An Excursion in Data Streams

Sudipto Guha

March 6, 2008

1 Useful Random Facts about Random Things

Let X, Y and $X_1, X_2, \ldots, X_k$ denote random variables. $E[X]$ denotes the expectation and $V[X]$ denotes the variance $V[X] = E[X^2] - E[X]^2$.

1. *Linearity of Expectation:* $E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i]$. Note that this holds even if $\{X_i\}$ are dependent. If $\{X_i\}$ are independent, we can further say that $V\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} V[X_i]$.

2. *(Pairwise) Independence:* Two (discrete) variables X, Y are defined to be pairwise independent if $Pr[(X = a) \wedge (Y = b)] = Pr[X = a]Pr[Y = b]$ for all a, b. Consequences of this are: first $E[XY] = E[X]E[Y]$, and secondly if $\{X_i\}$ are pairwise independent, i.e., all X_i, X_j are pairwise independent, then $V\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} V[X_i]$.

3. *Markov Inequality:* If $X \geq 0$ always, then $Pr\left[X > t\right] \leq E[X]/t$.

4. *Chebyshev Inequality:* $Pr\left[\ |X - E[X]| > t\right] \leq V[X]/t^2$; note X need not be positive.

5. *Chernoff Bounds:* Let $\{X_i\}$ be independent indicator (0/1) variable such that $Pr[X_i = 1] = p_i$. Let $\mu = \sum_i p_i$, $X = \sum_i X_i$ and $0 < \delta \leq 1$ then

$$Pr\left[|X - \mu| > \delta\mu\right] < 2e^{-\mu\delta^2/3}$$