SAMPLING-BASED SELECTIVITY ESTIMATION

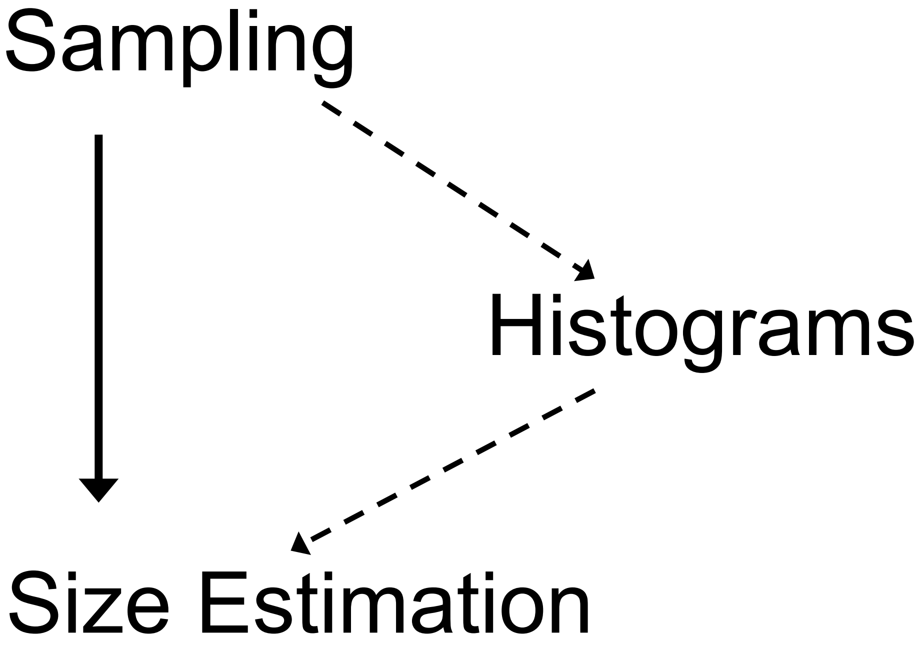
E0 261

Jayant Haritsa

Computer Science and Automation

Indian Institute of Science

BIG PICTURE



Sampling

Histograms

Size Estimation

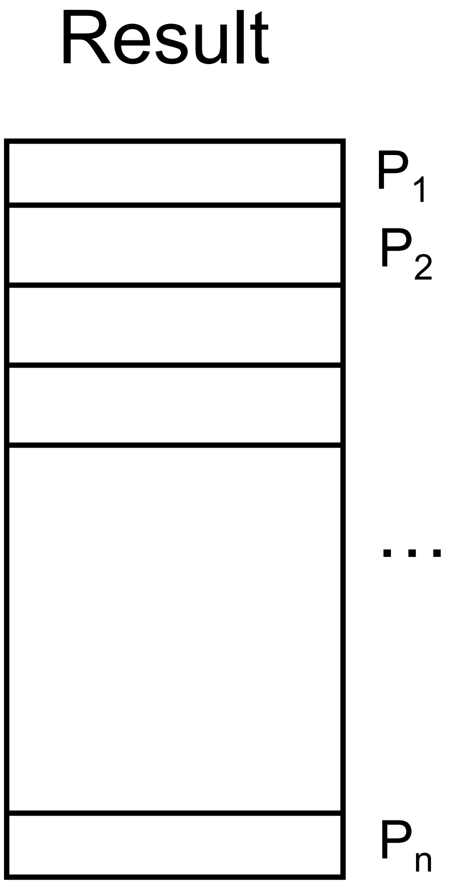
SAMPLING

- Pluses:
  - Fairly easy to compute and maintain samples
  - Multi-dimensional / correlations are not a problem
  - Gives whole tuple info
  - For queries with output from sample, "instant answers"
- Minuses:
  - Space consuming
  - Sensitive to data skew, especially low frequencies
  - May need to repeat for each query
  - Probabilistic guarantees (histogram gives deterministic guarantees)

TODAY's PAPER

- How to efficiently estimate selection and join result sizes using sampling
- Received Best Paper award in Sigmod 1990 !
  - Lipton is a top theoretician
- Basic Issues:
  - What kind and how many samples to take?
    - one disk I/O per sample, hence expensive and would like to minimize
  - What are the error bounds?

Partition Approach



- n partitions with $P_i \cap P_j = 0$
- Result = $\cup P_i$
- Randomly sample m partitions
- Let s = $\Sigma \mid P_i \mid$ of samples
- Estimate | Result | = (s/m) n
- What should n and m be ?
  - n : partition = tuple (or = page)
  - m: ?
- Partition result size (with tuple partitions)

  = 0 or 1 for selection

  = $t \bowtie S$ for join, 0 to |S|

Central Limit Theorem

- Number of samples = m
- Sample Mean (avg) = M
- Sample Variance = S^2

- True Mean = μ
- True Variance = σ^2

Then, CLT states that the (error) statistic

$$(M - \mu) / \sqrt{\ } \sigma^2 / m$$

will have N(0,1) distribution

$$\Phi(a) = \quad 1/\sqrt{2\pi}\int_{-\infty}^{a} e^{-x^2/2}\, d$$

(Approximate σ^2 by $S^{2)}$

What is remarkable is that this is true ***<u>independent of the sample distribution !</u>***

(assuming i.i.d.)

