Analysis of Approximate Median Selection

M. Hofri

*Department of Computer Science, WPI*

Collaborators:

Domenico Cantone & students
*Università di Catania, Dipartimento di Matematica*

Svante Janson
*Department of Mathematics, Uppsala University*
Finding the median efficiently — a difficult problem.

A deterministic algorithm for the exact median was improved in 5/99 by Dor & Zwick, requiring (in the worst case) \( \approx 2.942n \).

Extremely involved . . .

For expected number of comparisons: Floyd & Rivest showed (1975) it can be done in \((1.5 + o(1))n\).

Cunto & Munro (1989): this bound is tight.

Our algorithm was developed in 1998 by Cantone — and only much later we discovered that several formulated various analogues earlier — as early as 1978!

Deterministic, uses at most \(1.5n\) comparisons,

and the expected number is \(4/3n\).

Major virtue: extremely easy to implement

(and understand) — but it only approximates the median.
This is performed in situ.
Essentially the same algorithm can be done “on-line:” processing a stream of values and using work-area of $4\log_3 n$ positions.
Analysis — Cost of search

Finding median of three requires
   2 comparisons in 2 permutations,
   3 comparisons in 4 permutations,
— out of the 6 possible permutations.

Hence $E[C_3] = 8/3$.
The expected total number of comparisons when
looking in a list of size $n$:

$$C_3(n) = \frac{n}{3} \cdot \frac{8}{3} + C_3\left(\frac{n}{3}\right), \quad C_3(1) = 0$$

Result: $C_3(n) = \frac{4}{3}(n - 1)$.

The number of elements that are moved is similarly
$E_3(n) = \frac{1}{3}(n - 1)$.

The number of three-medians computed: $\frac{1}{2}(n - 1)$. 
Analysis — Probabilities of selection

To show that the selected median \( X_n \) is likely to be close to the true median we need to compute the distribution of the rank of the selected entry, \( X_n \).

Let \( n = 3^r \).

The key quantity is \( q_{a,b}^{(r)} \overset{\text{def}}{=} \) the number of permutations, out of the \( n! \) possible ones, in which the entry which is the \( a^{th} \) smallest in the array is:

(i) selected, and

(ii) has rank \( b \) ( = is the \( b^{th} \) smallest) in the next set, that has \( \frac{n}{3} = 3^{r-1} \) entries.

The counting is performed in two steps:

1. Count permutations in which \( a \) is chosen in the \( b^{th} \) triplet, and all the entries chosen in the first \( b - 1 \) triplets are smaller than \( a \), and all the items chosen in the rightmost \( n/3 - b \) triplets are larger that \( a \).

2. Compensate for this restriction: multiply the result of step one by the number of rearrangements of
such permutations: \( \frac{(n/3)!}{(b-1)!((n/3)-b)!} = \frac{n}{3} \binom{n-1}{b-1} \).

The first step is not that simple, and it produces the following expression,

\[
2n(a - 1)!(n - a)!3^{a-b} \sum_i \binom{b - 1}{i} \left( \frac{n}{3} - b \right) \left( \frac{n}{3} - b - a - 2b - i \right) \frac{1}{9^i}.
\]

We find:

\[
q_{a,b}^{(r)} = 2n(a - 1)!(n - a)! \left( \frac{n}{3} - \frac{1}{b-1} \right) 3^{a-b-1} \\
\times \sum_i \binom{b - 1}{i} \left( \frac{n}{3} - br \right) \left( \frac{n}{3} - b - a - 2b - i \right) \frac{1}{9^i}.
\]

The related probability: \( p_{a,b}^{(r)} = q_{a,b}^{(r)}/n! \):

\[
p_{a,b}^{(r)} = \frac{2}{3} \cdot \frac{3^{b-1}}{3-a n^{n-1}_{a-1}} \times \sum_i \binom{b - 1}{i} \left( \frac{n}{3} - b \right) \left( \frac{n}{3} - b - a - 2b - i \right) \frac{1}{9^i} \\
= \frac{2}{3} \cdot \frac{3^{b-1}}{3-a n^{n-1}_{a-1}} \times [z^{a-2b}](1 + \frac{z}{9})^{b-1}(1 + z)^{n/3-b}.
\]
Finally, $P_{a}^{(r)}$: the probability that the algorithm chooses $a$ from an array holding $1, \ldots, n = 3^{r}$.

