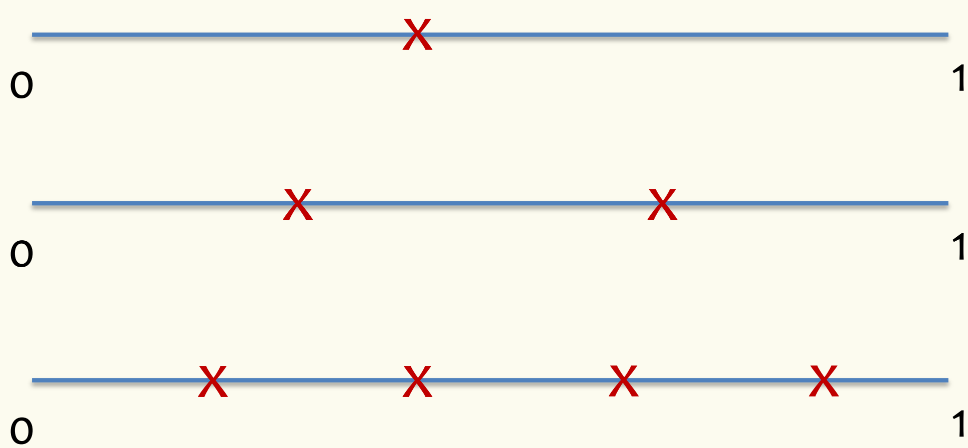
If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

"Evenly spread out"

$m = 1$

$m = 2$

$m = 4$



$E(\min(Y_1, \dots, Y_4)) = \frac{1}{5}$

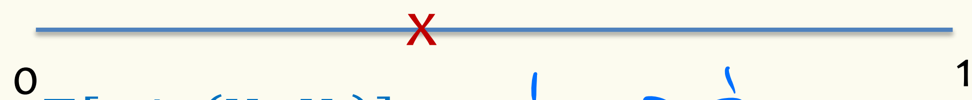
Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \text{Unif}(0,1)$ (iid) where do we expect the points to end up?

In general, $\mathbb{E}[\min(Y_1, \cdots, Y_m)] = \frac{1}{m+1}$

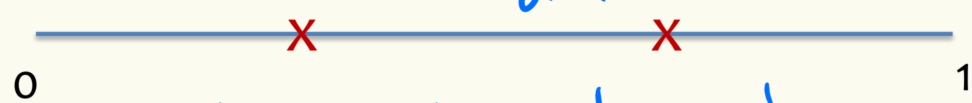
$\mathbb{E}[\min(Y_1)] = \frac{1}{1+1} = \frac{1}{2}$

