Distinct Elements

Kent Quanrud

September 2, 2020

This note is in pretty poor shape as I did not finish it before lecture, like I normally would. Still, I wanted to try to get it up in case anyone wanted to follow up on details. I will continue to edit this in the coming days.

1 Unique visitors

Imagine you made a website. You might like to know how many people have visited your website. You set up simple counter on your server to keep track of how many HTTP requests you have served. You are pleased to discover that this counter increases steadily - apparently there are many visits to your website. Upon investigation, however, you may find out that almost all of these visits are from the same bot, canvassing your web page for who-only-knows-what reason. In fact, what you are really interested in is not the number of visits of your webpage, but the number of *distinct visitors* to your website. Counting the number of distinct visitors is probably the most popular feature of *Google Analytics* and other similar software to help analyze websites.

How can we count the number of distinct visitors? Suppose each visitor has a unique identifier, such as an IP address[1]. We can store the set of different visitors in a dictionary, maybe even (or probably) using the linear probing hash table previously discussed. Even the best dictionaries, however, will ultimately require space proportional to the number of keys. For your website, this could be the total number of IP addresses in the world, and only so many IP addresses can fit in RAM. In fact, any *exact* implementation will always require a lot of space (see Exercise 12.1.)

Google Analytics has a help page where they explain how their unique visitors count works (`https://support.google.com/analytics/answer/2992042`). There is an interesting paragraph where they note that in 2017 they switched to a new algorithm to "more efficiently count users with high accuracy and low error rate (typically less than 2%)". From this we can extract some interesting observations. First, they are not reporting an exact count, but rather one with a "low error rate". That is, they are *approximating*. Second, they do not guarantee any particular rate of error. It is "typically" less than 2%, but apparently the error rate can vary and is not deterministically less than 2%. Lastly, they did something in early 2017 to improve their algorithm, and "more efficiently count users". This suggests that distinct elements is an active problem where having a better algorithm still matters[2]. A little more searching shows a 2014 publication out of Google about distinct elements [5], that explains some low-level enhancements for a well-known algorithm for the distinct elements problem called `HyperLogLog` [3].

[1] In reality, this may be a combination of IP address and cookies, partly because the same IP address can serve many people from the same network.

[2] Of course, there may be other details of the real-world problem, beyond the clean mathematical version we will discuss, that Google may have improved.

There are many natural applications for distinct elements besides internet-scale streaming. In databases, quick estimates of the number of distinct elements are used to optimize complex queries [1]. In general, having a crude estimate of the number of distinct elements can be useful in deciding what kind of data structure or algorithmic strategy to pursue when processing these elements. If the count is very small, then maybe an asymptotically poor approach with very good constants is actually faster. If the count is very large, then asymptotics kick in and one should go for the asymptotically optimal algorithms, even if the implementation is clumsy and the constants are bigger.

We briefly review the streaming model. We have elements coming in one at a time from a stream, from the set of integers $[n] = \{1, \ldots, n\}$. Let m denote the total number of elements in the stream. Items may be repeat. Our goal is to compute the *distinct* number of items in the stream. We will generally use k to denote the distinct number of items.

We mention in passing that the number of distinct elements can be interpretted as the **L_0-norm** of the frequency vector. Recall that the frequency vector is the vector with one coordinate per item counting the frequency of that item. Other norms, such as L_2 and L_p, and other quantities such as *entropy*, are helpful for understanding the "shape" of the data. We refer the reader to [2] for more on L_p-frequency estimation. Algorithms for these problems have found further use as extremely efficient data structures inside fast approximation algorithms, particularly for accelerating primitives from linear algebra.

2 Where to start

Where do we start in designing an efficient approximation algorithm for the distinct elements problem? The reader might guess based on previous discussions that hashing will be useful. Indeed, we will be using hashing. Let us take as a starting point the `count-min` data structure for the heavy hitters problem. One thing to point out about `count-min` is that some internal state changes with every element that is processed. On a given item e, `count-min` hashes e several times and increase a number of counters. Consider now the distinct elements problem. In this setting, the same item e can reappear again and again and again. The number of distinct elements does not change. Meanwhile, any algorithm similiar to `count-min` is always updating its state. While this mistmatch is not a formal proof that a `count-min` type of approach cannot work, the author suggests that it is a bad sign. One would prefer an approach that is totally impervious to repeating elements.