$$P_{a}^{(r)} = \sum_{b_{r}} P_{a,b_{r}}^{(r)} P_{b_{r}}^{(r-1)} = \sum_{b_{r},b_{r-1},\ldots,b_{3}} P_{a,b_{r}}^{(r)} P_{b_{r},b_{r-1}}^{(r-1)} \cdots P_{b_{3},2}^{(2)}$$

For $2^{j-1} \leq b_{j} \leq 3^{j-1} - 2^{j-1} + 1$.

$$P_{a}^{(r)} = \left(\frac{2}{3}\right)^{r} \frac{3^{a-1}}{\binom{n-1}{a-1}}$$

$$\times \sum_{b_{r},b_{r-1},\ldots,b_{3}} \prod_{j=2}^{r} \sum_{i_{j} \geq 0} \left(\begin{array}{c} b_{j} - 1 \\ i_{j} \end{array}\right) \left(\begin{array}{c} 3^{j-1} - b_{j} \\ b_{j+1} - 2b_{j} - i_{j} \end{array}\right) \frac{1}{9^{i_{j}}}$$

$b_{j} \in [2^{j-1} \ldots 3^{j-1} - 2^{j-1} + 1]$, $b_{2} = 2$ and $b_{r+1} \equiv a$.

No known reduction . . .

Numerical calculations produced:


<table>
<thead>
<tr>
<th>n</th>
<th>$r = \log_3 n$</th>
<th>Avg.</th>
<th>$\sigma_d$</th>
<th>$\sigma_d/n^{2/3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>2</td>
<td>0.428571</td>
<td>0.494872</td>
<td>0.114375</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>1.475971</td>
<td>1.184262</td>
<td>0.131585</td>
</tr>
<tr>
<td>81</td>
<td>4</td>
<td>3.617240</td>
<td>2.782263</td>
<td>0.148619</td>
</tr>
<tr>
<td>243</td>
<td>5</td>
<td>8.096189</td>
<td>6.194667</td>
<td>0.159079</td>
</tr>
<tr>
<td>729</td>
<td>6</td>
<td>17.377167</td>
<td>13.282273</td>
<td>0.163979</td>
</tr>
<tr>
<td>2187</td>
<td>7</td>
<td>36.427027</td>
<td>27.826992</td>
<td>0.165158</td>
</tr>
</tbody>
</table>

Variance ratios for the median selection as function of array size

$d$ is the error of the approximation:

$$d \equiv \left| X_n - \frac{n+1}{2} \right|$$

What can we expect when $n$ grows?
Plot of the median probability distribution for $n=27$
Plot of the median probability distribution for $n=243$
To answer the last question we look at a “similar” situation, where we look at \( n \) independent random variables:

\[
\Xi = (\xi_1, \xi_2, \ldots, \xi_n), \quad \xi_j \sim U(0, 1).
\]

\( \Xi \) is a permutation of their sorted order, \( S(\Xi) \):

\[
S(\Xi) = (\xi_{(1)} \leq \xi_{(2)} \leq \cdots \leq \xi_{(n)}).
\]

Observation:
If the Sicilian algorithm operates on this permutation of \( \mathbb{N}_n \), and returns \( X_n = k \),

then sicking it on \( \Xi \) would return \( Y_n = \xi_{(k)} \).

The idea: \( Y_n \) tracks \( \frac{X_n}{n} \), but—due to the independence of the \( n \) variables \( \xi_i \)—it has a simpler distribution.
How good is the tracking? Condition on the sampled value:

\[ E_S \left[ \left( Y_n - \frac{1}{2} \right) - \left( \frac{X_n - \frac{n+1}{2}}{n} \right) \right]^2 = E_S \left[ Y_n - \frac{X_n - 1/2}{n} \right]^2 \]

\[ = E_k \left[ \xi_{(k)} - \frac{k - 1/2}{n} \right]^2 \lesssim \frac{1}{4n}. \]

And the variance of \(|D_n|/n\) is larger, and decreases more slowly!

