

Counting Distinct Elements in a Stream

Notes 29 · Full lecture · 75 minutes

Sariel Har-Peled

CS 574 · University of Illinois, Urbana-Champaign · Fall 2026

Two approaches to distinct counting

First: the minimum of ideal random hash values. Then: bounded samples using pairwise-independent randomness.

Part	Minutes
Model and target	5
Minimum estimator and its moments	20
Pairwise sampling at one scale	20
Amplifying a mixed output	10
Unknown scale and finite seeds	15
Guarantees and discussion	5

The object being counted

$$\sigma = (s_1, \dots, s_m), \quad s_t \in [n], \quad S = \{s_t : 1 \leq t \leq m\}.$$
$$u = |S| = F_0(\sigma).$$

Repeated occurrences contribute only once to S .

The approximation guarantee

Given $0 < \varepsilon, \delta < 1$, return $\hat{u}$ such that

$$\Pr[(1 - \varepsilon)u \leq \hat{u} \leq (1 + \varepsilon)u] \geq 1 - \delta.$$

One pass; insertion-only stream; space measured in words. Return 0 on an empty stream. For $n = 1$, count exactly with a nonempty flag.

Randomness must be attached to values

A hash function h assigns one random value to each key. Every occurrence of key i reuses $h(i)$.

New randomness at every occurrence would reflect the stream length m . The target here is the number u of distinct values.

An ideal random function

Assume $h : [n] \rightarrow [0, 1]$ assigns independent uniform real values to distinct keys. Store

$$Y = \min_{1 \leq t \leq m} h(s_t) = \min_{i \in S} h(i).$$

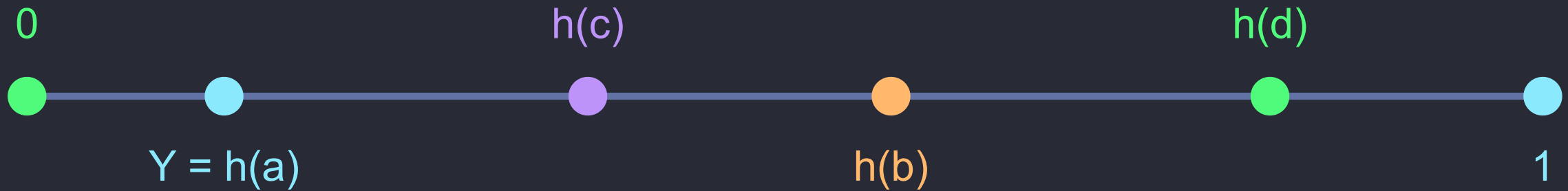
The repeated values disappear automatically under the minimum operation.

First order statistic

For $u \geq 1$, let

$$X_1, \dots, X_u \stackrel{\text{iid}}{\sim} \text{Unif}[0, 1], \quad Y = \min(X_1, \dots, X_u).$$

The minimum depends on distinct keys



Compute the survival probability

For $0 \leq t \leq 1$,

$$\begin{aligned}\Pr[Y > t] &= \Pr[X_1 > t, \dots, X_u > t] \\ &= \prod_{i=1}^u \Pr[X_i > t] = (1 - t)^u.\end{aligned}$$

Independence is essential to the product.

From survival to density

$$F_Y(t) = \Pr[Y \leq t] = 1 - (1 - t)^u.$$

Differentiate on $(0, 1)$:

$$f_Y(t) = u(1 - t)^{u-1}.$$

A density is not a point probability. For $0 \leq a \leq b \leq 1$, $\Pr[a \leq Y \leq b] = \int_a^b f_Y(t) dt$.