$m = 1$



$\mathbb{E}[\min(Y_1, Y_2)] = \frac{1}{2+1} = \frac{1}{3}$

$m = 2$



$\mathbb{E}[\min(Y_1, \cdots, Y_4)] = \frac{1}{4+1} = \frac{1}{5}$

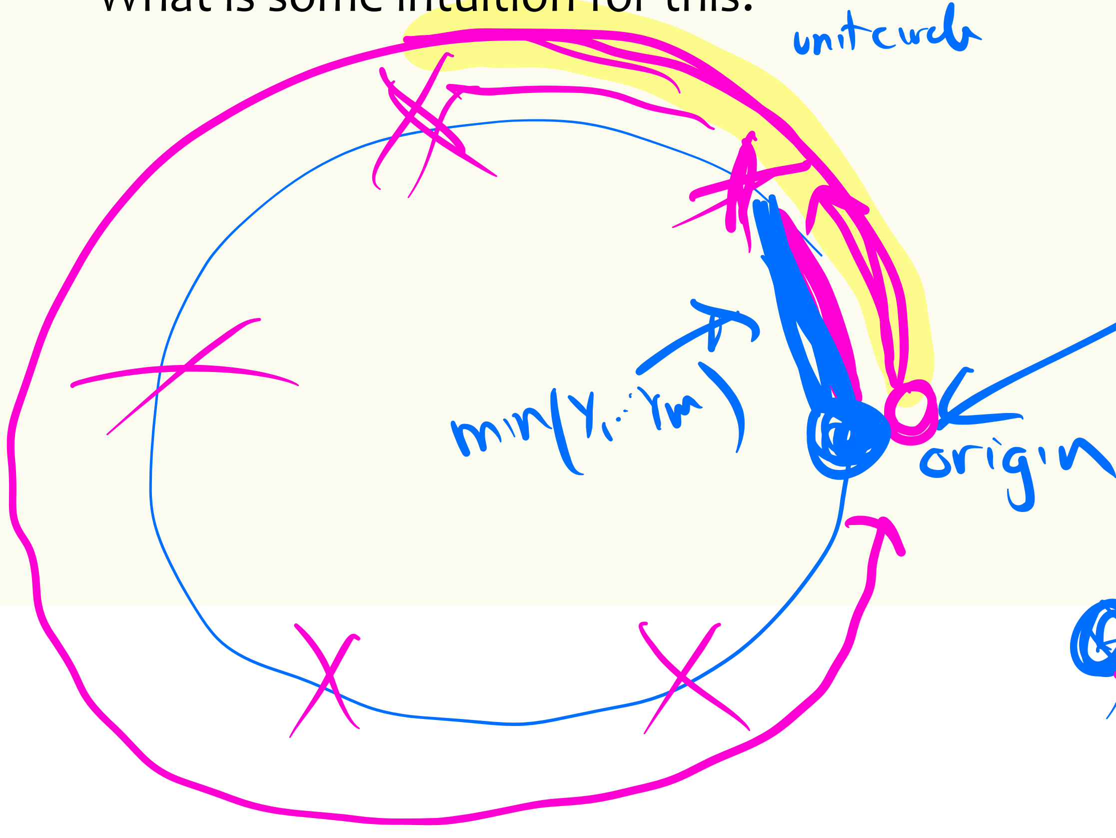
$m = 4$

Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (iid) where do we expect the points to end up?

In general, $\mathbb{E}[\min(Y_1, \cdots, Y_m)] = \frac{1}{m+1}$

What is some intuition for this?

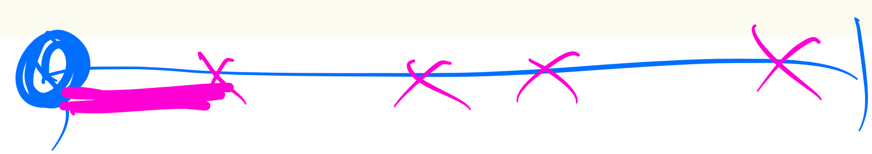


throw m pts unif at random along circle.

throw one more pt uniformly a

avg distance between pts

$\frac{1}{m+1} = E\left(\min(Y_1, \dots Y_m)\right)$



Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

e.g., what is $\mathbb{E}[\min\{Y_1, \cdots, Y_m\}]$?

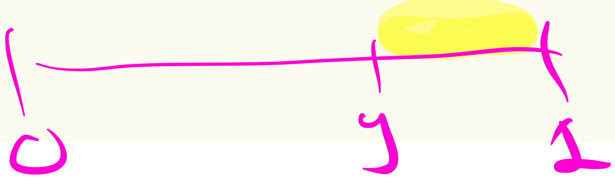
CDF: Observe that $\min\{Y_1, \cdots, Y_m\} \geq y$ if and only if $Y_1 \geq y, \ldots, Y_m \geq y$

$$P(\overset{Y}{\min\{Y_1, \cdots, Y_m\}} \geq y) = P(Y_1 \geq y, \ldots, Y_m \geq y)$$

$y \in [0,1]$

$$= P(Y_1 \geq y) \cdots P(Y_m \geq y) \quad \text{(Independence)}$$

$$= (1 - y)^m$$



$$\Rightarrow P(\min\{Y_1, \cdots, Y_m\} \leq y) = 1 - (1 - y)^m$$

$$F_Y(y) = P(Y \le y) = 1 - (1-y)^m$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = m(1-y)^{m-1}$$

$$E(Y) = \int_0^1 y f_Y(y)\, dy = \int_0^1 y m (1-y)^{m-1}\, dy$$

$$= \frac{1}{m+1}$$

Detour – Min of I.I.D. Uniforms

Useful fact. For any random variable Y taking non-negative values

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y)\mathrm{d}y$$

Proof

$$\mathbb{E}[Y] = \int_0^\infty x \cdot f_Y(x)\,\mathrm{d}x = \int_0^\infty \left(\int_0^x 1\,\mathrm{d}y\right) \cdot f_Y(x)\,\mathrm{d}x = \int_0^\infty \int_0^x f_Y(x)\,\mathrm{d}y\,\mathrm{d}x$$

$$= \int_0^\infty \int_y^\infty f_Y(x)\,\mathrm{d}x\,\mathrm{d}y = \int_0^\infty P(Y \geq y)\,\mathrm{d}y$$

Detour – Min of I.I.D. Uniforms

$Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.)

$Y = \min\{Y_1, \cdots, Y_m\}$

Useful fact. For any random variable Y taking non-negative values

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y)\mathrm{d}y$$

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y)\mathrm{d}y = \int_0^1 (1-y)^m \mathrm{d}y$$

$$= -\frac{1}{m+1}(1-y)^{m+1}\Big|_0^1 = 0 - \left(-\frac{1}{m+1}\right) = \boxed{\frac{1}{m+1}}$$

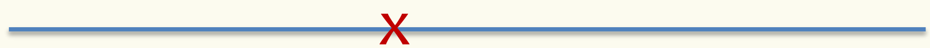
Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (iid) where do we expect the points to end up?

