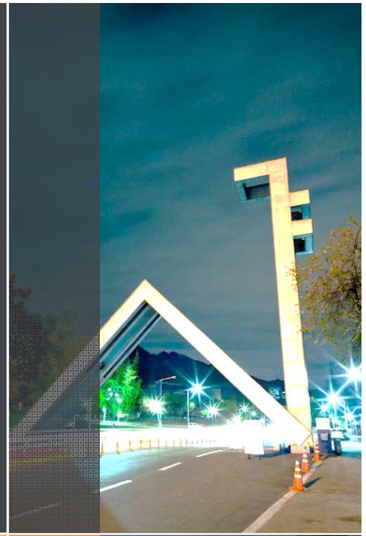
Seoul
National
University



Balls into Bins

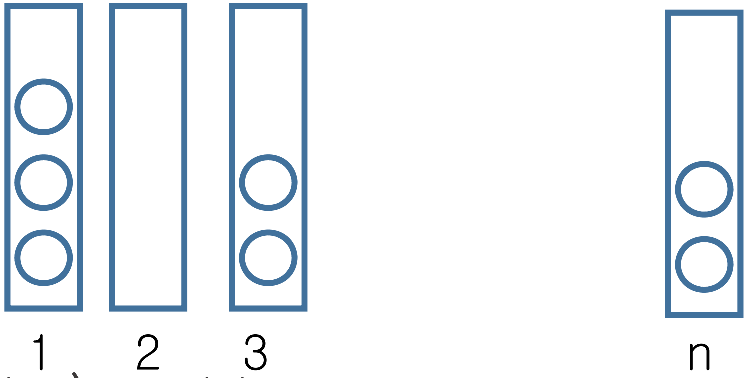
Name: Chong-kwon Kim

SCONE
Lab.

Balls & Bins

SCONE Lab.

- Put m balls into n bins uniformly at random



- Same (or similar) problems
  - Birthday paradox
  - Hash table
  - Coupon collection
  - Random allocation of requests to servers
  - Bucket sort, ……
- We are interested in various statistics such as
  - Number of (non)-empty bins
  - # balls in the most crowded bin
  - Max. # balls w/o collisions, ..

Difficult to obtain exact values
➔ Approximation

Load & Max. Load

- (Max) Load
  - Average (Maximum) load: Average (Maximum) # balls in a bin
  - If m=n, average load = m/n = 1
- Assume n balls into n bins. Then
  Pr(Max. load ≥ (3 ln n / ln ln n)) ≤ 1/n for large n
- Proof
  - Let E1: Event that bin1 receives at least M balls
    Pr(E1) = Pr(X_1 ≥ M) < $\binom{n}{M} \cdot (\frac{1}{n})^M$
  - Use inequalities and obtain $\binom{n}{M} \cdot (\frac{1}{n})^M < \frac{1}{M!} < (\frac{e}{M})^M$
  - Probability that any bin has at least M balls

$$= \Pr(\cup_{i=1}^{n} E_i) < n \ (\frac{e}{M})^M \quad \Leftarrow \text{For } M \geq 3 \ln n / \ln \ln n$$

$$\leq n \ (\frac{e \ln \ln n}{3 \ln n})^{3 \ln n / \ln \ln n}$$

$$\leq n \ (\frac{\ln \ln n}{\ln n})^{3 \ln n / \ln \ln n} = e^{\ln n}(e^{\ln \ln \ln n - \ln \ln n})^{3 \ln n / \ln \ln n}$$

$$= e^{-2 \ln n + 3 \ln n \cdot \frac{\ln \ln \ln n}{\ln \ln n}}$$

$$\leq 1/n \text{ (for large n such as } n = e^{e^{e^e}})$$

$$\frac{k^k}{k!} < \sum_{i=0}^{\infty} \frac{k^i}{i!} = e^k$$

$$k! > \left(\frac{k}{e}\right)^k$$

Union bound

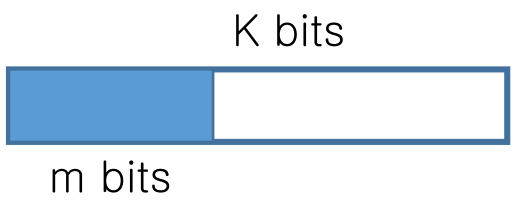
Compare to the prior approach that uses (Chernoff) inequalities