Integrate the tail to get the mean

Since $0 \leq Y \leq 1$,

$$Y = \int_0^1 \mathbf{1}[Y > t] dt.$$

Taking expectations,

$$\mathbb{E}Y = \int_0^1 \Pr[Y > t] dt = \int_0^1 (1 - t)^u dt = \frac{1}{u + 1}.$$

Compute the second moment

$$Y^2 = \int_0^1 2t \mathbf{1}[Y > t] dt.$$

$$\mathbb{E}Y^2 = 2 \int_0^1 t(1-t)^u dt$$

$$= 2 \int_0^1 (1-z)z^u dz$$

$$= 2 \left(\frac{1}{u+1} - \frac{1}{u+2} \right) = \frac{2}{(u+1)(u+2)}.$$

Subtract the squared mean

$$\begin{aligned}\text{Var}(Y) &= \frac{2}{(u+1)(u+2)} - \frac{1}{(u+1)^2} \\ &= \frac{u}{(u+1)^2(u+2)}.\end{aligned}$$

$$\boxed{\frac{\text{Var}(Y)}{(\mathbb{E}Y)^2} = \frac{u}{u+2} < 1.}$$

What are we actually estimating?

$$\mathbb{E}Y = \frac{1}{u+1}, \quad \mathbb{E}Y \neq u.$$

First estimate $\mu = 1/(u+1)$ accurately; then invert and subtract 1.

The variance-to-mean-square ratio is bounded independently of u .

Recall the averaging step

For independent copies $Y_1, \dots, Y_\alpha$, set

$$A = \frac{1}{\alpha} \sum_j Y_j.$$

Then

$$\mathbb{E}A = \mu, \quad \text{Var}(A) \leq \frac{\mu^2}{\alpha}.$$

Choose accuracy for the reciprocal

Set $\eta = \varepsilon/4$ and $\alpha = \lceil 64/\varepsilon^2 \rceil$.

$$\Pr[|A - \mu| > \eta\mu] \leq \frac{1}{\alpha\eta^2} \leq \frac{1}{4}.$$

Each block needs $O(\varepsilon^{-2})$ independent minimum sketches.

Recall the median step

Build $\beta = 2\lceil 12 \ln(1/\delta) \rceil + 1$ independent block averages. Let $Z = \text{median}(A_1, \dots, A_\beta)$.

If G counts good blocks, then $\mathbb{E}G \geq 3\beta/4$ and

$$\Pr[|Z - \mu| > \eta\mu] \leq \Pr[G < \beta/2] \leq e^{-\beta/24} \leq \delta.$$

All sketches fit in one pass

Choose independent ideal functions $h_{i,j}$. On item s_t , update

$$Y_{i,j} \leftarrow \min\{Y_{i,j}, h_{i,j}(s_t)\}.$$

At the end, average within blocks and take the median across blocks.

$$M = \alpha\beta = O\left(\varepsilon^{-2} \log \frac{2}{\delta}\right).$$

Convert the estimate back to u

On the success event, write

$$Z = \frac{1 + \theta}{u + 1}, \quad |\theta| \leq \varepsilon/4.$$

Output $\hat{u} = 1/Z - 1$. Then

$$\hat{u} - u = -\frac{(u + 1)\theta}{1 + \theta}.$$

Bound the reciprocal error

For $u \geq 1$ and $0 < \varepsilon < 1$,

$$\begin{aligned} \frac{|\hat{u} - u|}{u} &= \frac{u+1}{u} \frac{|\theta|}{|1+\theta|} \\ &\leq \frac{2|\theta|}{1-|\theta|} \leq \frac{2\varepsilon}{4-\varepsilon} < \varepsilon. \end{aligned}$$

The idealized theorem

Assuming perfectly random real-valued hashes and their stored values cost $O(1)$ words,

$$\Pr[|\hat{u} - u| > \varepsilon u] \leq \delta,$$

using $O(\varepsilon^{-2} \log(2/\delta))$ words in one pass.

The constant-space hash assumption is idealized. We next give an explicit finite-seed construction.

A sample of distinct values

For each $i \in [n]$, let $B_i \in \{0, 1\}$, with

$$\Pr[B_i = 1] = p.$$

Assume the bits are pairwise independent and reproducible from a small seed.

Retain the set $R = \{i \in S : B_i = 1\}$.

The update rule

On arrival of s_t , evaluate B_{s_t} . If it is 1, insert s_t into a set R ; duplicates do nothing.

One persistent sampling decision per key



The sample size is a sum of indicators

$$X = |R| = \sum_{i \in S} B_i.$$

By linearity,

$$\mathbb{E}X = \sum_{i \in S} \mathbb{E}B_i = pu.$$

The sum runs over distinct values, not stream positions.

