CS168: The Modern Algorithmic Toolbox
Lecture #13: Sampling and Estimation

Tim Roughgarden & Gregory Valiant*

May 12, 2021

This week, we will cover tools for making inferences based on random samples drawn from some distribution of interest (e.g. a distribution over voter priorities, customer behavior, ip addresses, etc.). We will also learn how to use sampling techniques to solve hard problems—both problems that inherently involve randomness, as well as those that do not.

As a warmup, to get into the probabilistic mindset, we will see a very cute, and useful algorithm for drawing samples from a datastream.

1 Reservoir Sampling

Problem: How can one efficiently sample k uniformly random elements from a datastream of length $N \gg k$? The two main issues are 1) N might be huge—we certainly don't want to store a significant fraction of the stream in memory, 2) N might be unknown—we might not know the length of the stream ahead of time, but we still hope that whenever the stream ends, we have a uniformly random sample. (Recall that we are asking this problem in a similar "streaming" setting as that of the Count-Min-Sketch algorithm we saw earlier.)

As a motivating example, suppose a router might want to sample 1,0000 random ip addresses from a given day's traffic. The router obviously doesnt want to store all of the ip addresses, and the router does not know ahead of time the total amount of traffic that it will see on a given day.

Algorithm 1
Reservoir Sampling [Vitter '85]
Given a number k, and a datastream $x_1, x_2, \ldots$ of length greater than k:

- Put the first k elements of the stream into a "reservoir" $R = (x_1, \ldots, x_k)$.
- For $i \geq k+1$
  - With probability $\frac{k}{i}$ replace a random entry of R with x_i.
- At the end of the stream, return the reservoir R.

We now show that at any time $t \geq k$, the reservoir R consists of a uniformly random subset of k of the entries of $x_1, \ldots, x_t$. To do this, let R_t denote the reservoir after the tth datastream element has been seen. It suffices to show that for all $t \geq i$, $\Pr[x_i \in R] = \frac{k}{t}$, and the event that $x_i \in R$ is independent of the contents of the reservoir at times $t < i$.

Claim 1.1 *For all $t \geq i$, $\Pr[x_i \in R_t] = \frac{k}{t}$, where R_t denotes the reservoir after time t.*

Proof: We prove this by induction on t. If $i \leq k$, the base case will be when $t = k$ and $\Pr[x_i \in R_k] = 1 = k/t$. If $i > k$, the base case will be $t = i$, and recall that the algorithm specifies that we include x_i in R with probability exactly $k/i = k/t$. For the induction hypothesis, assume that the statement holds for some fixed value of $t \geq i$. We now consider $\Pr[x_i \in R_{t+1}]$ and leverage our induction hypothesis and conditional probability:

$$\begin{aligned}
\Pr[x_i \in R_{t+1}] &= \Pr[x_i \in R_t] \cdot \Pr[x_i \text{ not replaced at time } t+1 | x_i \in R_t] \\
&= \frac{k}{t} \cdot \Pr[x_i \text{ not replaced at time } t+1 | x_i \in R_t] \qquad \text{by our induction hyp.} \\
&= \frac{k}{t} \cdot \left(1 - \frac{k}{t+1} \cdot \frac{1}{k}\right) = \frac{k}{t+1}.
\end{aligned}$$

Note that in the last line, we computed $\Pr[x_i \text{ not replaced at time } t+1 | x_i \in R_t]$ as $1 - \Pr[x_{t+1}$ replaces $x_i | x_i \in R_t] = 1 - \frac{k}{t+1}\frac{1}{k}$, where the $\frac{k}{t+1}$ term is the probability that x_{t+1} is stored in R_{t+1}, and the $1/k$ term is the probability that x_i is chosen as the random element of R_t to remove. ■

2 Basic Probability Tools: Our good friends Andrey [Markov] and Pafnuty [Chebyshev]

We begin with two of the most basic tools from probability theory that are extremely helpful for designing sampling schemes, and interpreting the significance of results gleaned from random samples.

