Sampling and Streaming

Admin Corner

Sampling

Want to know: "Do you approve of the US Congress?"

YES = 1 NO = 0

Goal: Estimate fraction of population who say YES

Alg 1: Ask all 330 million Americans

Alg 2:
1. Sample k people at random
2. Collect their answers $X_1, \ldots, X_k \in \{0,1\}$
3. Output $= \frac{1}{k} \sum_{i=1}^{k} X_i = \hat{p}$ = estimate of fraction of YES

Goal: Pick k s.t. w/prob $1-\delta$, $|\hat{p} - p^{\text{true value}}| \leq \varepsilon$

Economist / YouGov Poll:

8. Approval of U.S. Congress

Overall, do you approve or disapprove of the way that the United States Congress is handling its job?

		Gender		Race			Age				Income		
	Total	Male	Female	White	Black	Hispanic	18-29	30-44	45-64	65+	<50K	50-100K	100k +
Strongly approve	4%	4%	4%	2%	7%	7%	7%	7%	2%	1%	6%	2%	5%
Somewhat approve	12%	14%	10%	11%	17%	14%	21%	12%	8%	10%	9%	14%	13%
Approve	**16%**	**18%**	**14%**	**14%**	**24%**	**21%**	**28%**	**19%**	**10%**	**11%**	**15%**	**16%**	**18%**
Neither approve nor disapprove	17%	18%	16%	15%	26%	19%	23%	21%	13%	12%	17%	17%	13%
Somewhat disapprove	22%	20%	24%	26%	9%	15%	15%	16%	25%	31%	22%	21%	25%
Strongly disapprove	37%	39%	35%	40%	33%	31%	22%	33%	46%	45%	39%	38%	38%
Disapprove	**59%**	**59%**	**59%**	**66%**	**41%**	**46%**	**36%**	**48%**	**71%**	**76%**	**61%**	**60%**	**63%**
Not sure	8%	5%	10%	6%	9%	14%	13%	11%	6%	1%	7%	7%	5%
Totals	100%	100%	99%	100%	101%	100%	101%	100%	100%	100%	100%	99%	99%
Unweighted N	(1,496)	(692)	(804)	(999)	(170)	(213)	(202)	(303)	(651)	(340)	(607)	(431)	(304)

Margin of error ± 3.1% (adjusted for weighting)
± 3% (Registered voters)

answer is correct up to error ±3.1% w/prob 95%

aka: $|\hat{p} - p| \leq \varepsilon$ w/prob $1-\delta$, $\varepsilon = .031$, $\delta = .05$

Alg 2: 1. Sample k people at random

2. Collect their answers $X_1, \ldots, X_k \in \{0,1\}$

3. Output $= \frac{1}{k} \sum_{i=1}^{k} X_i = \hat{p}$ = estimate of fraction of YES

Thm (Chernoff bound):

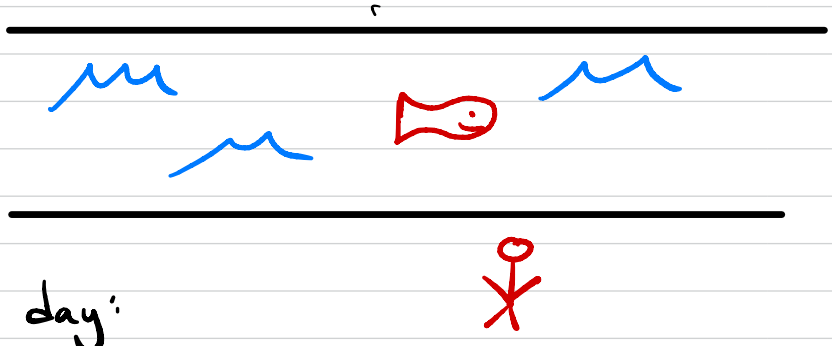
Suppose $X_1, \ldots, X_t \in \{0,1\}$ are independent, identically distributed random variables so that each $\Pr[X_i = 1] = p$, $\Pr[X_i = 0] = 1-p$.

$$\Pr\left[\left|\frac{1}{t}\sum_{i=1}^{t} X_i - p\right| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 t}$$

- $\frac{1}{t}\sum_{i=1}^{t} X_i$: average of samples
- p: expectation
- $e^{-2\varepsilon^2 t}$: decays exponentially in # of samples

To get $\Pr[\text{error} > \varepsilon] \leq \delta$, pick $t = \lceil \frac{1}{2\varepsilon^2} \cdot \log_2\left(\frac{1}{2\delta}\right) \rceil$.

Streaming



At end of day:

1) What fraction of fish were red?
2) How many fish swam by?
3) How many fish species?

Data: 1. Too large to store
2. Comes in a stream
3. No rewind

Sampling from a stream

Input: a stream $s_1, \dots, s_i, \dots \in \{1, \dots, N\}$

Goal: output one uniformly random element from stream.

Note: If L = length of stream is known

1.) Pick a random index $i \in \{1, \dots, L\}$

2.) Output s_i

Reservoir Sampling

1. "Reservoir" $r = s_1$
2. For each $i \geq 2$, flip a biased coin
   - if (H), set $r = s_i$
   - if (T), do nothing.
3. When stream stops, output r.

$$\begin{cases} \text{(H) w/prob } p_i = 1/i \\ \text{(T) w/prob } 1 - p_i \end{cases}$$

Claim: If stream has length L, $\Pr[\text{output } s_i] = 1/L$, $\forall i$

Pf: To output s_i need (H) on step i, (T) on steps $i+1, \ldots, L$.

$$\Pr[\underset{i}{H}\ \underset{i+1}{T} \ldots \underset{L}{T}] = \frac{1}{i}\left(1 - \frac{1}{i+1}\right)\left(1 - \frac{1}{i+2}\right)\cdots\left(1 - \frac{1}{L-1}\right)\left(1 - \frac{1}{L}\right)$$

$$= \frac{1}{i}\left(\frac{i}{i+1}\right)\left(\frac{i+1}{i+2}\right)\cdots\left(\frac{L-2}{L-1}\right)\left(\frac{L-1}{L}\right) = \frac{1}{L} \quad \square$$

Distinct elements

Input: stream $s_1, \dots, s_i, \dots \in \{1, \dots, N\}$
Goal: estimate # of distinct elements in stream

Algorithm (ideal)

1. Pick a random hash function $h: \{1, \dots, N\} \to [0,1]$
2. Compute (as the stream goes by)
   minimum of $h(s_1), \dots, h(s_L) = \alpha$ ← only need to store one number!
3. Output $1/\alpha$.

Intuition

Suppose k distinct elements in stream.
Then k hash values $r_1, \dots, r_k$. α = min of these

Fact: Can prove $E[\alpha] = \frac{1}{k+1}$!

Problems with random $h: \{1, \ldots, N\} \to [0,1]$

1. Computers can't store arbitrary real numbers

Sol^n: Pick $h: \{1, \ldots, N\} \to \{1, \ldots, R\}$, R is large

So $h(i)/R \approx$ random number in $[0,1]$

2. If $h: \{1, \ldots, N\} \to \{1, \ldots, R\}$ is uniformly random, need $N \log_2 R$ bits to store

Sol^n: Make h a pseudorandom hash function