Pairwise independence is enough for variance

For $i \neq j$,

$$\text{Cov}(B_i, B_j) = 0.$$

Therefore

$$\text{Var}(X) = \sum_{i \in S} \text{Var}(B_i) = up(1 - p) \leq up.$$

Rescale the sample count

For $p > 0$, let $\rho = X/p$. Then

$$\mathbb{E}\rho = u, \quad \text{Var}(\rho) = \frac{u(1-p)}{p}.$$

$$\Pr[|\rho - u| > \varepsilon u] \leq \frac{1-p}{\varepsilon^2 p u} \leq \frac{1}{\varepsilon^2 p u}.$$

Why introduce a scale?

Accurate sampling needs pu on the order of ε^{-2} . But u is the unknown target.

First solve a promised problem: given a scale N with $u > N/4$, return an estimate or a valid declaration that $u > 2N$.

A precise one-scale contract

Given $0 < \varepsilon < 1/4$ and $u > N/4$, output one of:

1. The declaration **large**: $u > 2N$.
2. A number $\rho \in [(1 - \varepsilon)u, (1 + \varepsilon)u]$.

The output is correct with probability at least $7/8$.

Both output types can be correct when $u > 2N$.

Set the probability and memory cap

In the ideal sampling calculation, choose

$$p = \min \left\{ 1, \frac{32}{N\epsilon^2} \right\}, \quad T = pN.$$

Keep distinct sampled keys until $|R| > 8T$.

The minimum with 1 is necessary at small scales.

The one-scale algorithm

1. Maintain the sampled set R .
2. If its size exceeds $8T$, output **large** and discard this copy's set.
3. If no overflow occurs, output $\rho = |R|/p$ at the end.

The maximum stored size is $\lfloor 8T \rfloor + 1 = O(\varepsilon^{-2})$.

Analyze the exact branch first

If $p = 1$, then $N \leq 32/\varepsilon^2$ and R contains every distinct key seen.

No overflow: return $|R| = u$ exactly.

Overflow: more than $8N$ distinct keys appeared, so $u > 8N > 2N$.

Both outputs are correct deterministically.

Analyze the sampling branch

Suppose $p = 32/(N\varepsilon^2) < 1$ and $u > N/4$. Then

$$pu > \frac{32}{N\varepsilon^2} \frac{N}{4} = \frac{8}{\varepsilon^2}.$$

$$\Pr[|X - pu| > \varepsilon pu] \leq \frac{\text{Var}(X)}{\varepsilon^2 p^2 u^2} \leq \frac{1}{\varepsilon^2 pu} < \frac{1}{8}.$$

Use one success event for both output types

Let X be the size of the final conceptual sample, even if storage stopped at overflow. Define

$$\mathcal{G} = \{|X - pu| \leq \varepsilon pu\}.$$

Then $\Pr[\mathcal{G}] \geq 7/8$.

On $\mathcal{G}$, prove correctness separately for a number and for **large**.

A numerical output is accurate

If there was no overflow, the stored set is the full conceptual sample. On $\mathcal{G}$,

$$(1 - \varepsilon)pu \leq X \leq (1 + \varepsilon)pu.$$

Divide by p :

$$(1 - \varepsilon)u \leq \rho \leq (1 + \varepsilon)u.$$

An overflow declaration is valid

Samples grow monotonically, so overflow implies $X > 8pN$. On $\mathcal{G}$, $X \leq (1 + \varepsilon)pu$. Thus

$$u \geq \frac{X}{(1 + \varepsilon)p} > \frac{8N}{1 + \varepsilon} > 2N.$$

This is an unconditional success proof. We do not condition the probability bound on observing overflow.

One-scale lemma proved

Under $u > N/4$, the capped sampler uses $O(\varepsilon^{-2})$ words and returns a correct output with probability at least $7/8$.

Small scale: exact counting until overflow.

Larger scale: pairwise sampling, rescaling, and Chebyshev.

Independent repetitions at one scale

Run r independent copies, each with its own seed and bounded sampled set. Each output is either a number or **large**.

The amplification rule must handle both types of output.

Aggregate the mixed outputs

1. If at least half the copies say **large**, return **large**.
2. Otherwise, return the median of the numerical outputs.

For the proof, we will ensure fewer than $r/4$ copies are incorrect.