In general, $\mathbb{E}[\min(Y_1, \cdots, Y_m)] = \frac{1}{m+1}$

$$\mathbb{E}[\min(Y_1)] = \frac{1}{1+1} = \frac{1}{2}$$

$m = 1$



$$\mathbb{E}[\min(Y_1, Y_2)] = \frac{1}{2+1} = \frac{1}{3}$$

$m = 2$

$$\mathbb{E}[\min(Y_1, \cdots, Y_4)] = \frac{1}{4+1} = \frac{1}{5}$$

$m = 4$

Back to counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

N = # of IDs in the stream = 11, m = # of distinct IDs in the stream = 5

Want to compute number of **distinct** IDs in the stream.

How to do this <u>without</u> storing all the elements?

Distinct Elements – Hashing into $[0, 1]$

Hash function $h: U \rightarrow [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \mathrm{Unif}(0,1)$ and mutually independent

32, 12, 14, 32, 7, 12, 32, 7

h(32), h(12), h(14), h(32), h(7), h(12), h(32), h(7)

0.38 0.71 0.15 0.38 0.25 0.71

Distinct Elements – Hashing into $[0, 1]$

Hash function $h: U \to [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \mathrm{Unif}(0,1)$ and mutually independent

32, 12, 14, 32, 7, 12, 32, 7

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

h(32), h(12), h(14), h(32), h(7), h(12), h(32), h(7)

4 distinct elements

$\longrightarrow$ 4 i.i.d. RVs $\quad h(32), h(12), h(14), h(7) \sim \mathrm{Unif}(0,1)$

$\longrightarrow \mathbb{E}[\min\{h(32), h(12), h(14), h(7)\}] = \frac{1}{4+1} = \frac{1}{5}$

$\frac{1}{m+1}$

Distinct Elements – Hashing into $[0, 1]$

Hash function $h: U \to [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \text{Unif}(0,1)$ and mutually independent

$x_1, x_2, \dots, x_N$ contains m distinct elements

⬇

$h(x_1), h(x_2), \dots, h(x_N)$ contains m i.i.d. rvs $\sim \text{Unif}(0,1)$
and $N - m$ repeats

⬇

$$\mathbb{E}[\min\{h(x_1), \dots, h(x_N)\}] = \frac{1}{m+1}$$

A super duper clever idea!!!!

$$\mathbb{E}[\min\{h(x_1), \ldots, h(x_N)\}] = \frac{1}{m+1}$$

So $m = \frac{1}{\mathbb{E}[\min\{h(x_1),\ldots,h(x_N)\}]} - 1$

What if $\min\{h(x_1), \ldots, h(x_N)\}$ is $\approx \mathbb{E}[\min\{h(x_1), \ldots, h(x_N)\}]$?

The MinHash Algorithm – Idea

$$m = \frac{1}{\mathbb{E}[\min\{h(x_1), \dots, h(x_N)\}]} - 1$$

val

1. Compute $\mathrm{val} = \min\{h(x_1), \dots, h(x_N)\}$
2. Assume that $\mathrm{val} \approx \mathbb{E}[\min\{h(x_1), \dots, h(x_N)\}]$
3. Output as estimate for m: $\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)$

The MinHash Algorithm – Implementation

Algorithm **MinHash**$(x_1, x_2, \ldots, x_N)$

$\boxed{\text{val} \leftarrow \infty}$

for $i = 1$ **to** N **do**

$\quad \text{val} \leftarrow \min\{\text{val}, h(x_i)\}$

return $\boxed{\text{round}\left(\frac{1}{\text{val}} - 1\right)}$

estimate for m.

Memory cost = just remember val (with sufficient precision)

$\text{val} = \min(h(x_1), \ldots, h(x_N))$

MinHash Example

1. Compute $\mathrm{val} = \min\{h(x_1), \ldots, h(x_N)\}$
2. Assume that $\mathrm{val} \approx \mathbb{E}[\min\{h(x_1), \ldots, h(x_N)\}]$
3. Output $\boxed{\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)}$

Stream: 13, 25, 19, 25, 19, 19

Hashes: 0.51, 0.26, 0.79, 0.26, 0.79, 0.79

val = 0.26

$\mathrm{round}\left(\frac{1}{0.26} - 1\right)$

What does MinHash return?