CLT
(de Moivre, Laplace, Lyapunov) 1733

Sir Francis Galton (1889):

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Estimation

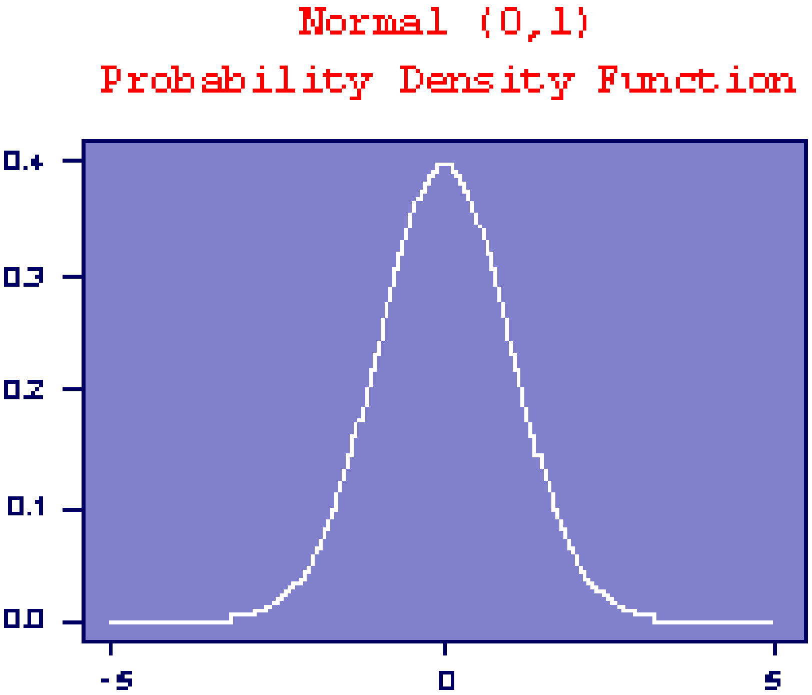
- Then $\mu = M \pm t_{1-\alpha/2,\, m-1} \sqrt{\sigma^2} / m$
  - that is, with a confidence of $(1 - \alpha)$, the true value of the distribution mean is within the term in the $\pm$

- Rewritten version of CLT in Defn. 2.1 with different symbols and equation organization

Definition 2.1 ≡ CLT

- $$\frac{\sum_1^m \mathrm{X_i} - mE}{\sqrt{mV}} = \frac{\sum_1^m \frac{\mathrm{X_i}}{m} - E}{\sqrt{\frac{V}{m}}} = \frac{M - \mu}{\sqrt{\frac{\sigma^2}{m}}}$$

Unit Normal Distribution



Student t distribution

1– α M-1	80%	90%	95%	99%
1	1.376	3.078	6.314	31.82
10	0.879	1.372	1.812	2.821
100	0.845	1.290	1.660	2.364
∞	0.842	1.282	1.645	2.326

Algorithm (Figure 1)

Average case

Worst case

```
- $s = 0$   $m = 0$
  while ( ($s < k_1 b d (d+1)$) and ($m < k_2 e^2$) ) do
      $s = s + \text{RandomSample}()$
      $m = m + 1$
  end
  **$\tilde{A} = n s / m$**
```

- within $(A/d, bn/e)$ of true A with prob p
- d and e are accuracy indicators
- b is upper bound on value of Random Sample()
- k_1 and k_2 are functions of p

Algorithm Parameter Settings

- d, e and p are specified by the user
- $k_1 = [\Phi^{-1} ((1 + \sqrt{p}) / 2)]^2$
- $k_2 = [\Phi^{-1} ((1 + p) / 2)]^2$

Theorem 2.2

Suppose that in a run of the algorithm in Figure 1, the while loop terminates because $s \geq k_1\, b\, d\, (d+1)$, and the CLT applies. Then, for $0 \leq p < 1$, if

$$k_1 \geq [\Phi^{-1}\,((1 + \sqrt{p})/\, 2)]^2$$

the error in $\tilde{A}$ is less than A/d with probability p.