As a first hint towards a new approach, suppose the reader was given access to ideal hash functions $h : [n] \to [0, 1]$. That is, for each item e, we independently sample a unique continuous $h(e) \in [0, 1]$. (Later, after developing the main algorithmic ideas, we will return to this assumption on h and replace it with a more realistic alternative.) Does an ideal and continuous function such as h inspire any ideas?

As a second hint, suppose the reader was presented with the following statistical facts. How might the following lemma be employed in the service of estimating the number of distinct elements?

Lemma 1. *Let $Y_1, \ldots, Y_k \in [0, 1]$ be independent and distributed uniformly at random. Let $X = \min\{Y_1, \ldots, Y_k\}$.*

1. $\mathrm{E}[X] = \dfrac{1}{k+1}$.
2. $\mathrm{E}\left[X^2\right] = \dfrac{2}{(k+1)(k+2)}$.

3 An interlude on continuous random variables

We stated Lemma 1 before properly introducing continuous random variables. As with discrete random variables, we assume the reader has had some familiarity with continuous random variables. Here we briefly review the basic notions, which should confirm the reader's common sense.[3]

Consider a uniformly random variable $X \in [0,1]$. That is, X has "equal probability" of being any particular value in $[0,1]$. Now, the "probabilities" associated with X are a little less straight forward then with finite and discrete random variables, but the reader will likely find it to still be totally natural.

For any fixed $\alpha \in [0,1]$, say $\alpha = .51691$, the probability that X is *exactly* α is 0. The fact that a continuous random variable has "zero probability" at every point may seem odd but it is not a paradox. It is completely analogous to a line segment having no area in two dimensions, or a plane having no volume in three dimensions. For a continuous random variable such as X, the probability at a point is the wrong question to ask. Instead, let consider any two fixed values $a, b \in [0,1]$ with $0 \leq a \leq b \leq 1$. Then we have

$$\mathrm{P}[a \leq X \leq b] = b - a.$$

This is totally intuitive. We note that from the formal perspective of *measure theory*, understanding the probability of X lying in any particular (open or closed) interval suffices to define X. (See, e.g, [4].)

Any continuous random variable $X \in \mathbb{R}$ we consider will implicitly be equipped with a **density function** $f : \mathbb{R} \to [0,1]$. Then the probability of X lying in an interval $[a,b]$ is given by

$$\mathrm{P}[a \leq X \leq b] = \int_a^b f(t)\, dt.$$

We are implicitly assuming that f is integrable. The expected value of a continuous random variable $X \in \mathbb{R}$ with density function $f(t)$ is given by the following.

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} t f(t)\, dt.$$

This is the continuous analogue of the definition of expected value for discrete random variables, with the (discrete) sum replaced by a (continuous) integral.[4]

We note that Markov's inequality applies to continuous random variables as well. The argument is the same.

Lemma 2. *Let $X \geq 0$ be a nonnegative random variable (discrete or continuous). Then for any $\alpha \geq 1$,*

$$\mathrm{P}[X \geq \alpha\, \mathrm{E}[X]] \leq \frac{1}{\alpha}.$$

[3] The basic definitions and axioms of probability theory are natural and confirmed by every day experience. The challenge is in *sticking to these simple rules* when there are many moving parts and the conclusions are nonintuitive.

[4] In measure theory, discrete and continuous random variables are unified under the notion of *measurable* random variables. For finite variables, the summation over its values is interpreted as an integral over an appropriate *discrete* topology. See [4].

4 Back to distinct elements

Let us return to our discussion on developing an (idealized) algorithm for counting the distinct elements. The reader was presented with the following statistical facts about the minimum of independent uniform continuous random variable. How can one employ the following statistical facts algorithmically?

Lemma 1. *Let $Y_1, \ldots, Y_k \in [0,1]$ be independent and distributed uniformly at random. Let $X = \min\{Y_1, \ldots, Y_k\}$.*

1. $\mathrm{E}[X] = \dfrac{1}{k+1}$.
2. $\mathrm{E}\left[X^2\right] = \dfrac{2}{(k+1)(k+2)}$.

One can use the idealized, continuous hash function $h : [m] \to [0,1]$ to assign to each distinct element an indpendent and uniformly sampled value in $[0,1]$. Consider the minimum hash value seen in the stream. The set of hash values seen, over k distinct elements, is precisely a set of k independent $[0,1]$-random variables. We start with three basic observations.

1. The minimum hash does not change in the face of duplicate elements. It is strictly a function of the set of distinct elements that have appeared in the stream.
2. We only have to keep track of one number, which is space friendly.
3. By part 1 of Lemma 1, there is an explicit connection between the number of distinct elements.

