Variance Bounds on Binary Data Sets

Dennis Walsh
Middle Tennessee State University

I. Variance Formula

Let $D = \; < x_1,\, x_2, ...,\, x_n >$ denote an ordered data set of size n. The variance of D, denoted Var(D), is given by

$$\text{Var}(D) = \frac{1}{n^2} \sum_{i<j} (x_i - x_j)^2. \tag{1}$$

For binary data sets consisting only of ones and zeros, such as $< 0, 0, 0, 0, 1, 1, 1, 1, 1, 1 >$, we derive the variance and bounds on its maximum and minimum value.

For a binary data set, the summand $(x_i - x_j)^2$ in (1) can take on only values of 0 and 1. In particular, since the data is ordered, $(x_i - x_j)^2 = 1$ if and only if $x_i = 0$ and $x_j = 1$. Therefore, if $D = \; < 0, .., \;\; 0, 1, ..., 1 >$ with k ones and $(n-k)$ zeros, we have $\sum_{i<j} (x_i - x_j)^2 = \; k(n-k)$ and hence

$$\text{Var}(D) = \; \frac{k(n-k)}{n^2}. \tag{2}$$

II. Minimum Variance

If a data set D consists of all ones or all zeros, the $\text{Var}(D) = 0$. Otherwise, if $\text{Var}(D) > 0$, the data set must contain both one(s) and zero(s). For such data sets, the minimum nonzero variance occurs when there is exactly one "0" or exactly $(n-1)$ "0's". Therefore, for binary size-n data sets,

$$\text{Var}(D) \begin{cases} \geq 0 & \\ \geq \frac{n-1}{n^2} & \text{for data sets with digits 0 and 1.} \end{cases}$$

III. Maximum Variance

For fixed even n, the variance is maximized when $k(n-k)$ is maximized. But $k(n-k)$ is quadratic in k and is maximized when $k = \; -n/(-2) = n/2$. For fixed odd n, the variance is similarly maximized when $k = (n-1)/2$ or $k = (n+1)/2$. Therefore, for binary size-n data sets,

$$\text{Var}(D) \begin{cases} \leq \frac{1}{4} & \text{for even } n \\ \leq \frac{n^2-1}{4n^2} & \text{for odd } n. \end{cases}$$

IV. Table of $k(n-k)$

Values for $k(n-k)$

$n \backslash k$	0	1	2	3	4	5	6	7	8
0	0								
1	0	0							
2	0	1	0						
3	0	2	2	0					
4	0	3	4	3	0				
5	0	4	6	6	4	0			
6	0	5	8	9	8	5	0		
7	0	6	10	12	12	10	6	0	
8	0	7	12	15	16	15	12	7	0

The table above provides the numerator of the variance for binary data set D that has k ones and $(n-k)$ zeros. [We note that table entries appear as integer sequence A004247 in the *On-Line Encyclopedia of Integer Sequences* (http://www.research.att.com/~njas/sequences/.]

V. Extension of Results

Let r and m be nonzero real numbers. For binary data set D with $(n-k)$ zeros and k ones, let $E = rD + m$ denote the data set $< m, \ldots, m, r+m, \ldots, r+m >$ that consists of $(n-k)$ $m's$ and k $(r+m)'s$. The variance of E is given by

$$\text{Var}(E) = \text{Var}(rD+m)$$

$$= \text{Var}(rD) \qquad \text{[since } m \text{ is a constant]}$$

$$= r^2\text{Var}(D).$$

Note that r is the range of the data set E. The bounds derived for binary data sets can be extended to these new data sets. In particular,

$$\text{Var}(E) \geq \frac{r^2(n-1)}{n^2} \text{ for data sets with distinct digits } m \text{ and } r+m,$$

and

$$\text{Var}(E) \begin{cases} \leq \frac{r^2}{4} & \text{for even } n \\ \leq \frac{r^2(n^2-1)}{4n^2} & \text{for odd } n. \end{cases}$$

Example. Let $E = \; < -2, -2, -2, 3, 3, 3, 3, 3 >$. Here $n = 8$, $m = -2$ and $r = 5$. We obtain $\text{Var}(E) = \frac{25(5)(3)}{8^2} = \frac{375}{64}$ and note that $\frac{r^2(n-1)}{n^2} = \frac{175}{64} \leq \frac{375}{64} \leq \frac{25}{4} = \frac{r^2}{4}$.