Part 1 Proof

Yilin Guo

March 7, 2016

1 Proof

Provide a proof to derive the formulas for "SELECT AVG(X) FROM D WHERE c" query under "Fixed-size without Replacement" (i.e., row 3 and column 3) in Table 2 of the following paper: `http://web.eecs.umich.edu/~mozafari/php/data/uploads/approx_chapter.pdf`

θ_c is the estimator of approximating $\overline{X}_c$ using S (`AVG(X) FROM D WHERE C`), then θ_c equals the mean of sample tuples that satisfies condition, i.e. $\theta_c = \overline{Y_c}$.

$W_k = \frac{\binom{N_c}{k}\binom{N-N_c}{n-k}}{\binom{N}{n}}$ is the probability that select n samples $Y_1, Y_2, \ldots, Y_n$ among which exactly k samples $Y_{c1}, Y_{c2}, \ldots, Y_{ck}$ satisfying the condition. In the best case, there are at most $b = \min\{n, N_c\}$ samples to satisfy the condition. In the worst case, there are at least $a = \max\{1, n-(N-N_c)\} = \max\{1, n-N+N_c\}$ to satisfy the condition. Therefore, the expected value of the estimator could be calculated through the condition mean in D times the total probability of selecting n samples with possible satisfying conditions, i.e. $E[\theta_c] = \overline{X}_c \sum_a^b W_k = \overline{X_c} W$. If $N \leq N - N_c$, then $W \neq 1$, $E[\theta_c] - \overline{X}_c = \overline{X}_c W - X_c \neq 0$; in this case the estimator θ_c is biased. Otherwise, if $N < N - N_c$, then $W = 1$, $E[\theta_c] - \overline{X}_c = \overline{X}_c W - X_c = 0$; in this case the estimator θ_c is unbiased.