Hash Estimation

Preliminaries

Combinations → Order Does not Matter
Permutations → Order Matters
A Permutation is an **ordered** Combination.

Permutations with Repetition

When we have n things to choose from ... we have n choices each time! When choosing r of them, the permutations are: n × n × … (r times) (In other words, there are n possibilities for the first choice, THEN there are n possibilites for the second choice, and so on, multiplying each time.

$$n \times n \times \ldots \text{(r times)} = n^r$$

Permutations without Repetition

Now we cannot replenish our choices. So unlike above our choices are reduced every time we choose, so instead of n^r we have to deal with a new quantity n! Meaning n factorial. So the number is reduced each time we pull to

$$P(n,r) = {}^nP_r = {}_nP_r = \frac{n!}{(n-r)!}$$

Combinations without Repetition

So we adjust our permutations formula to **reduce it** by how many ways the objects could be in order (because we aren't interested in their order any more):

$$\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!(n-r)!}$$

As well as the "big parentheses", people also use these notations:

$$C(n,r) = {}^nC_r = {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Combinations with Repetition

Where ***n*** is the number of things to choose from, and we choose ***r*** of them
(Repetition allowed, order doesn't matter)

$$\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$$

Generalizing (use n,r with n=5, r = 3)

Permutation with Repetition	>	Permutation without Repetition	>	Combination with Repetition	>	Combination without Repetition
125	>	60	>	35		10

Probability

Definition
An **experiment** is a situation involving chance or probability that leads to results called outcomes.
An **outcome** is the result of a single trial of an experiment.
An **event** is one or more outcomes of an experiment.
Probability is the measure of how likely an event is.

With these definitions then

$$P(A) = \frac{\text{The Number Of Ways Event A Can Occur}}{\text{The total number Of Possible Outcomes}}$$

And of course, the reasons why we presented combinations and permutations above is so that we can count all the ways an event can occur with all the possible outcomes.