Lemma 2.1

Squared coefficient of variation

Let $m_{sufficient} = \beta\left(\frac{\mathrm{V}}{\mathrm{E}^2}\right)$

and $s_{sufficient} = \alpha\left(\frac{\mathrm{V}}{\mathrm{E}}\right)$

α, β are constants

coefficient of dispersion

Then,

$$P\left[\sum_{i=1}^{m} X_i > \alpha\frac{V}{E}\right] = P\left[\frac{\sum_{i=1}^{m} X_i - mE}{\sqrt{mV}} > \frac{\alpha\frac{V}{E} - mE}{\sqrt{mV}}\right]$$

$$= P\left[\frac{\sum_{i=1}^{m} X_i - mE}{\sqrt{mV}} > \frac{\alpha - \beta}{\sqrt{\beta}}\right]$$

$$= 1 - \Phi\left(\frac{\alpha - \beta}{\sqrt{\beta}}\right)$$

Lemma 2.2

$$P\left[\left|\frac{n}{m}\left(\sum_{i=1}^{m} X_i\right) - A\right| \le \frac{A}{d}\right]$$

$$= P\left[\left|\frac{n}{m}\left(\sum_{i=1}^{m} X_i\right) - nE\right| \le \frac{nE}{d}\right] \qquad A = nE$$

$$= P\left[\left|\left(\sum_{i=1}^{m} X_i\right) - mE\right| \le \frac{mE}{d}\right] \qquad * \frac{m}{n}$$

$$= P\left[\left|\frac{\sum_{i=1}^{m} X_i - mE}{\sqrt{mV}}\right| \le \frac{mE}{d\sqrt{mV}}\right] \qquad * \frac{1}{\sqrt{mV}}$$

$$= P\left[\left|\frac{\sum_{i=1}^{m} X_i - mE}{\sqrt{mV}}\right| \le \frac{\sqrt{\beta}}{d}\right]$$

Now $P[|Z| \le \delta] = \Phi(\delta) - \Phi(-\delta)$

$$= \Phi(\delta) - (1 - \Phi(\delta))$$

$$= 2\Phi(\delta) - 1$$

$$\therefore \quad P\left[\left|\frac{\sum_{i=1}^{m} X_i - mE}{\sqrt{mV}}\right| \le \frac{\sqrt{\beta}}{d}\right] = 2\Phi\left(\frac{\sqrt{\beta}}{d}\right) - 1$$

Theorem 2.2

$$P(success) = P\left(\frac{success}{sufficient\ samples}\right)P(sufficient\ samples)$$

$\downarrow$ $\downarrow$

Lemma 2.2 Lemma 2.1

$$= \left[2\Phi\left(\frac{\sqrt{\beta}}{d}\right)-1\right] * \Phi\left(\frac{\alpha-\beta}{\sqrt{\beta}}\right)$$

Given $MN = k$, min value of $M + N$

is when $M = N = \sqrt{k}$

$\Rightarrow$ fewest required samples

$\therefore$ if user wants confidence level p,

$$\left(2\Phi\frac{\sqrt{\beta}}{d}-1\right)=\sqrt{p} \qquad \text{and} \qquad \Phi\left(\frac{\alpha-\beta}{\sqrt{\beta}}\right)=\sqrt{p} \qquad \Rightarrow$$

$\Downarrow$

$$\beta = \left[\Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)\right]^2 d^2$$

$$\begin{aligned}
\alpha &= \Phi^{-1}\left(\sqrt{p}\right)\sqrt{\beta} + \beta \\
&= \sqrt{\beta}\left(\Phi^{-1}\left(\sqrt{p}\right)+\sqrt{\beta}\right) \\
&= \Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)d\left(\Phi^{-1}\left(\sqrt{p}\right)+\Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)d\right) \\
&\leq \Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)d\left(\Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)+\Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)d\right) \\
&\leq \left[\Phi^{-1}\left(\frac{1+\sqrt{p}}{2}\right)\right]^2 d(d+1) \\
&\leq k_1 d(d+1)
\end{aligned}$$

Theorem 2.2 (contd)

$\therefore$ if the algorithm samples until $s \geq \alpha \dfrac{V}{E}$

i.e., $s \geq k_1 d(d+1)\dfrac{V}{E}$, then the desired accuracy holds with probability p.

Now, we prove $b \geq \dfrac{V}{E}$

$$V = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i \right)^2 \right]$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} X_i b$$

$$\leq b\frac{1}{n}\sum_{i=1}^{n} X_i \qquad \leq bE$$

$$\Rightarrow b \geq \frac{V}{E}$$

$\therefore$ if $s \geq k_1 bd(d+1)$, it automatically implies $s \geq k_1 \dfrac{V}{E} d(d+1)$

Skewed Data

- b >> E, meaning "10 percent of the samples produced 90 percent of the output tuples"
- 1 tuple out of a million
  - Expected size of a random sample is 1/million
  - Sampling requires $10^6 * k_1 \, d \, (d+1)$ samples
    - more than the number of tuples itself!
    - happens because of sampling with replacement!

Theorem 3.2

Suppose that in a run of the algorithm of Figure 1, the while loop terminates because $m > k_2 e^2$. Then for $0 \leq p < 1$, if

$$k_2 \geq [\Phi^{-1} ((1 + p)/ 2)]^2$$

the error in Ã is less than A_{max} / e with probability p.

Implementation

- Selection:
  - Sampling via Index (on any attribute of relation) for Selection
- Join:
  - Sampling via Index (on any attribute of source relation) and via Index (on join attribute of target relation)

END SAMPLING

E0 261