In lecture we discussed some applications—the punch line was that data is everywhere, and that there is a huge difference between what is possible with a careful, mathematically principled analysis, versus a sloppy/crude analysis. Just ask Nate Silver, who has essentially made a career out of analyzing publicly available sports data and political poll data. I have added a link from the course page to a page on his website explaining his political prediction poll-aggregation system, in case you are interested.

Markov's inequality expresses the basic fact that "at most 10% of the population can have an income that is more than 10× the average income of the population":

Theorem 2.1 (Markov's Inequality) *For a real-valued random variable X s.t. $X \geq 0$, for any $c > 0$,*

$$\Pr\left[X \geq c\boldsymbol{E}[X]\right] \leq \frac{1}{c}.$$

Proof: Assume for the point of contradiction that $\Pr\left[X \geq c \cdot \mathbf{E}[X]\right] > \frac{1}{c}$, then $\mathbf{E}[X] > \frac{1}{c}c\dot{\mathbf{E}}[X] = \mathbf{E}[X]$, which is a contradiction. ■

Markov's inequality tells us something very simple about a distribution over real numbers, if all we know about the distribution is its expectation, and that it is non-negative. In many applications, it gives extremely weak bounds (e.g. the probability that a student's GPA is more that twice the average GPA is at most 1/2.....not very surprising). Nevertheless, Markov's inequality is the key tool that lets us prove more powerful theorems that give sharper characterizations of distributions for which we have more information. For example, it lets us prove Chebyshev's inequality, which tells us that the probability that a random variable is more than c standard deviations from its expectation, is at most $1/c^2$.

Theorem 2.2 (Chebyshev's Inequality) *For a real-valued random variable X, and any $c > 0$,*

$$\Pr\left[|X - \boldsymbol{E}[X]| \geq c\sqrt{\boldsymbol{Var}[X]}\right] \leq \frac{1}{c^2}.$$

Proof: Chebyshev's inequality is proved by applying Markov's inequality to a cleverly chosen random variable. Let $Y = (X - \mathbf{E}[X])^2$. Note that Y is a perfectly good random variable, satisfying $Y \geq 0$. To apply Markov's inequality, we need to know $\mathbf{E}[Y]$. First, note that $\mathbf{E}[Y] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{Var}[X]$.

$$\begin{aligned}
\Pr\left[|X - \mathbf{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] &= \Pr[Y \geq c^2\mathbf{Var}[X]] \\
&= \Pr\left[Y \geq c^2\mathbf{E}[Y]\right] \leq \frac{1}{c^2} \qquad \text{(by Markov's inequality)}
\end{aligned}$$

■

What is this good for? Lets see an example:

Example 2.3 How many people must we poll to estimate the percentage of people who support candidate C? Specifically, say we want our answer to be accurate to $\pm 1\%$, with probability at least 3/4.

Let $p \in [0,1]$ denote the true probability that a randomly chosen person supports candidate C. Suppose we poll n people, and output the fraction that support the candidate. Let $X_1, \ldots, X_n$ denote the 0/1 indicator random variables with X_i indicating that the ith person polled supports the candidate. Let $Z = \sum_i X_i$ denote the number of people polled who support the candidate.

$$\mathbf{Var}[Z] = \sum_i \mathbf{Var}[X_i] = n \cdot p(1-p).$$

In the above calculation, we used 1) the fact that the variance of a sum of independent random variables is the sum of the variances, and 2) a quick calculation showing that if X is 1 with probability p, and 0 otherwise, $\mathbf{Var}[X] = p(1-p)^2 + (1-p)p^2 = p(1-p)$.

An error in our estimate of 1% corresponds to estimating $\mathbf{E}[Z]$ up to an error of $0.01 \cdot n$. From Chebyshev's inequality,

$$\begin{aligned}
\Pr[|Z - \mathbf{E}[Z]| \geq 0.01n] &= \Pr\left[|Z - \mathbf{E}[Z]| \geq \frac{0.01n}{\sqrt{\mathbf{Var}[Z]}}\sqrt{\mathbf{Var}[Z]}\right] \\
&= \Pr\left[|Z - \mathbf{E}[Z]| \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\sqrt{\mathbf{Var}[Z]}\right] \\
&\leq \frac{p(1-p)}{n/100^2} \qquad \text{from Chebyshev's inequality.}
\end{aligned}$$

For any $p \in [0,1]$, the numerator $p(1-p) \leq 1/4$, and hence if $n \geq 100^2$, this probability is at most 1/4, as desired. Note that if we know that p is very small (i.e. $p < 0.1$) or very close to 1, then the numerator $p(1-p)$ is very small, and the required number of samples is correspondingly smaller.