Consider the third observation. If there are k distinct elements, which generate k uniformly random numbers $Y_1, \ldots, Y_k \in [0,1]$, then the minimum, $X = \min\{Y_1, \ldots, Y_k\}$, has expected value

$$\mathrm{E}[X] = \frac{1}{k+1},$$

per Lemma 1. This *suggests* using $(1/X) - 1$ as an estimator for k. However, we have not shown that $\mathrm{E}[(1/X) - 1]$ equals k, nor are we going to. Instead, our analysis is based on showing that X is close enough to $1/(k+1)$ to guarantee that $(1/X) - 1$ is closed to k.

Lemma 3. *Let $\epsilon > 0$ be sufficiently small. If $\dfrac{1-\epsilon}{k+1} \leq X \leq \dfrac{1+\epsilon}{k+1}$, then*

$$(1-3\epsilon)k \leq \frac{1}{X} - 1 \leq (1+3\epsilon)k.$$

Proof. We have the following equivalent inequalities.

$$\begin{aligned} \frac{1-\epsilon}{k+1} \leq X \leq \frac{1+\epsilon}{k+1}. \\ \frac{k+1}{1+\epsilon} \leq \frac{1}{X} \leq \frac{k+1}{1-\epsilon} \\ \frac{k-\epsilon}{1+\epsilon} \leq \frac{1}{X} - 1 \leq \frac{k+\epsilon}{1-\epsilon}. \end{aligned}$$

In the last equation, we observe that the LHS is $\geq (1-3\epsilon)k$ and the RHS is $\leq (1+3\epsilon)k$ for $k \geq 1$. ∎

Recall that $\mathrm{E}[X] = 1/(k+1)$. If one could show that

$$|X - \mathrm{E}[X]| \leq \frac{\epsilon}{k+1}$$

with high probability, then $1/X - 1$ would give an accurate estimate from k with high probability. Unfortunately, $\epsilon/(k+1)$ is a very small margin of error. For example, suppose we tried Markov's inequality. Markov's inequality implies that

$$\mathrm{P}\left[|X - \mathrm{E}[X]| \geq \frac{\epsilon}{k+1}\right] \leq \left(\frac{k+1}{\epsilon}\right) \mathrm{E}[|X - \mathrm{E}[X]|].$$

The RHS is challenging for two reasons. First, there is a potentially large $(k+1)/\epsilon$ factor that means $\mathrm{E}[|X - \mathrm{E}[X]|]$ needs to be much smaller then $\epsilon/(k+1)$ to obtain interesting probabilities. The second issue is that the absolute value is hard to analyze. Consider instead the following attempt that at least addresses the second issue by squaring both sides. We have

$$\mathrm{P}\left[|X - \mathrm{E}[X]| \geq \frac{\epsilon}{k+1}\right] \overset{\text{(a)}}{=} \mathrm{P}\left[(X - \mathrm{E}[X])^2 \geq \frac{\epsilon^2}{(k+1)^2}\right] \overset{\text{(b)}}{\leq} \frac{\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]}{\epsilon^2/(k+1)^2}.$$

We square both sides in step (a) because analyzing $(X - \mathrm{E}[X])^2$ is easier than analyzing the absolute value. (b) is by Markovs inequality. It remains to analyze the numerator. We have

$$\mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \overset{\text{(c)}}{=} \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X\,\mathrm{E}[X]] + \mathrm{E}\left[\mathrm{E}[X]^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \tag{1}$$

where (c) applies linearity of expectation. By our given statistics on the minimum of k uniform random variables, we have

$$\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{2}{(k+1)(k+2)} - \frac{1}{(k+1)^2} = \frac{k}{(k+1)^2(k+2)}.$$

Plugging back in, we have

$$\mathrm{P}\left[|X - \mathrm{E}[X]| \geq \frac{\epsilon}{k+1}\right] \leq \frac{\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]}{\epsilon^2/(k+1)^2} = \frac{k}{\epsilon^2(k+2)}.$$

Alas, $\dfrac{k}{\epsilon^2(k+2)}$ is bigger than 1 for any interesting ϵ and k, so the bound we obtained is not very useful.

5 Variance and Chebyshev's inequality

While our attempt above may not have totally worked out, it used some interesting tools that are worth pointing out.

The **variance** of a random variable X, sometimes denoted $\mathrm{Var}[X]$, is the quantity

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

We always have the identity,

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2,$$

which appears in (1).