Define a bad copy by its output

A copy is bad if it gives a false declaration or an inaccurate number.

$$B = \sum_{j=1}^r \mathbf{1}[\text{copy } j \text{ is bad}], \quad \Pr[\text{copy } j \text{ bad}] \leq 1/8.$$

The indicators are independent across copies.

Bound the bad-copy tail

Using Markov on 2^B ,

$$\Pr[B \geq r/4] \leq 2^{-r/4} \mathbb{E}[2^B] \leq \left(\frac{9/8}{2^{1/4}} \right)^r.$$

Since $\ln(9/8) - (\ln 2)/4 < -1/24$,

$$\Pr[B \geq r/4] \leq e^{-r/24}.$$

A majority declaration contains a correct witness

Suppose $B < r/4$ and at least $r/2$ copies say **large**. At least one of those declarations is correct.

Correct declaration $\implies u > 2N$.

So the aggregate declaration is correct.

The numerical median has a good majority

Otherwise, the number q of numerical outputs satisfies $q > r/2$. At most $B < r/4 < q/2$ of them are bad.

More than half of the numerical outputs lie in $[(1 - \varepsilon)u, (1 + \varepsilon)u]$. Their median lies there too.

Amplified one-scale lemma

For a failure budget $0 < \eta < 1$, choose

$$r = \lceil 24 \ln(2/\eta) \rceil.$$

Then the aggregate output is correct with probability at least $1 - \eta$, using

$$O\left(\varepsilon^{-2} \log \frac{2}{\eta}\right) \text{ words.}$$

Implement the bits with a finite seed

Choose a prime $n \leq Q < 2n$ for $n \geq 2$. Choose independent uniform $a, b \in \mathbb{F}_Q$ and define

$$h(i) = a(i - 1) + b \pmod{Q}.$$

Two field elements specify a pairwise-independent hash function.

Why the affine hash is pairwise independent

For distinct keys i, j and prescribed outputs x, y ,

$$a(i - 1) + b = x, \quad a(j - 1) + b = y \pmod{Q}.$$

Subtracting gives a unique a because $i - j \neq 0$ in the field; then b is unique.

$$\Pr[h(i) = x, h(j) = y] = Q^{-2}.$$

Round the sampling probability upward

$$p_0 = \min\{1, 32/(N\varepsilon^2)\}, \quad K = \lceil Qp_0 \rceil.$$
$$B_i = \mathbf{1}[h(i) < K], \quad p' = K/Q \geq p_0.$$

Use the actual probability p' in both $T = p'N$ and $\rho = |R|/p'$.

Rounding preserves concentration

If $p_0 < 1$ and $u > N/4$,

$$p'u \geq p_0u > 8/\varepsilon^2.$$

Therefore the same Chebyshev bound applies.

If $p' = 1$, the sample is exact. Its size is still capped, so the exact/overflow proof applies.

Rounding preserves the space bound

For $p_0 < 1$,

$$T = p'N \leq \frac{32}{\varepsilon^2} + \frac{N}{Q}.$$

At the dyadic scales used below, $N < 2n$ and $Q \geq n$, so

$$T < \frac{32}{\varepsilon^2} + 2.$$

The sample cap is $O(\varepsilon^{-2})$; the seed and threshold use $O(1)$ words.

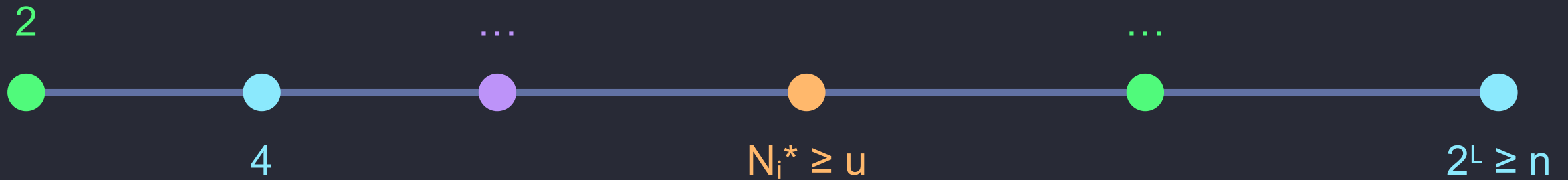
Cover the possible scales

For $n \geq 2$, define

$$L = \lceil \log_2 n \rceil, \quad N_i = 2^i \quad (1 \leq i \leq L).$$

Run an amplified sampler at every scale in parallel.