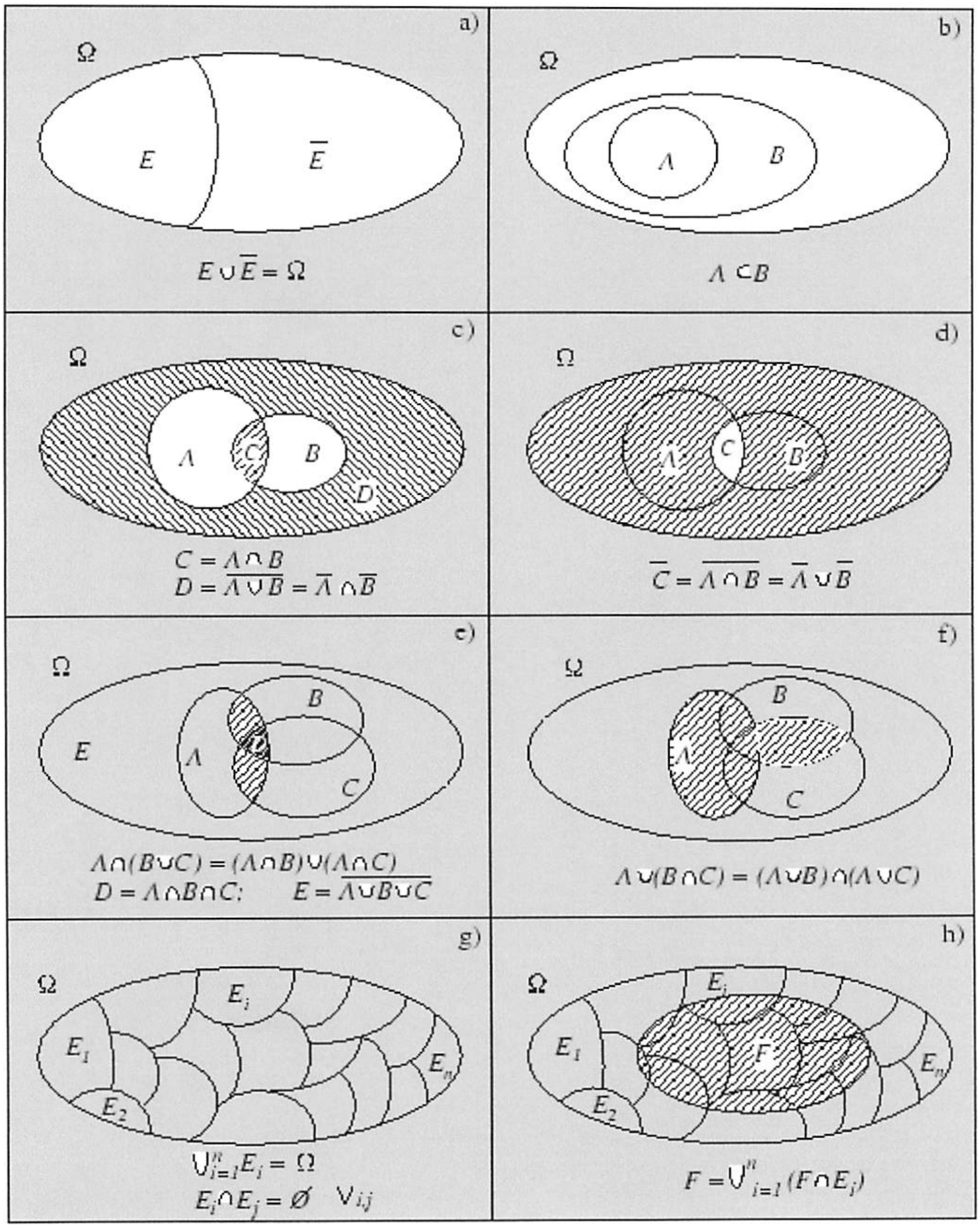
Summary of Probabilities

Event	Probability
A	$P(A) \in [0, 1]$
not A	$P(A^c) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive
A and B	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B)$ if A and B are independent

[illegible] 10/14/14

A given B $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B|A)P(A)}{P(B)}$

Putting some concepts of this in a Venn Diagram



Probability Union Bound

In probability theory, **Boole's inequality**, also known as the **union bound**, says that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events. Boole's inequality is named after George Boole.

Formally, for a countable set of events $A1, A2, A3, \ldots$, we have

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

Approximations

The number e occurs naturally in connection with many problems involving asymptotics. A prominent example is Stirling's formula for the asymptotics of the factorial function, in which both the numbers e and π enter:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Another useful formula to the following limit for $1/e$:

$$\frac{1}{e} = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n . \rightarrow (1\text{-}i/n) \text{ approx} = e^{-i/n}$$

Series Summations

It can be shown (for instance by mathematical induction) that

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Solution

C = Collision Occurs (Event) $\qquad \overline{C}$ = Collision does not occur (Event)

$\overline{C}_i \rightarrow i^{th}$ hash does not collide

Given m hashes and N positions calculate the probability of no collisions i.e. $P(\overline{C})$ $\quad m < N$

Counting technique

$$P(\overline{C}) = \frac{\text{Permutations without repetition}}{\text{Permutations with repetitions}}$$

$$= \frac{{}_{N}P_{m}}{N^{m}}$$

Conditional Probability

$$P(\overline{C}) = P\left(\bigcap_{i=1}^{m} \overline{C}_i\right) = P(\overline{C}_1) \cdot P(\overline{C}_2 / C_1) \cdot P(C_3 / C_2 \cap C_1)$$
$$\cdots P\left(C_m / \bigcap_{i=1}^{m-1} C_i\right)$$

$$= \left(1 - \frac{0}{N}\right) \cdot \left(1 - 1/N\right) \cdot \left(1 - 2/N\right) \cdots \left(1 - \frac{m-1}{N}\right)$$

$$= \prod_{i=0}^{m-1} \left(1 - i/N\right)$$

$$e^{-i/N} \approx (1 - i/N) \qquad \text{so}$$

$$\prod_{i=0}^{m-1} e^{-i/N} = e^{-1/N \sum_{i=0}^{m-1} i}$$

$$-1/N \sum_{i=0}^{m-1} i = \frac{-m(m-1)}{2N} \qquad \text{so}$$

Analyze

$$e^{\frac{-m(m-1)}{2N}} = 1/2$$

$$\begin{aligned} \ln(\quad) &= -\ln 2 \\ &= -m(m-1)/2N \end{aligned}$$

$$\begin{aligned} m^2 &\approx (2 \ln 2) N \\ m &\approx \sqrt{2 \ln 2}\ N \\ &\approx 1.177\sqrt{N} \end{aligned}$$