Lemma 4 (Chebyshev's inequality)**.** *Let $X \in \mathbb{R}$ be a random variable. For any $\alpha > 0$,*

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

In the above analysis, we implicitly proved and then applied Chebyshev's inequality for $\alpha = \epsilon/(k+1)$. (The proof, to recap, is to square both sides and then apply Markov's inequality.) Chebyhsev's inequality, combined with our analysis of the variance of X, gave us

$$\mathrm{P}\left[|X - \mathrm{E}[X]| \geq \frac{\epsilon}{k+1}\right] \leq \left(\frac{k+1}{\epsilon}\right)^2 \mathrm{Var}[X] = \frac{k}{\epsilon^2(k+2)}.$$

Our attempt to analyze $X = \min\{Y_1, \ldots, Y_k\}$ above fell short because $\mathrm{Var}[X]$ was too large.

6 Amplification

How do we reduce the variance of a randomized experiment? This question extends far beyond distinct elements. In our own lives, every day is a randomized experiment. Some days are encouraging, and other days are more discouraging. Certainly we don't want to judge our life on the most recent day, or we would drive ourselves nuts! In this case, we might remind ourselves to take a step back, and take a *longer view over a greater time period* by taking the *average* of many days, to come to a more reliable assessment.

Averaging is precisely the technique that we will use to reduce the variance. The following lemma states that *the variance of the average of n independent variables is $(1/n)$th their average variance*. The key point is that *averaging reduces variance*. If we want to reduce the variance of an experiment by a factor of 10, simply repeat the experiment (independently) 10 times, and take the average!

Lemma 5. *Let $X_1, \ldots, X_\ell$ be ℓ independent and identically distributed random variables. Let*

$$\overline{X} = \frac{1}{\ell}\sum_{i=1}^{\ell} X_i$$

be the average of the X_i's. Let

$$\overline{V} = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathrm{Var}[X_i]$$

be the average variance over the X_i's. Then

$$\mathrm{Var}[\overline{X}] \leq \frac{1}{\ell}\overline{V}.$$

Proof. We have

$$\begin{aligned}\mathrm{Var}[\overline{X}] &= \frac{1}{\ell^2}\mathrm{E}[(X_1 + \cdots + X_\ell - \mathrm{E}[X_1] - \cdots - \mathrm{E}[X_\ell])] \\ &\overset{\text{(a)}}{=} \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])].\end{aligned}$$

(a) applies linearity of expectation. Consider each term in the sum for fixed i and j. If $i = j$, then

$$\mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])] = \mathrm{E}\left[(X_i - \mathrm{E}[X_i])^2\right] = \mathrm{Var}[X_i]$$

If $i \neq j$, then

$$\mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[(X_j)])] \overset{\text{(b)}}{=} \mathrm{E}[X_i - \mathrm{E}[X_i]]\,\mathrm{E}[X_j - \mathrm{E}[X_j]] = 0,$$

where (b) is by independence of X_i and X_j. Thus

$$\mathrm{Var}[\overline{X}] = \frac{1}{\ell^2}\sum_{i=1}^{\ell}\mathrm{Var}[X_i] = \frac{1}{\ell}\overline{V},$$

as desired. ■

7 Variance reduction for distinct elements

Let us return to the distinct elements problem. Recall that we have a random variable X with expected value $1/(k+1)$, where k is the number of distinct elements. X also has variance $2k/(k+1)^2(k+2)$. Chebyshev's inequality gives a bound of the form

$$\mathrm{P}\left[|X - \mathrm{E}[X]| \leq \frac{\epsilon}{k+1}\right] \leq \frac{(k+1)^2}{\epsilon^2}\,\mathrm{Var}[X] = \frac{k}{\epsilon^2(k+2)},$$

which is isn't very interesting since it is > 1 for any interesting value of ϵ. Now, however, we are equipped with a tool to reduce the variance.

Suppose we run ℓ independent copies of our experiment, producing ℓ independent random variables $X_1, \ldots, X_\ell$ each with expected value $1/(k+1)$ and variance $2k/(k+1)^2(k+2)$. Consider their *average*,

$$\overline{X} = \frac{1}{\ell}(X_1 + \cdots + X_\ell).$$

We have $\mathrm{E}[\overline{X}] = 1/(k+1)$ of course. The real point is that the variance decreases at a linear rate:

$$\mathrm{Var}[\overline{X}] = \frac{1}{\ell}\cdot\frac{2k}{(k+1)^2(k+2)}.$$

In particular, when we repeat the analysis as above except w/r/t the average $\overline{X}$, we get

$$\mathrm{P}\left[\left|\overline{X} - \frac{1}{k+1}\right| \geq \frac{\epsilon}{k+1}\right] \leq \frac{1}{\ell}\cdot\frac{k}{\epsilon^2(k+2)}$$