We said \(Y_n\) is simpler... How simple is it?

\(F_r(x) \equiv \Pr(Y_n - 1/2 \leq x), \quad -1/2 \leq x \leq 1/2, \quad n = 3^r.\)

\(F_0(x) = x + 1/2.\)

Now we need a recurrence:

\(Y_{3n}\) is the median of 3 independent values \(\sim Y_n\), hence

\(F_{r+1}(x) = \Pr(Y_{3n} \leq x + 1/2) = 3F_r^2(x)(1 - F_r(x)) + F_r^3(x)\)

\[ = 3F_r^2(x) - 2F_r^3(x). \]
A simpler form is obtained by shifting $F_r(\cdot)$ by $1/2$;

$$G_r(x) \equiv F_r(x) - 1/2 \implies G_0(x) = x,$$

We get our first key equation:

$$G_{r+1}(x) = \frac{3}{2}G_r(x) - 2G^3_r(x).$$

But it is not interesting! it is satisfied by

$$G_r(x) = \begin{cases} 
-\frac{1}{2} & x < a \\
0 & x = a \\
\frac{1}{2} & x > 0 
\end{cases}$$

This says: $\frac{D_n}{n} \rightarrow 0,$ $D_n \overset{\text{def}}{=} X_n - \frac{n+1}{2}.$

Need change of scale.

We showed,

$$\mu^{2r}E \left[ (Y_n - \frac{1}{2}) - D_n/n \right]^2 \rightarrow 0 \quad \forall \mu \in [0, \sqrt{3}).$$

Hence we can track $\mu^r(D_n/n)$ with $\mu^r(Y_n - 1/2)$.

We pick a convenient value, $\mu = 3/2$ and show:
Theorem [Svante Janson]

Let \( n = 3^r, \ r \in \mathbb{N}. \)

\( X_n \) — approximate median of random permutation of \( N_n. \)

Then a random variable \( X \) exists, such that

\[
\left( \frac{3}{2} \right)^r \frac{X_n - \frac{n+1}{2}}{n} \to X,
\]

where \( X \) has the distribution \( F(\cdot); \)
with the same shift

\[
F(x) \equiv G(x) + 1/2,
\]

we get the equation

\[
G\left( \frac{3}{2}x \right) = \frac{3}{2}G(x) - 2G^3(x), \quad -\infty < x < \infty
\]

Moreover:

The distribution function \( F(\cdot) \) is strictly increasing throughout.

The value \( 3/2 \) is inherent in the problem!
The proof of the Theorem uses the technical lemma

**Lemma** Let $a \in (0, \infty)$ and $\phi$ that maps $[0,a]$ into $[0,a]$.
For $x > a$ we define $\phi(x) = x$. Assume

(i) $\phi(0) = 0$
(ii) $\phi(a) = a$
(iii) $\phi(x) > x$, for all $x \in (0,a)$.
(iv) $\phi'(0) = \mu > 1$, and continuous there;

$\phi(\cdot)$ is continuous and strictly increasing on $[0,a)$.
(v) $\phi(x) < \mu x$, $x \in (0,a)$.

Let $\phi_r(t) = \phi(\phi_{r-1}(t))$, the $r$th iterate of $\phi(\cdot)$.
Then

$$\text{as } r \to \infty, \quad \phi_r(x/\mu^r) \to \psi(x), \quad x \geq 0.$$ 

$\psi(x)$ is well defined, strictly monotonic increasing for all $x$, increases from 0 to $a$, and satisfies the equation $\psi(\mu x) = \phi(\psi(x))$.