Poll: **www.slido.com/2226110**

a. 1
b. 3
c. 5
d. No idea

MinHash Example II

Stream: 11, 34, 89, 11, 89, 23

Hashes: 0.5, 0.21, 0.94, 0.5, 0.94, 0.1

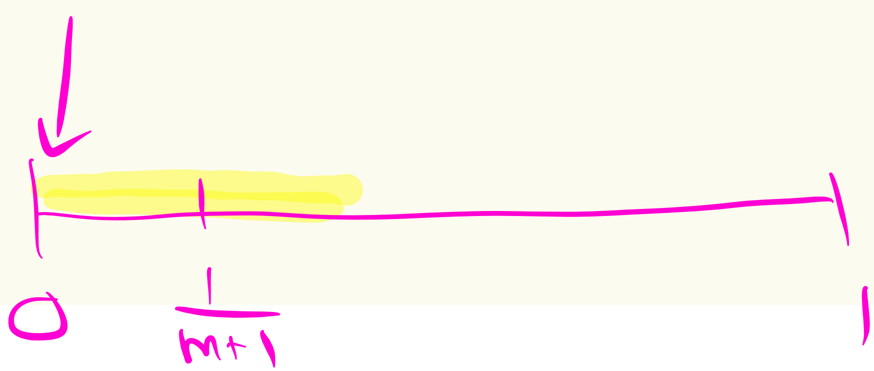
Output is $\frac{1}{0.1} - 1 = 9$

Clearly, not a very good answer!

Not unlikely: $P(h(x) < 0.1) = 0.1$

$\mathrm{Var}(val) \approx \frac{1}{(m+1)^2}$

$\min(h(x_1), \ldots h(x_m))$



st dev $\sigma(\text{val}) \approx \frac{1}{m+1}$

The MinHash Algorithm – Problem

Algorithm **MinHash**$(x_1, x_2, \ldots, x_N)$

$\text{val} \leftarrow \infty$

for $i = 1$ **to** N **do**

$\text{val} \leftarrow \min\{\text{val}, h(x_i)\}$

return $\text{round}\left(\frac{1}{\text{val}} - 1\right)$

But val is not $\mathbb{E}[\text{val}]$!

How far is val from $\mathbb{E}[\text{val}]$?

$$\text{Var}(\text{val}) \approx \frac{1}{(m+1)^2}$$

$\text{val} = \min\{h(x_1), \ldots, h(x_N)\}$

$$\mathbb{E}[\text{val}] = \frac{1}{m+1}$$

How can we reduce the variance?

Idea: Repetition to reduce variance!

Use k **independent** hash functions $h^1, h^2, \cdots h^k$

$$val^1 = \min(h^1(x_1), \ldots, h^1(x_N)) = \min(y_1^1, \ldots y_m^1)$$

$$val^2 = \min(h^2(x_1), \ldots h^2(x_N)) = \min(y_1^2, \ldots y_m^2)$$

$$\vdots$$

$$val^k =$$

$$\boxed{\overline{val} = \frac{1}{k}\sum_{i=1}^{k} val^i}$$

$$E(\overline{val}) = \frac{1}{k}\sum_{i=1}^{k} \underbrace{E(val^i)}_{\frac{1}{m+1}} = \frac{1}{m+1}$$

$$Var(\overline{val}) = \frac{1}{k^2} Var\left(\sum_{i=1}^{k} val^i\right) = \frac{1}{k^2}\cdot\sum_{i=1}^{k}\frac{1}{(m+1)^2}$$

$= \frac{1}{k^2} \frac{k}{(m+1)^2} = \frac{1}{k(m+1)^2}$

How can we reduce the variance?

Idea: Repetition to reduce variance!

Use k **independent** hash functions $h^1, h^2, \cdots h^k$

Algorithm **MinHash**$(x_1, x_2, \ldots, x_N)$

$\mathrm{val}_1, \ldots, \mathrm{val}_k \leftarrow \infty$

for $i = 1$ **to** N **do**

$\cancel{\mathrm{val}_1 \leftarrow \min\{\mathrm{val}_1, h^1(x_i)\}, \ldots, \mathrm{val}_k \leftarrow \min\{\mathrm{val}_k, h^k(x_i)\}}$

for $j = 1$ to k

$\mathrm{val}_j \leftarrow \min(\mathrm{val}_j, h^j(x_i))$

$\mathrm{val} \leftarrow \frac{1}{k} \sum_{i=1}^{k} \mathrm{val}_i$

return $\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)$

$$\mathrm{Var}(\mathrm{val}) = \frac{1}{k} \frac{1}{(m+1)^2}$$

$m = \frac{1}{E(\text{minhash})} - 1$

MinHash and Estimating # of Distinct Elements in Practice

- MinHash in practice:
  - One also stores the element that has the minimum hash value for each of the k hash functions
    - Then, just given separate MinHashes for sets A and B, can also estimate
      - what fraction of $A \cup B$ is in $A \cap B$; i.e., how similar A and B are