The key is that we have an additional parameter of ℓ to decrease the RHS. When we set $\ell \geq 1/\epsilon^2$, for example, we start to obtain interesting upper bounds that are less than 1. Suppose we want our probability of error to be δ, where $\delta \in (0,1)$. Setting $\ell = 1/\delta\epsilon^2$, we have

$$\frac{1-\epsilon}{k+1} \leq \overline{X} \leq \frac{1+\epsilon}{k+1} \text{ with error probability } \leq \delta.$$

In turn, we have

$$(1-3\epsilon)k \leq \frac{1}{\overline{X}} - 1 \leq (1+3\epsilon)k$$

with probability of error $\leq \delta$.

The new (idealized) algorithm is as follows. Suppose we want $(1 \pm \epsilon)$-multiplicative error with probability of error $\leq \delta$. We make $\ell = O\big(1/\epsilon^2\delta\big)$ ideal hash functions $h_1, \ldots, h_\ell : [m] \to [0,1]$. For each i, we let X_i be the minimum hash value produced by h_i over all the elements in the stream. We return

$$\frac{1}{\overline{X}} - 1 \text{ where } \overline{X} = \frac{1}{\ell}\sum_{i=1}^{\ell} X_i$$

is the average of the X_i's seen so far.

Thus, ignoring the assumption that we have access to such hash functions, we need space proportional to $1/\epsilon^2\delta$ to obtain an $(1 \pm \epsilon)$-approximation to the number of distinct elements with probability of error $\leq \delta$. This is far better than where we started. Suppose, however, that we wanted extremely small probability of error. Then the $1/\delta$ term starts to become expensive. Here, we will take inspiration from the following coin tossing experiment.

8 An interlude on coin tosses

If, out of 100 coin tosses, you were told that 50 of them were heads, would you be surprised? Actually, you should be a little surprised. The odds of getting exactly 50 heads is about 8%. But if you were told that the number was in the range, say, 45 to 55, you probably wouldn't think much of it.

If you were told that all 100 coin tosses came up heads, you wouldn't believe it. The odds of that, we know, is $1/2^{100}$. If you bet money and lost on this event, you would be outraged (and, at even odds, certainly broke for the rest of eternity).

Suppose you were told that at least 75 coin tosses came up heads. Should you be surprised? On one hand, this is 50% larger than the average. On the other hand, the claim is not that there was exactly 75 heads, but at least 75 heads. There could be 75, 76, 77, etc., up to 100. Even though the event of getting any one of these counts should be low, being far from average, there are also 26 of these events. Do they add up to very much?

Let $X_1, X_2 \ldots, X_n \in \{0,1\}$ be independent random variables where for each i, $\mathrm{P}[X_i = 1] = 1/2$. We have

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] \overset{\text{(c)}}{=} \sum_{i=1}^{n} \mathrm{E}[X_i] = \frac{n}{2}$$

by (c) linearity of expectation. We want to find an upper bound on the following:

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \geq \frac{3}{4}n\right] \leq ???.$$

Note that Markov's inequality gives us the bound,

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \geq \frac{3}{4}n\right] \leq 2/3.$$

This really doesn't tell us much. To do better, we amplify Markov's inequality. Previously we amplified Markov's inequality by squaring (Chebyshev's inequality) and by taking the fourth power

(for 4-wise independent variables). This time, we *exponentiate*. We first raise both sides by powers of 2 and then apply Markov's inequality:

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \geq .75n\right] = \mathrm{P}\left[2^{\sum_{i=1}^{n} X_i} \geq 2^{.75n}\right] \leq \frac{\mathrm{E}\left[2^{\sum_{i=1}^{n} X_i}\right]}{2^{.75n}}.$$

The good news, at this stage, is that the denominator is exponentially small in n. It remains to analyze the numerator, which looks a little complicated. We can make our lives easier by taking advantage of independence.

$$\mathrm{E}\left[2^{\sum_{i=1}^{n} X_i}\right] = \mathrm{E}\left[2^{X_1} \cdot 2^{X_2} \cdots 2^{X_n}\right] \overset{\text{(d)}}{=} \mathrm{E}\left[2^{X_1}\right] \mathrm{E}\left[2^{X_2}\right] \cdots \mathrm{E}\left[2^{X_n}\right].$$