**Proof:**

From Property (v):

$$\phi(x/\mu^{r+1}) < x/\mu^r,$$

Since iteration preserves monotonicity,

$$\phi_{r+1}(x/\mu^{r+1}) = \phi_r(\phi(x/\mu^{r+1})) < \phi_r(x/\mu^r).$$

Hence a limit $\psi(\cdot)$ exists.
The properties of $\psi(x)$ depend on the behavior of $\phi(\cdot)$ near $x = 0$. Since $\phi'(x)$ is continuous at $x = 0$, $\psi(\cdot)$ is continuous throughout. Since it is bounded, the convergence is uniform on $[0, \infty]$. Hence, since $\phi(\cdot)$ and all its iterates are strictly monotonic, so is $\psi(\cdot)$ itself.

We have then the equation

$$G\left(\frac{3}{2}x\right) = \frac{3}{2}G(x) - 2G^3(x), \quad -\infty < x < \infty$$

but we have no explicit solution for it.

What can we do?

Several things.
We can calculate a power expansion for it; From $G_0(\cdot)$ and the iteration, all $G_r(\cdot)$ are odd, hence we can write

$$G(x) = \sum_{k \geq 1} b_k x^{2k-1}.$$  

$b_1$ is available from the iteration: The derivatives of $G_r(x/\mu^r)$ are all 1, hence this is also the derivative there of $G(x)$.

Successive calculations are easy:
The fit of $F(\cdot)$—calculated using the first 150 $b_k$—to the distribution of $X_n/n$ is poor for $n$ in the low hundreds but improves very fast.

We show an example later.
Fact: it is very close to Normal, with mean zero and
\[ \sigma = \frac{1}{\sqrt{2\pi}} \] — but not quite!

We can investigate how similar it is by looking at the
tail of the distribution — the complementary function
\( g(x) \equiv 1 - F(x) \).
It satisfies
\[
g(x\mu) = 3g^2(x) - 2g^3(x) \implies 3g(x\mu) = (3g(x))^2 \left(1 - \frac{2}{3}g(x)\right).
\]

since \( 0 < g(x) < 1 \):
\[
\frac{1}{3} (3g(x))^2 < 3g(x\mu) < (3g(x))^2
\]
This is all we need in order to show that the tail of
the distribution of \( X_n/n \) is
\[
e^{-dt^v} < g(t) < \frac{1}{3} e^{-ct^v} \quad c \approx 3.8788, \quad d \approx 4.9774 \quad v \approx 1.70951
\]
\[
c = \ln(1 - F(1)); \quad v = \ln 2 / \ln(3/2), \quad d = c + \ln 3,
\]
whereas the Normal distribution decays much faster:
itself is
\[
1 - \Phi(x) \sim e^{-0.5x^2} / (2\pi x).
\]
Example: From simulation, at $n = 1000$, 95% of the values of $D_{1000}$ fell in the interval $[-58, 58]$.

From the “tracking claim” we have $D_n \sim nX/\mu_r$.

Also $n/\mu_r = 3^r/(3/2)^r = 2^r = n^{\ln(2)/\ln(3)} \approx n^{0.63092975}$.

And then

\[
\Pr[|D_n| \leq d] \approx \Pr[|X| \leq d\mu_r/n] \\
\implies \Pr[|D_{1000}| \leq 58] \approx \Pr[|X| \leq 0.7424013] \approx 0.934543.
\]

This was calculated using the power series development.

When using the upper bound on the tail we similarly find

\[
\Pr[|D_{1000}| \leq 58] = 1 - 2g(0.7424013) \\
\approx 1 - \frac{2}{3} \exp\left(-3.878797 \times 0.7424^{1.7095113}\right) \approx 0.935205.
\]
Open problems:

1. A better characterization of the solution for $G(x)$.

2. An explicit value for the variance of the limiting distribution; From the relation $\mu'(Y_n - 1/2) \xrightarrow{d} X$ we can numerically iterate the transformation and find that it is about 2–3% larger than $1/2\pi$, but an exact value is not easy.

3. A much taller order: compute the quality of a derivative algorithm, that produces an approximate fractile $X_k/n$ for any $1 \leq k \leq n$.

This can be done by filtering the initial values: for example, by picking the 23rd from each set of 28 initial values, and then finding their median, we approximates the fractile $X_{0.8n}$ of the original data.