What we have just shown is that, in general, to estimate the expectation of a 0/1 random variable to error $\pm\epsilon$, one needs roughly $O(1/\epsilon^2)$ independent samples.

If Chebyshev's inequality tells us something about a distribution given a bound on its variance, you might suspect that, given information on a distributions' higher moments (i.e. $\mathbf{E}[X^3], \mathbf{E}[X^4]$,, etc.) one might be able to obtain even sharper insights into the distribution. This intuition is true, and can be made rigorous via the same approach as the proof of Chebyshev's: apply Markov's inequality to a cleverly chosen random variable, corresponding to one of the higher moments of the distribution you care about. This is also how one proves "central limit" style inverse exponential bounds on tail probabilities (i.e. showing that the probability you flip fewer than 400 'heads' in 1000 tosses of a fair coin is miniscule.)

3 Importance Sampling

We mentioned "importance sampling" extremely briefly. The basic idea is that we can estimate properties of distribution A, based on samples from distribution B. Sometimes, A and B are fixed, and sometimes we can design distribution B to let us answer questions about A using fewer samples than we would need if we were to directly sample from A. This savings is achieved by having distribution B place higher weight on the "important" elements of the domain. The following examples will clarify this high level intuition.

Example 3.1 Suppose we are conducting a poll to estimate the fraction of the population that supports candidate X. Suppose we have some prior knowledge (e.g. from the census) that the population consists of 50% 'young' people and 50% 'old' people. Additionally, suppose we have a hunch that roughly 90% of the old people support candidate X, but only maybe 50% of the young people do. If our hunch is true, and we were to poll n random people, we would get an unbiased estimate of the fraction who support candidate X, and the variance would be $0.7(1-0.7)/n = 0.21n$. According to our hunch, however, the young people are responsible for a higher proportion of the variance than the older people, hence we could improve the variance by sampling more young people, and then reweighting the results so that our estimate ends up being unbiased. For example, suppose we poll $0.6n$ random young people, and $0.4n$ random old people. Letting X_{young} denote the number of young people polled who support candidate X and X_{old} denote the number of old people polled who support the candidate, an unbiased estimate of the true fraction is

$$X_{young}\frac{0.5}{0.6n} + X_{old}\frac{0.5}{0.4n}.$$

The factors of 0.5 come from our knowledge that 50% of the population is young/old, and the denominators are the number of young and old people we sampled. This is an unbiased estimate, just like the naive approach. But what is the variance? If our hunch is correct, the variance would be:

$$(0.5)(1-0.5)0.6n\left(\frac{0.5}{0.6n}\right)^2 + (0.9)(1-0.9)0.4n\left(\frac{0.5}{0.4n}\right)^2 \approx 0.16.$$

This is almost a 25% reduction in variance just by cleverly designing our sampling! Of course, if our hunch about how the young and old people tend to vote is wrong, this variance calculation is incorrect, though this might be alright, since 1) the overall estimate is still unbiased, and 2) after conducting the poll, we can see whether this hunch was correct or not, update our estimates of the voting behaviors of the young/old demographics, and update our estimate of the variance.

Example 3.2 Suppose we want to estimate the average income of the population. We know that the distribution of incomes has a long tail; i.e. there is a small fraction of people who have very large incomes, and hence have a large effect on the average income. This long, but important tail, is what makes the estimation task hard—if we don't get any samples from

the tail, then our estimate will likely be too low, but we need to take many samples in order to get any representatives from the tail.