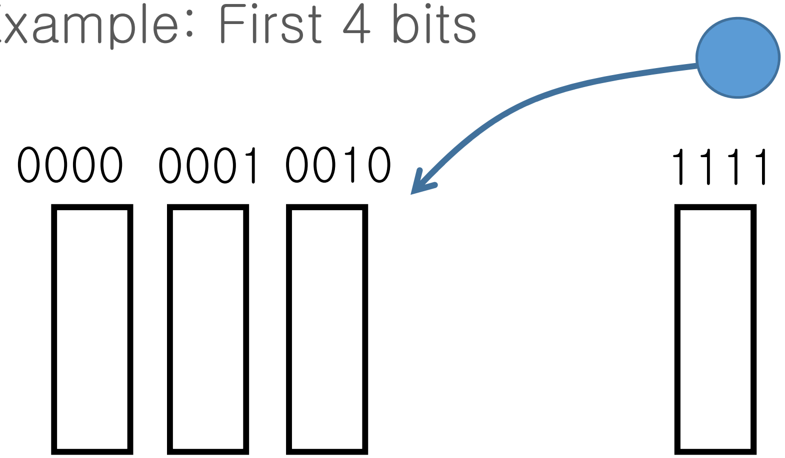
Application – Bucket Sort

SCONE Lab.

- Lower bound for average complexity of comparison based sorting is Ω(n log n)
  - Example: Quick Sort

- Complexity of bucket sort is O(n)
- Input: $n = 2^m$ elements each chosen uniformly at random from $[0, 2^K)$ where $K \geq m$
- First, sort based on left-most m bits
  - Example: First 4 bits





Assume national IDs consist of 13 digits
Each ID can be expressed as a 47 bit long number
There are $n \approx 2^{26}$ IDs
Sort based on the first 26 bits

Application – Bucket Sort

1. Prepare n buckets (bins) and put input elements based on first m bits
   - All elements in j th bucket is larger than those in i –th bucket if j > *i*
   - If we can put an element into bucket in O(1) ➔ O(n) for this stage
2. Sort elements in each bin and combine all bins

   ➔ Total complexity = O(n)

- Complexity of stage 2
  - X_j: # elements in bucket j

    X_j: B(n, 1/n) ➔ E[X_j] = 1, E[${X_j}^2$] = $\frac{n(n-1)}{n^2}$ + 1 = 2− 1/n < 2

    Var[X_j]

  - Complexity of sorting X_j elements = c ${X_j}^2 < cX_j log X_j$
  - E[$\sum_{i=1}^{n} {X_j}^2$] = 2·c·n

BB and Poisson Distribution

- Again m balls into n bins
- X: Number of empty bins
  - Let $X_j = \begin{cases} 1, & if\ bin\ j\ is\ empty \\ 0, & o.w \end{cases}$
  - E$[X_j]$ = $(1 - \frac{1}{n})^m$
  - E[X] = E$[\sum_{i=1}^{n} X_i]$ = $n \cdot (1 - \frac{1}{n})^m \approx n \cdot e^{-m/n}$

Poisson was a prolific French mathematician
First paper at 19 and more than 300 papers

Fraction of empty bins $= \boldsymbol{e^{-m/n}}$

- Generalize ➔ Prob. that a bin has exactly r balls, p_r

$$p_r = \binom{m}{r} \cdot (\frac{1}{n})^r \cdot (1 - \frac{1}{n})^{m-r}$$

$$= \frac{1}{r!} \frac{m(m-1)\cdots(m-r+1)}{n^r} \cdot (1 - \frac{1}{n})^{m-r}$$

$$\approx \frac{(m/n)^r e^{-m/n}}{r!}$$ (For r << m)

Poisson Distribution

- A discrete Poisson random variable X with parameter μ, Poi(μ), is

$$\Pr(X{=}j) = \frac{\mu^j e^{-\mu}}{j!}$$

- Properties
  - $\sum_{j=0}^{\infty} \Pr(X = j) = \sum_{j=0}^{\infty} \frac{\mu^j e^{-\mu}}{j!}$
  - $E[X] = \sum_{j=0}^{\infty} j \cdot \frac{\mu^j e^{-\mu}}{j!} \quad = \mu$

- Lemma: The sum of finite number of independent Poisson random variables is Poisson
- Proof
  - Let X and Y be independent Poisson with means μ_1 and μ_2
  - $\Pr(X{+}Y{=}j) = \sum_{k=0}^{j} \Pr(X = k) \cap (Y = j - k) \rightarrow \frac{(\mu_1+\mu_2)^j e^{-(\mu_1+\mu_2)}}{j!}$

Any other methods?

