Ex: To estimate the size of set S of labeled elements, sample with replacement from S from a uniform distribution until the first duplicate. If k elements have been drawn, then an unbiased estimator is $|S| = 2k^2/\pi$.

- -

Ex: Counting distinct words Flajolet & Martin, "Probabilistic Counting Algorithms for Database Applications", *J. Comp. Syst. Sciences*, 1985 Assume there are n (not necessarily distinct) words in a text, and you want to estimate the # of distinct words. Let there be a hashing function $h:\{\text{words}\} \to \{0,1\}^{5+\lg n}$. If s is a string of bits, let $\pi(x,b)$, $b\in\{0,1\}$, denote the index of the leftmost bit of x equal to b

signature $= s_1 s_2 \ldots s_{5+\lg n} = 00\ldots0$

for each word x **do** $signature_{\pi(h(x),1)} \leftarrow 1$

return $\left(\dfrac{2^{\pi(signature,0)}}{1.574}\right)$

If $\pi(signature,0)=4$, then leftmost bits of *signature* =1110. If there are 16 distinct words on the tape, Pr{none of hash codings begins 0001}=$\left(\frac{15}{16}\right)^{16}$ =.356

****

Hash Coding - Chapter 11 Assume n balls uniformly independently distributed among $m \ge n$ bins. What is probability of no collisions?

$$\prod_{1\le j\le n-1}\left(1-\frac{j}{m}\right) \approx \prod_{1\le j\le n-1} e^{-\frac{j}{m}} = e^{-\sum_{1\le j\le n-1}\frac{j}{m}} = e^{-\frac{n(n-1)}{2m}} \approx e^{-\frac{n^2}{2m}}$$

where first approximation justified by $e^{-j/m} = \left(-\frac{j}{m}\right)^0 + \frac{(-j/m)^1}{1!} + \frac{(-j/m)^2}{2!} + \ldots \approx 1-\frac{j}{m}$

for $j \ll m$. For Birthday Paradox, $e^{-\frac{n^2}{2m}} \approx \frac{1}{2}$ implies $\frac{n^2}{2m} \approx \ln 2 \Rightarrow n = \sqrt{2m\ln 2}$ and

m=365$\Rightarrow n = 22.49$

A *hash function* $h: U \to [0\ldots m-1]$ where $|U| \gg m$. A *good* h distributes sampled subset $S = \{a_1,\ldots,a_n\}$ of U uniformly among m *buckets* (bins).

Ex: (Probabilistic set membership) Want to test if $p \in S = \{a_1,\ldots,a_n\}$ $\quad h: U \to [2^b]$

Need Pr{false negative}=0 and low Pr{false positive}.

Preprocess $\{h(a_1),\ldots,h(a_n)\}$.

for $i \leftarrow 1$ **to** 2^b **do** $A[i] \leftarrow 0$

for $i \leftarrow 1$ **to** n **do** $A[h(a_i)] \leftarrow 1$

Given p, test if $\exists x \in S\; h(x) = h(p)$.

return $A[h(x)]$

ANALYSIS: Fix $x \in S$ $\Pr\{h(x) = h(p)\} = 2^{-b}$. $\Pr\{h(x) \ne h(p)\} = 1-2^{-b}$.

$\Pr\{\text{false positive}\} \geq 1-\left(1-2^{-b}\right)^{n} \geq 1-e^{-n/2^{b}}$ If we want $\Pr\{\text{false positive}\} \leq c \Rightarrow$

$1-e^{-n/2^{b}} \geq 1-c \Rightarrow b \geq \lg \dfrac{n}{\ln\left(1/\left(1-c\right)\right)}$ We need $\Omega\left(\lg n\right)$ bits.

<u>EX</u>: Bloom Filters Boolean $A\left[0..m-1\right]$ boolean, initialized to 0

k independent hash functions $h_1,\ldots,h_k : U \rightarrow \left[0,\ldots,m-1\right]$

for $i \leftarrow 1$ **to** *n* **do**
  for $j \leftarrow 1$ **to** *k* **do**
    $A\left[h_j\left(a_i\right)\right] \leftarrow 1$

MEMBER?(*p*, *A*)
  for $j \leftarrow 1$ **to** *k* **do**
    if $A\left[h_j\left(a_i\right)\right] = 0$ **then return** $p \notin S$
  return $p \in S$

$\Pr\{\text{false negative}\}=0$. $\Pr\{\text{false positive}\}$ – $\forall j \ \Pr\left\{A\left[j\right]=0\right\} = \left(1-\dfrac{1}{m}\right)^{kn} \approx e^{-kn/m}$

So $\Pr\{\text{false positive}\} = \left(1-\left(1-\dfrac{1}{m}\right)^{kn}\right)^{k} \approx \left(1-e^{-kn/m}\right)^{k}$.

Choose larger or smaller *k*?
- · Larger *k*$\Rightarrow$ For $p \notin S$ more chances to find a 0 in *A*
- · Larger *k*$\Rightarrow$ more 0's in *A*

Want to choose *k* to minimize $\left(1-e^{-kn/m}\right)^{k}$? Taking derivative, global minimum is $k = \left(\ln 2\right)\left(m/n\right) \Rightarrow$ So $\Pr\{\text{false positive}\}=2^{-k} \approx 0.6185^{m/n}$, which falls exponentially with *m*/*n*.

****

FINDING AVERAGE GRADE *n*>2 friends want to compute average grade while conveying no knowledge about their grades, $s_1,\ldots,s_n$. No player cheats; no trusted $(n+1)^{\underline{st}}$ party.

Choose *m*> largest possible $\sum_{n \geq i \geq 1} s_i$

Each player *i* chooses $x_{i,1},\ldots,x_{i,n-1} \sim U\left[0,m-1\right]$

Each player *i* chooses $x_{i,n}$ to satisfy $\sum_{n \geq k \geq 1} x_{i,k} = s_i \bmod m$

Each player *i* distributes one of $\left\{x_{i,1},\ldots,x_{i,n}\right\}$ (without replacement) to every player

Each player *i* computes and announces sum of numbers held (mod *m*) (say S_i)

Average grade $= \sum_{n \geq i \geq 1} s_i / n = \sum_{n \geq i \geq 1} S_i \text{ (mod m)} \big/ n$

Claim: For any set T of friends, $|Y| < n-1$, the only information they can infer about other grades is $\sum_{i \notin T} S_i$

DEF: Events *A*, *B* *independent* if $\Pr[A] = \Pr[A|B]$. *k-wise independent*.

(Return to) Claim: $(\forall i)(\{x_{i1},...x_{in}\}$ is $(n-1)$ – wise independent$)$