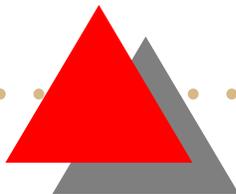




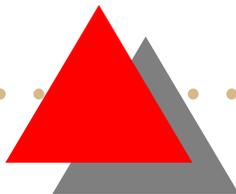
Approximating median in large data vectors

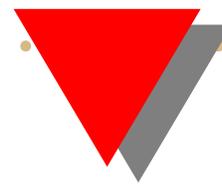
Reza Hosseini

Department of Statistics
University of British Columbia
Vancouver, BC



- 
- Data vector (x_1, \dots, x_n) , n large
 - partition n to m equal vectors of length l
 - Is median of medians a good approximation of the median?
 - Is the median of medians a good approximation if we let m and/or l be large? (Should not depend on n)
 - Good approximation in what sense?
 - Answer: The approximation should be within a reasonable range of quantiles of the data $(1/2 - \epsilon, 1/2 + \epsilon)$.





The median of medians can be bad!

partition number	Partition	Median of the partition
1	$1, 2, \dots, b, b + 1, 10^b, \dots, 10^b$	$b + 1$
2	$1, 2, \dots, b, b + 1, 10^b, \dots, 10^b$	$b + 1$
.		
.		
.		
a	$1, 2, \dots, b, b + 1, 10^b, \dots, 10^b$	$b + 1$
a+1	$1, 2, \dots, b, b + 1, 10^b, \dots, 10^b$	10^b
a+2	$10^b, 10^b, \dots, 10^b$	10^b
.		
.		
.		
2a+1	$10^b, 10^b, \dots, 10^b$	10^b

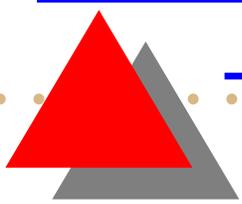
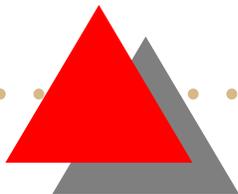


Table 1: The table of data

- 
- Median of medians is not that bad!
 - It is going to be within the range (0.25,0.75)
 - $m = 2a$ and $l = 2b$
 - Let MM be the median of the medians
 - Order the obtained medians of each partition and show them by M_1, \dots, M_m . By definition $MM \geq M_j, j \leq a$.
 - Each M_j is greater than b data points.
 - Hence, MM is greater than ab number of data points
 - $ab/4ab = 0.25$



- 
- How to improve?
 - For each partition take the 1st quartile, median and 3rd quartile
 - The approximation is improved to $(3/8, 5/8) = (0.375, 0.625)$
 - In general take $1/q, 2/q, \dots, q - 1/q$ quantiles then approximation is improved to $(1/2(q/q+1), 1/2(q+2/q+1))$
 - To get an approximation as good as $(0.4, 0.6)$ only need to let $q=4$
 - Note that this does not depend on m, l ($m, l > 2$)
 - We can pick m, l based on our computing abilities