Poisson Distribution

- Alternatively, we can show Poi(μ_1) + Poi(μ_2) = Poi(μ_1+μ_2) using MGFs

Uniqueness of MGF

- First, show that MGF of a Poisson is

$$M_X(\mathrm{t}) = e^{\mu(e^t-1)}$$

- Let X and Y be two Poisson with means $\mu_1$and μ_2

$$M_{X+Y}(\mathrm{t}) \ = M_X(\mathrm{t})\ M_Y(\mathrm{t}) = e^{\mu_1(e^t-1)}e^{\mu_2(e^t-1)}$$

Bounds on Poisson Distribution

- Let X be a Poisson r.v. with mean μ
  1. If x > μ, then $\Pr(X \geq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$
  2. If x < μ, then $\Pr(X \leq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$

What method would you use?
Chernoff inequal. & find a proper t

- Proof
  - For any t > 0, and x > μ

$$\Pr(X \geq x) = \Pr(e^{tX} \geq e^{tx}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{tx}} = e^{\mu(e^t-1)-xt}$$

Min. at t = $\ln(x/\mu) > 0$

$$\rightarrow \Pr(X \geq x) \leq e^{x-\mu-x\ln(x/\mu)}$$

$$= \frac{e^{-\mu}(e\mu)^x}{x^x}.$$

  - For any t < 0 and x < μ

$$\Pr(X \leq x) = \Pr(e^{tX} \geq e^{tx}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{tx}}$$

Again, let t = $\ln(x/\mu) < 0$

Poisson as Limit of Binomial

- Already showed that Bins and Balls model can be approximated with a Poisson distribution
- Let Xn be a Binomial w/ parameters *n* and *p*, where *p* is function of *n* and λ= $\lim_{n\to\infty} np$ is a constant and is independent of *n*. Then for any k

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

More rigorous proof

- Proof
  - $\Pr(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$

$$\Pr(X_n = k) \le \frac{n^k}{k!} p^k \frac{(1-p)^n}{(1-p)^k}$$

$$\le \frac{(np)^k}{k!} \frac{e^{-pn}}{1-pk}$$

$$= \frac{e^{-pn}(np)^k}{k!} \frac{1}{1-pk}.$$

From Taylor expansion, for |x|≤1
$e^x(1- x^2) \le 1+x \le e^x$ (−p as x)

For k > 0, $(1-p)^k \ge 1\text{-pk}$

Poisson as Limit of Binomial

SCONE Lab.

$e^{x}(1-x^2) \le 1+x \le e^{x}$

- Also, $\Pr(X_n = k) \ge \dfrac{(n-k+1)^k}{k!} p^k (1-p)^n$

$$\ge \frac{((n-k+1)p)^k}{k!} e^{-pn}(1-p^2)^n$$

$$\ge \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n).$$

For k > 0, $(1-p)^k \ge 1\text{-pk}$

- Combining, we have

$$\frac{e^{-pn}(np)^k}{k!}\frac{1}{1-pk} \le \Pr(X_n = k) \le \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n)$$

$$\lim_{n\to\infty} \frac{e^{-pn}(np)^k}{k!}\frac{1}{1-pk} = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\lim_{n\to\infty} \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n) = \frac{e^{-\lambda}\lambda^k}{k!}$$

New Inequality

- We already learnt three bounds
- (Most) Analysis of a single bin can be done by applying the three inequalities
- Analysis of multiple bins requires additional bound
  - ➔ Poisson approximation

Poisson Approximation

- In Balls into Bins model, one difficulty is the *dependency* between bins
- If bin i is empty, then the probability that other bins are empty decreases
- Again throw m balls into n bins
- AND, Consider two sets of random variables

$\{\boldsymbol{X_1^{(m)}}, \boldsymbol{X_2^{(m)}}, \cdots, \boldsymbol{X_n^{(m)}}\}$ and $\{\boldsymbol{Y_1^{(m)}}, \boldsymbol{Y_2^{(m)}}, \cdots, \boldsymbol{Y_n^{(m)}}\}$

  - $X_i^{(m)}$ be the # balls in i–th bin
  - $Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_n^{(m)}$ are independent Poisson with mean m/n.