[illegible] 10/14/14

Scheduling m jobs over N processors

Worse case, random assignment

$X_i \to i^{th}$ processor $k \to$ # of jobs assigned

Calculate

$$P(X_i < k) = 1 - P(X_i \geq k)$$

easier

Look at Union Bound

$$P\left(\bigcup_{i=1}^{N} (X_i \geq k)\right) \leq \sum_{i=1}^{N} P(X_i < k)$$

$$\frac{1}{2} = N * \frac{1}{2N}$$

$$P(X_i \geq k) \leq 1/2N$$

Looking at a biased coin

$$P(X = k) \to \binom{N}{d} p^{d} (1-p)^{N-d}$$

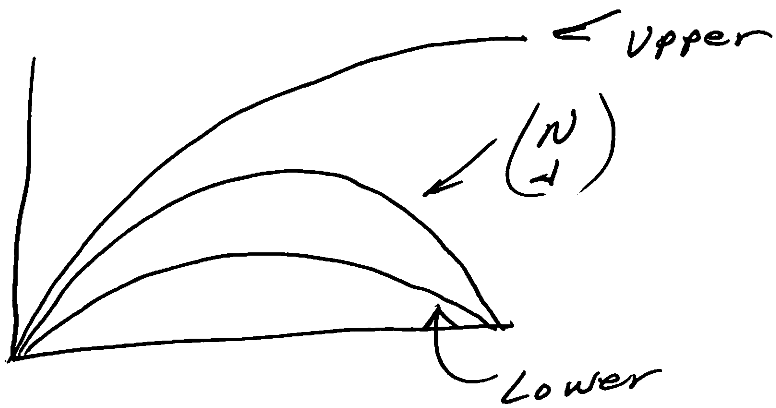
in our case $P = 1/N$

So

$$P(x \geq k) = \sum_{\cancel{k} = k}^{N} \binom{N}{j} \left(\tfrac{1}{N}\right)^{j} \left(1 - \tfrac{1}{N}\right)^{N-j}$$

Use bound for

$$\left(\frac{N}{j}\right)^{j} \leq \binom{N}{j} \leq \left(\frac{Ne}{j}\right)^{j}$$



$$\left(1 - \tfrac{1}{N}\right)^{N-j} \approx 1$$

$$\leq \sum_{j=k}^{N} \left(\frac{Ne}{j}\right)^{j} \left(\tfrac{1}{N}\right)^{j}$$

$$\leq \sum_{j=k}^{N} \left(\frac{e}{j}\right)^{j}$$

Upper Bound Proof

$$\binom{N}{k} \le \left(\frac{Ne}{k}\right)^k$$

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \qquad \text{but } \frac{{}_NP_k}{k!} = \binom{N}{k}$$

from previous problem

a)
$$\frac{{}_NP_k}{N^k} \approx e^{-k(k-1)/2N} \quad <1$$

$$_NP_k \, N^k \approx e^{-k(k-1)/2N} N^k$$

and

b)
$$e^k = \sum_{N=0}^{\infty} k^N/N! \rightarrow e^k > k^k/k!$$

$$\text{so } 1 < \frac{k!\,e^k}{k^k}$$

So with a) and b)

$$\binom{N}{k} < \frac{N^k}{k!} \cdot \frac{k!\,e^k}{k^k} = \left(\frac{Ne}{k}\right)^k$$

How to simplify

$$\le \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \left(\frac{e}{k}\right)^3 \cdots\right)$$

let $k > 2e$

$$\le 2\left(\frac{e}{k}\right)^k$$

$$2\left(\frac{e}{k}\right)^k = \frac{1}{2N}$$

$$\left(\frac{e}{k}\right)^k = \frac{1}{4N}$$

$$\ln\left(\frac{e}{k}\right)^k = \ln\left(1/4N\right) = -\ln 4 - \ln N$$

$$= k(1 - \ln k)$$

$$k(\ln k - 1) = \ln N + \ln 4$$

Randomly email to 300×10^6 americans,
No one would get 14

Mark 10/14/14 11