(d) uses the fact that the variables $\left\{2^{X_1}, \ldots, 2^{X_n}\right\}$ are independent because $\{X_1, \ldots, X_n\}$ is an independent set of variables. (A function applied to each of a set of independent random variables begets a set of independent random variables, which is easily verified.) The net effect is similar to that of linearity of expectation, though to be clear it is not linearity of expectation. We have reduced the probabilistic analysis from a complicated function of many variables, $\mathrm{E}\left[2^{\sum_{i=1}^{n} X_i}\right]$, to a much simpler term, $\mathrm{E}\left[2^{X_i}\right]$, where $X_i \in \{0, 1\}$ models a fair coin toss. This second term can be analyzed explicitly. For each i, we have

$$\mathrm{E}\left[2^{X_i}\right] = \mathrm{P}[X_i = 0]2^0 + \mathrm{P}[X_i = 1]2^1 = \frac{3}{2}.$$

Plugging back in, we have

$$\frac{\mathrm{E}\left[2^{\sum_{i=1}^{n} X_i}\right]}{2^{.75n}} \leq \frac{(3/2)^n}{2^{.75n}} = 2^{(\log_2(3)-1.75)n} \leq 2^{-.165n}.$$

Thus,

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \geq .75n\right] \leq 2^{-.165n}.$$

If we plug in $n = 100$, we see that the probability of getting at least 75 heads out of 100 is $\leq .00001079$, which is very small after all. The effect is greater with larger n. For $n = 1000$ fair coin tosses, for example, the probability of getting at least 75% heads is $\leq 1/2^{165}$ – the odds of getting 165 consecutive heads! (And our analysis isn't even that tight.) This phenomena is much more general. One can have different coins with different probabilities of flipping heads. One can also have random variables that take continuous values between 0 and 1. The critical thing is that each "coin toss" is bounded and nonnegative in value.

Theorem 6 (Multiplicative Chernoff bound)**.** *Let $X_1, \ldots, X_n \in [0, 1]$ be independent random variables, and let $\mu = \mathrm{E}[X_1 + \cdots + X_n]$ be their expected sum. For $\epsilon \in (0, 1)$.*

$$\mathrm{P}[X_1 + \cdots + X_n \geq (1 + \epsilon)\mu] \leq e^{-\epsilon^2 \mu/2}$$

and

$$\mathrm{P}[X_1 + \cdots + X_n \leq (1 - \epsilon)\mu] \leq e^{-\epsilon^2 \mu/2}.$$

Observe that the multiplicative Chernoff bound above is *independent of* n, and directly a function of the expected sum, μ. In this sense the above Chernoff bound is said to be "dimension-free". (That said, we are still constrained by $\mu \leq n$.) Later we will also come across the *additive* Chernoff bound, (proven essentially the same way), that is not dimension free and has a (square root) dependence on n.

The moral of the story: *many small, independent parts adding up to a large sum in expectation is extremely well concentrated around the mean.*

9 Amplification amplification

Let us return to estimating distinct elements. We showed earlier that by using space proportional to $1/\delta\epsilon^2$, we get $(1 \pm \epsilon)$-error with probability $\geq 1 - \delta$. In particular, we can get any desired constant error with the average of $O\big(1/\epsilon^2\big)$ samples. We pay a *linear cost* for decreasing the error: to decrease the error by half, we double the number of experiments. This is helpful, but not as dramatic as with coins. In the coin tossing experiment above, going from 100 coin tosses to 1000 coin tosses (a factor of 10) takes the odds of getting $\geq 75\%$ heads from $2^{-16.5}$ to 2^{-165} – much more than a 10 fold decrease. How can we use coin tosses to inspire a more efficient experiment design?

Suppose we ran the experiment 100 times independently (in parallel), where in each trial we take the average of enough trials so that the error probability is $< .1$. We expect 90 of them to be correct (enough), and since they are independent, and as independent coin tosses, it should be well concentrated. For example, more likely than not, at least 50 of the estimators have additive error $\leq \epsilon/(k+1)$. Indeed, by the Chernoff bound (with $\epsilon = 4/9$), the probability of getting < 50 correct is less than

$$e^{-(4/9)^2 100/2} \leq .00005135.$$

Thus, almost certainly, at least half of the estimators are correct. How can we pluck out one of the correct ones? Of the remaining ones, so are too big, and so are too small. We are looking for something in the "middle".

The answer is *not* the mean. A single estimator that is atmospherically larger than all others will influence the mean too much. This is inherently unstable.[5] Rather, we take the *median* of our estimators.