Exact Case: # balls in a bin when m balls are thrown to n bins
Poisson Case: # balls in a bin is Poisson with mean m/n

Poisson Approximation

- Theorem: The distribution of $\{Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_n^{(m)}\}$ conditioned on $\sum_i Y_i^{(m)} = k$ is the same as $\{X_1^{(k)}, X_2^{(k)}, \cdots, X_n^{(k)}\}$, regardless of m.

$Y_i^{(m)}$ is Poisson with mean m/n

$\forall\, i,\ Y_i^{(m)} = k_i \Leftrightarrow X_i^{(k)} = k_i$

- Proof
  - Throwing k balls into n bins, the probability that $\{X_1^{(k)}, X_2^{(k)}, \cdots, X_n^{(k)}\} = (k_1, k_2, \ldots, k_n)$, $\sum_i k_i$=k is given by
  - $\dfrac{\binom{k}{k_1;\, k_2;\ldots,k_n}}{n^k} = \dfrac{k!}{(k_1!)(k_2!)\cdots(k_n!)n^k}$.
  - For any $k_1, k_2, \ldots, k_n$ with $\sum_i k_i$=k , consider the probability that $\{Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_n^{(m)}\} = (k_1, k_2, \ldots, k_n)$ conditioned on $\sum_i Y_i^{(m)}$=k

Poisson Approximation

SCONE Lab.

$$\Pr\left((Y_1^{(m)}, \ldots, Y_n^{(m)}) = (k_1, \ldots, k_n) \;\middle|\; \sum_{i=1}^{n} Y_i^{(m)} = k\right)$$

$$= \frac{\Pr((Y_1^{(m)} = k_1) \cap (Y_1^{(m)} = k_2) \cap \cdots \cap (Y_n^{(m)} = k_n))}{\Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)}.$$

$Y_i^{(m)} \sim$ Poi(m/n) ⇨ $\sum_i Y_i^{(m)} \sim$ Poi(m)

$$\frac{\Pr((Y_1^{(m)} = k_1) \cap (Y_1^{(m)} = k_2) \cap \cdots \cap (Y_n^{(m)} = k_n))}{\Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)} = \frac{\prod_{i=1}^{n} e^{-m/n}(m/n)^{k_i}/k_i!}{e^{-m} m^k / k!}$$

$$= \frac{k!}{(k_1!)(k_2!)\cdots(k_n!)n^k},$$

Poisson Approximation

SCONE Lab.

Proved that Joint distributions of Exact Case and conditioned Poisson Case are the same. How about individual random variables? Or more generally

- Let f(x_1,.., x_n) be a nonnegative function. Then

$$\mathrm{E}[\mathrm{f}(X_1^{(m)}, X_2^{(m)}, \cdots, X_n^{(m)})] \leq \mathrm{e}\sqrt{m} \cdot \mathrm{E}[f(Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_n^{(m)})]$$

- Proof

$$\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})] = \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = k\right] \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)$$

Count k= m case only

$$\geq \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = m\right] \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = m\right)$$

Previous theorem

$$= \mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \Pr\left(\sum Y_i^{(m)} = m\right),$$

$$= \mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \frac{m^m \mathrm{e}^{-m}}{m!}$$

Note: $m! < \mathrm{e}\sqrt{m}\left(\frac{m}{\mathrm{e}}\right)^m$

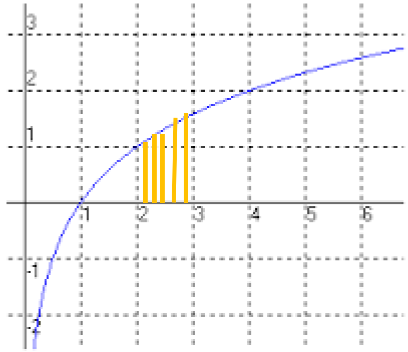
$$\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})] \geq \mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \frac{1}{\mathrm{e}\sqrt{m}}.$$

Poisson Approximation

SCONE Lab.

