Lecture 9

$D = \langle R \rangle$

1° $f(D) = SUM\ x$

2° $f(D) = SUM\ g(x)$

$\rightarrow$ Selections

3° Generalizations

- more than one relation
- generalize to fct. that do not look like $\mathcal{F}\left(\sum_{t \in R} g(t)\right)$ $\rightarrow$ $AVG = \frac{SUM}{CNT}$

one possible generalization $\mathcal{F} : \mathbb{R} \to \mathbb{R}$

- Totally general fcts $f(R)$

Ex: $f(R) = \prod_{t \in R} g(t) = Exp\left(\sum_{t \in R} \log g(t)\right)$

$\log f(R)$

Two relations now. $D = \langle R, S \rangle$

Equi-join

Q: Is there a fundamental difference w.r.t. sampling between join and fct. over crossproducts?

$$f(D) = \sum_{(t_R, t_S) \in R \bowtie S} g(t_R, t_S)$$

$$R \bowtie S = \sigma(R \times S)$$

$$f(D) = \sum_{t_R \in R} \sum_{t_S \in S} \bar{g}(t_R, t_S)$$

$$\sum_{(t_R, t_S) \in R \times S}$$

$\bar{g}(t_R, t_S) = g(t_R, t_S) \cdot I_{t_R.A = t_S.A}(t_R, t_S)$

Problem:

Given R' sample of R, S' sample of S } samples without replacement

estimate $f(D) = \sum_{t_R \in R} \sum_{t_S \in S} g(t_R, t_S)$

for given fct $g(t_R, t_S)$

Sol:

$$X = \frac{|R|}{|R'|} \frac{|S|}{|S'|} \sum_{t_R \in R'} \sum_{t_S \in S'} g(t_R, t_S)$$

α, β

$$= \alpha \beta \sum_{t_R \in R} \sum_{t_S \in S} X_{t_R} Y_{t_S} g(t_R, t_S)$$

0/1

$$E[X] = \alpha \beta \sum_{t_R \in R} \sum_{t_S \in S} E[X_{t_R} Y_{t_S}] g(t_R, t_S) = f(D)$$

$$E[X_{t_R} Y_{t_S}] = E[X_{t_R}]\, E[Y_{t_S}]$$

$\frac{1}{\alpha}$ $\frac{1}{\beta}$

Side note:

Sup $R = S$

R' smp R
S' smp $S = R$) s.w. repl.

Q: Is this a complete generative model?

Comm: $R' = S'$

$\boxed{R' \text{ indep of } S'}$

$X(R')$ indep $Y(S')$

$$X(R') = \begin{cases} 1 & \text{if } t_R \in R' \\ 0 & \text{oth} \end{cases}$$

Side note

$R' = S'$

$$E[X_{t_R} \cdot Y_{t_S}] = \begin{cases} \frac{1}{\alpha} & t_R = t_S \\ \frac{1}{\alpha}\frac{1}{\alpha_1} & t_R \neq t_S \end{cases}$$

$$E[\lambda^2] = \nu^2 \beta^2 \sum_{t_R \in R} \sum_{t_R' \in R} \sum_{t_S \in S} \sum_{t_S' \in S}$$

$$\underbrace{E[X_{t_R} X_{t_R'}] \cdot E[Y_{t_S} Y_{t_S'}]}_{\text{indep}} \cdot \delta(t_R, t_S)\,\delta(t_R', t_S')$$

$\text{Prob}(U_{t_R, t_R'})$ $\qquad$ $\text{Prob}(S_{t_S, t_S'})$

$$\text{Prob}(U_{t_R, t_R'}, S_{t_S, t_S'}) =$$

$$\text{Pr} - \ldots \delta_{t_R, t_R'} + \ldots \delta_{t_S, t_S'} - \ldots \delta_{t_R t_R'} \delta_{t_S t_S'}$$

$\Rightarrow$ Some big term