Indeed, consider the median estimate. The median estimate is too high only if at least half of the estimates are too high. The median estimate is too low only if at least half the esimates are too low. But, as our coin-flipping analysis has shown, the majority of our estimates is correct with high probability. In this event, the median is *always correct*.

More generally, we can repeat the experiment $O(\log(1/\delta))$ times, rather than $O(1/\delta)$ times, to achieve probability of error $\leq \delta$.

This is called the *median trick*, or the *median of means*. By taking the *median of means*, we can efficiently reduce the error probability at an *exponential* rate. A colorful interface is given in Figure 1.

10 Distinct elements with pairwise independent hash functions

We have now developed a number of good ideas for the distinct elements problem. However we have continued to cheat in one critical way. Namely, we assumed access to ideal hash functions from $h : [m] \to [0, 1]$.

[5] The answer is also not the "mode", which is pretty useless in general.

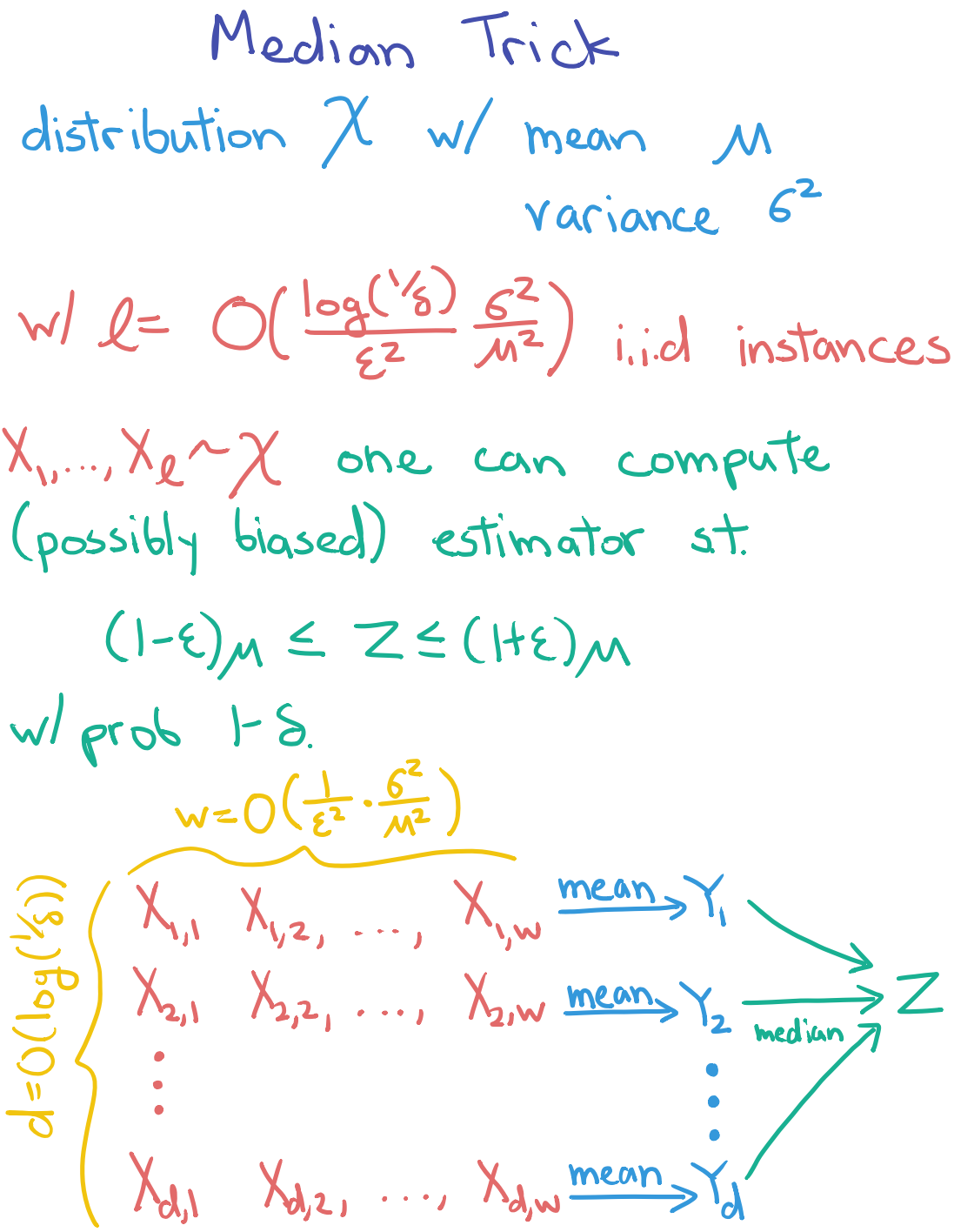

Figure 1: The median trick.