The first scale at least u is a witness, not an input



Share the failure budget across levels

Give each level failure budget $\eta = \delta/L$. Its space is

$$O\left(\varepsilon^{-2} \log \frac{2L}{\delta}\right).$$

Across all L levels:

$$O\left(\varepsilon^{-2} L \log \frac{2L}{\delta}\right) \text{ words.}$$

Select the first numerical output

After the stream ends, inspect levels in increasing order. Skip **large** declarations and return the first number.

We must prove that a numerical output exists before leaving the range where the one-scale promise holds.

Choose a witness level for the proof

For $u \geq 1$, let

$$i_* = \min\{i : N_i \geq u\}.$$

This exists since $N_L \geq n \geq u$.

$$u \leq N_{i_*} \leq 2u.$$

Every earlier level satisfies its promise

For every $j \leq i_*$,

$$N_j \leq N_{i_*} \leq 2u, \quad N_j/4 \leq u/2 < u.$$

Thus the one-scale guarantee applies to every level we need.

Levels beyond i_* need not satisfy the promise, and their outputs are irrelevant.

Union bound only over valid levels

Each level through i_* is wrong with probability at most δ/L .

$$\Pr[\text{some level } j \leq i_* \text{ is wrong}] \leq i_* \delta / L \leq \delta.$$

Cross-level independence is unnecessary for this union bound. Independence is required among the repetitions within each level.

The witness forces a numerical output

On the event that all levels through i_* are correct, level i_* cannot return **large**:

$$u \leq N_{i_*} \implies u > 2N_{i_*} \text{ is false.}$$

It must return an accurate number. Any earlier number is also accurate.

Complete the scale-selection proof

The first numerical output occurs at some $j \leq i_*$. On the simultaneous success event,

$$(1 - \varepsilon)u \leq \hat{u} \leq (1 + \varepsilon)u.$$

Hence

$$\Pr[|\hat{u} - u| > \varepsilon u] \leq \delta.$$

Distinct-count theorem with finite seeds

For $n \geq 2$, $0 < \varepsilon < 1/4$, and $0 < \delta < 1$, a one-pass algorithm uses

$$O\left(\varepsilon^{-2} \log n \log \frac{2^{\lceil \log_2 n \rceil}}{\delta}\right) \text{ words}$$

and returns a $(1 \pm \varepsilon)$ estimate of F_0 with probability at least $1 - \delta$.

Use exact handling for the empty stream and for $n = 1$. For larger requested ε , run with $\min\{\varepsilon, 1/8\}$.

The high-probability form in the notes

Fix any constant $a > 0$ and take $\delta = n^{-a}$. Then

$$L = O(\log n), \quad \log(2L/\delta) = O(\log n).$$

$$O(\varepsilon^{-2} \log^2 n) \text{ words,} \quad \Pr[\text{success}] \geq 1 - n^{-a}.$$

Keep the two randomness models distinct

Method	Randomness	Space
Minimum + average + median	Ideal fully random real hashes	$O(\varepsilon^{-2} \log(2/\delta))$ ideal words
Capped sampling at dyadic scales	Finite pairwise hash seeds	$O(\varepsilon^{-2} L \log(2L/\delta))$ words

Both ignore repeated keys consistently; the second has an explicit small seed.

Reconstruct the argument

1. Why does a minimum estimate $1/(u + 1)$?
2. Why is pairwise independence enough for the sampling bound?
3. Why does amplification use fewer than a quarter bad copies?
4. Which scale ensures that the first numerical answer is trustworthy?

Source and references

Based on Notes 29, *Approximating the Number of Distinct Elements in a Stream*.

Prerequisites: median estimation from Notes 28, Chebyshev's inequality, Chernoff amplification, and polynomial hashing. Background cited in the notes: (Motwani & Raghavan, 1995).

Motwani, R., & Raghavan, P. (1995). *Randomized algorithms*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511814075>