We can use importance sampling as follows: Suppose we know that computer scientists have higher income than average, and that computer scientists compose exactly a 0.05 fraction of the total population. Rather than taking n random samples from the population, suppose we take $0.8n$ samples of non-computer scientists, and $0.2n$ samples of random computer scientists. If a_1 is the average salary of the non-computer scientists in our sample, and a_2 is the average salary of the sample of computer scientists, we can estimate the population average salary as:

$$\hat{avg} = 0.95a_1 + 0.05a_2.$$

The benefit of this is that by taking more samples from higher earners, we obtained a better estimate of this important tail. Of course, the final calculation needs to reweight these two samples so that our answer is still an unbiased estimate of the overall population average.

Formally, how does one characterize the improvement that we obtained by over sampling the computer scientists? It does not change the expectation of our answer—our estimate is still the unbiased estimate of the population average. The benefit is that we have reduced the variance of our estimate, by focussing our samples on the "important" portion of the distribution—the part of the distribution that contributes more variance towards the quantity we are trying to estimate.

4 Knowing the unknowns: Estimating the missing mass

Thus far we have discussed interpreting random samples drawn from distributions over 0/1, or distributions over real numbers (as in estimating average salary). In many applications, particularly in natural language processing, and genetics, one obtains samples from a distribution that is supported on a huge number of incomparable items. For example, a distribution over words, or a distribution over rare genetic variants. In these settings, one might not even know the support of the distribution.

While it is generally not possible to know what elements of the support are missing from a given sample, we can still hope to estimate the fraction of the distribution that is composed of the missing elements. One concrete question is the following:

> Given a set of random samples drawn from a distribution, what is the probability that the next sample we draw is a "new" domain element that we have not seen previously? Equivalently, what is the amount of probability mass in the distribution composed of domain elements that are not represented in our sample (i.e. the "missing mass")?

This question was first studied by I.J. Good and Alan Turing, in their work at Bletchley Park, in their work during WWII. (They wanted to estimate the probability that the next enigma machine ciphertext would be a previously unseen ciphertext.) They proposed the

following extremely simple estimate of the missing mass, known as the *Good–Turing* frequency estimation scheme. Given n independent draws from a distribution, they proposed the following estimate:

$$\Pr[\text{next draw is something new}] \approx \frac{\#\text{ elmts seen exactly once}}{n}.$$

There is a large literature analyzing and extending this basic estimate—just search for "Good-Turing frequency estimation".

Example 4.1 Assuming radio stations play songs in a random order, if you have been listening to the radio for the past 5 hours, the probability that the next song you here is a song you haven't heard in the past 5 hours can be estimated as the fraction of songs that you have heard exactly once in the past 5 hours.

Example 4.2 An estimate of the probability that the next word you read is a new word that you haven't seen before, is the number of words that you have seen exactly once, divided by the total number of words that you have seen. Keep in mind that these lecture notes are just an abbozzo.

Here is a quick sketch/derivation of this estimate: Assume that we have n samples from a distribution p supported on some unknown domain X, with $p(x)$ denoting the probability that the distribution assigns to element $x \in X$.

$$\begin{aligned}
E[\text{"unseen mass"}] &= \Pr[\text{next draw is new}] = \sum_{x\in X} p(x)\cdot \Pr[x \text{ unseen}] \\
&\approx \sum_{x\in X} p(x)\cdot \Pr[Binomial(n,p(x))=0] \\
&= \sum_{x\in X} p(x)(1-p(x))^n \\
&= \frac{1}{n+1}\sum_{x\in X}(n+1)\cdot p(x)(1-p(x))^n \\
&= \frac{1}{n+1}\sum_{x\in X}\Pr[Binomial(n+1,p(x))=1] \\
&\approx \frac{1}{n}\sum_{x\in X}\Pr[Binomial(n,p(x))=1] = E\left[\frac{\#\text{ elmts seen exactly once}}{n}\right].
\end{aligned}$$

To conclude the proof sketch of why the amount of unseen probability mass is close to $\frac{\#\text{ elmts seen exactly once}}{n}$, we just need to argue that the "unseen mass" and the number of elements seen exactly once will both be closely concentrated about their expectations. This takes a bit of work, but is at least intuitively plausible.