Instead, suppose we used pairwise independent hash function $h : [n] \to \{1/n^3, 2/n^3, \ldots, 1\}$. (That is, we create pairwise independent hash functions from $[n]$ to $[n^3]$, and divide the output by n^3.) Rather than track the smallest hash, we will track the rth smallest hash for $r = O(1/\epsilon^2)$. Let X be the rth smallest value be X for a single hash fuction. We return r/X as our estimate for the number of disinct elements.

We first analyze this procedure for a single hash function. Afterwards, we can take the median of means to get much more accurate and reliable estimators, like before...

11 Takeaways

- Like the heavy hitters problem discussed previously, keeping track of the number of distinct elements is easy to do when the data fits in memory, but impossible to do exactly with sublinear space in streaming settings. Thus we consider approximations.
- Assuming access to an ideal and continuous hash function h into the interval $[0, 1]$, the minimum hash over all elements leads does encode the number of distinct elements
- By taking the mean of a few independent trials, we can get small error with *constant* probability of success. However, the mean is not concentrated well enough to be able to amplify it directly.
- Rather, we take the *median* of several independently sampled means to amplify the success probability at an exponential rate. In general, the *median is more consistent then the mean.*
- To argue that the median is concentrated, we applied the *Chernoff inequality*, a generalization of the law of large numbers.
- Ultimately, we produce an *accurate and well-concentrated* estimate of the number of distinct elements, but the estimate is *not unbiased*. This was also true of the `count-min` data structure for heavy hitters.
- To adapt the approach to pairwise independent hash functions, we instead find that the rth minimum hash, for $r \approx 1/\epsilon^2$, is more helpful.

12 Exercises

Exercise 12.1. Consider the special case of the distinct elements streaming problem where there are $n+1$ total items in a stream, each of which is one of n different possible items $\{1, \ldots, n\}$. Show that any algorithm that maintains the number of distinct elements exactly throught the stream has to use at least n bits of memory. *Hint: argue that the algorithm must account for 2^n different possible states after the first n items in the stream.*

Exercise 12.2. Let $X \geq 0$ be a continuous and nonnegative random variable. Prove that

$$\mathrm{E}[X] = \int_0^\infty \mathrm{P}[X \geq t]\, dt.$$

As a helpful hint, here are the first two steps towards deriving the claim.

$$\mathrm{E}[X] \overset{\text{(e)}}{=} \int_0^\infty t f(t)\, dt \overset{\text{(f)}}{=} \int_0^\infty \int_{s=0}^t f(t)\, ds\, dt.$$

(e) is by definition of $\mathrm{E}[X]$. (f) expands out t to an integral.

Exercise 12.3. Prove Lemma 1.

Exercise 12.4. This exercise is about how for many intents and purposes, we approximately have the extremely convenient identity, "$1 + x = e^x$".

1. Prove that for all $x \in \mathbb{R}$, $1 + x \leq e^x$.

   Hint: At $x = 0$, both sides are equal. What are their respective rates of change moving away from 0?

2. Prove that for all $x \leq 1$, $e^x \leq 1 + x + x^2$.

Exercise 12.5 (Chernoff)**.** Recall that we proved a special case of the Chernoff bound in class. Here we prove the more general statement.

1. Let $X \in [0, 1]$ be a random variable and $t \leq 1$ a fixed value. Show that

$$\mathrm{E}\left[e^{tX}\right] \leq e^{(1+t)t\,\mathrm{E}[X]}.$$

2. Prove the following slightly more convenient form of the Chernoff bound. The following bounds lead to Theorem 6 (maybe with slightly different constants).

   Chernoff bound. *Let $X_1, \ldots, X_n \in [0, 1]$ be independent random variables and let $\mu = \mathrm{E}\left[\sum_{i=1}^{n} X_i\right]$. Let $\epsilon \in [0, 1]$. For any $\gamma > 0$, we have*

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \geq (1 + \epsilon)\mu + \gamma\right] \leq e^{-\epsilon\gamma}$$

   and

$$\mathrm{P}\left[\sum_{i=1}^{n} X_i \leq (1 - \epsilon)\mu - \gamma\right] \leq e^{-\epsilon\gamma}.$$