- To prove that

$$m! < e\sqrt{m}\left(\frac{m}{e}\right)^m$$



- Proof
  - Note that ln n is concave

$$\Rightarrow \quad \int_{i-1}^{i} \ln x \, dx \geq \frac{\ln(i-1) + \ln i}{2}.$$

$$\int_{1}^{n} \ln x \, dx \geq \sum_{i=1}^{n} \ln i - \frac{\ln n}{2}$$

$$n \ln n - n + 1 \geq \ln(n!) - \frac{\ln n}{2}.$$

  - Taking exponentiation to both sides, we obtain the result

Poisson Approximation

- Theorem: Any event that takes place with probability p in the PC takes place with probability at most $\mathrm{pe}\sqrt{m}$ in the EC

- Proof
  - Let f be the indicator function of the event (i.e f = 1 if the event occurs, f = 0 ow). Then E[f] is the probability that the event occurs.

Poisson Approximation

- Theorem: Let f(x_1,.., x_n) be a nonnegative function such that E[$f(X_1^{(m)}, X_2^{(m)}, \cdots, X_n^{(m)})$] is either monotonically increasing or decreasing in m. Then

$$E[f(X_1^{(m)}, X_2^{(m)}, \cdots, X_n^{(m)})] \le 2\, E[f(Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_n^{(m)})]$$

$$E[f(X_1^{(m)}, X_2^{(m)}, \cdots, X_n^{(m)})] \le E[f(X_1^{(m+1)}, X_2^{(m+1)}, \cdots, X_n^{(m+1)})]$$

Exercise 5.13 & 5.14

- Corollary: Let ε be an event whose probability is monotonically increasing (or decreasing) in the number of balls. If ε has probability p in the PC, then ε has probability at most 2p in the EC.

Lemma 5.12

- Claim: Assume n balls into n bins. The maximum load is at least ln n /ln ln n with probability at least 1−1/n.
- Proof:
  - Pr(Max. load is at least M) $\geq$ 1-1/n
    ➔ Pr(Max. load < M) $\leq$ 1/n
  - Consider Poisson case
    - Prob. that bin 1 has at least M balls $\geq 1/eM!$
    - Prob. that no bins has $\geq$ M balls = $(1 - 1/eM!)^n \leq e^{-n/eM!}$

Event that Max. load < M

  - Now, from Pr(EC) $\leq$ e$\sqrt{n}$ Pr(PC),
  need to prove that e$\sqrt{n}\,e^{-n/eM!}$ < 1/n
    ➔ Sufficient to prove that $e^{-n/eM!} < n^{-2}$
    ➔ M! < n / 2e ln n
    M! $\leq$ e$\sqrt{M}\,(\frac{M}{e})^M \leq$ M$(\frac{M}{e})^M$ when n are suitably large (also M are quite large)
    ➔ln M! $\leq$ M ln M – M + ln M

Lemma 5.12

- ln M! ≤ M ln M − M + ln M

$$= \frac{\ln n}{ln ln\, n}(\ln\ln n - \ln\ln\ln n) - \frac{\ln n}{ln ln\, n} + (\ln\ln n - \ln\ln\ln n)$$

$$\leq \ln n - \frac{\ln n}{ln ln\, n}$$

$$\leq \ln n - \ln\ln n - \ln(2e)$$

Appendix: Revisit Birthday Paradox

- Sequential selection
  - First person selects a birthday out of 365 days
  - Second person selects a birthday out of (365−1) days
  - ..
  - i-th person selects a birthday out of (365−i+1) days
- Pr(All m people have different birthdays) (Assume n different days)

$$= \prod_{j=1}^{m-1}(1-\frac{j}{n})$$

$$= \prod_{j=1}^{m-1} e^{-\frac{j}{n}}$$

$$= \exp\{-\sum_{j=1}^{m-1}\frac{j}{n}\}$$

$$\approx e^{-m^2/2n}$$

m << n

- Let Pr(No birthday match) = ½ ➔ $m^2/2n$ = ln 2

➔ m= $\sqrt{2n \cdot ln2}$ = $\Omega(\sqrt{n})$

  - For n=365, m ≈23