LINEAR SEARCH VERSUS BINARY SEARCH: A STATISTICAL COMPARISON FOR BINOMIAL INPUTS

Anchala Kumari[1], Rama Tripathi[2], Mita Pal[3] and Soubhik Chakraborty[4*]

[1]University Department of Statistics, Patna University, Patna-800005, India
[2]Department of Information Technology, BIT Mesra, Ranchi-835215, India
[3, 4]Department of Applied Mathematics, BIT Mesra, Ranchi-835215, India

*corresponding author's email: soubhikc@yahoo.co.in (S. Chakraborty)

ABSTRACT

For certain algorithms such as sorting and searching, the parameters of the input probability distribution, in addition to the size of the input, have been found to influence the complexity of the underlying algorithm. The present paper makes a statistical comparative study on parameterized complexity between linear and binary search algorithms for binomial inputs.

KEY WORDS

Linear search; binary search; parameterized complexity; statistics; factorial experiments

1 INTRODUCTION

Two of the popular search algorithms are linear search and binary search. While linear search (also called sequential search) scans each array element sequentially, a binary search in contrast is a *dichotomic divide and conquer search algorithm*. For an extensive literature on searching, see Knuth [1].

In the present paper we investigate the effect of parameters n and p of a Binomial distribution input on the number of comparisons in linear search and binary search. Using factorial experiment, it is observed that both the main effects n and p and the interaction effects n*p are highly significant for linear and significant but comparatively less for binary search. The result clearly suggests that apart from the size of the input, the parameters of the input distribution need also be taken into account to explain the behavior of certain algorithms. In an earlier work on parameterized complexity, Anchala and Chakraborty [2] used factorial experiment to explain software complexity for insertion sort. The same authors have used factorial experiments with a

DOI : 10.5121/ijcsea.2012.2203

response surface design in [3] to examine the nature of the fast and popular quicksort. The first researchers to work on parameterized complexity are Downey and Fellows [4].

2. Experimental results

2.1 Results using factorial experiment

Binomial variates are independently filled in an array of size k = 2000 (fixed) and we make a linear search for an element which is in the array. We are interested in finding the number of comparisons expected to ascertain that the searched element is present. To ensure the searched element is indeed available in the array, one array index was randomly selected and the key of this index is the searched element. The code is omitted.

Binomial distribution (definition): Let X be a Binomial variate with parameters n and p. The probability $P(X = x) = {}^{n}C_{x}\, p^{x}(1-p)^{n-x}$ where x=0, 1, 2, …n and $0<p<1$. Binomial distribution has three assumptions:

1. Trials are independent.
2. Each trial can result in one of two possible outcomes which we call "success" and "failure" respectively.
3. The probability of success p is fixed in each trial. The expression for P(X=x) gives the probability of getting x successes in n trials made under the three above-mentioned assumptions. The distribution is so called as the expression for P(X=x) is the general term, i.e., (x+1)-th term in the Binomial expansion of $(q+p)^n$, q=1-p is the probability of failure in each trial. Since we assume p as fixed, q is fixed as well. Further literature on Binomial distribution can be found in Gupta and Kapoor [5].

To study the main effects n and p as well as the interaction effects n*p of the parameters n and p of Binomial (n, p) distribution input on the number of comparisons, a 3^2 factorial experiment was conducted with two factors n and p each at three levels (3000, 6000, 9000 for n and 0.2, 0.5 and 0.8 for p). Table 1 gives the data for the desired factorial experiment (n is written as N and p as P; this is what MINITAB will print) for linear search while table 2 gives the same for binary search. Tables 3 and 4 give the ANOVA tables depicting the results of factorial experiment on linear search and binary search respectively.

2.2 Other experimental results

Tables 5-8 and figures 1-8 summarize our other experimental results. These results were obtained for fixed array size k = 2000.

3. Discussion

It can be theoretically argued that the parameters of the Binomial distribution, in addition to the array size k (here fixed at 2000), will affect the number of comparisons in linear search (the same for binary search is under investigation). Since the searched element can be present in more than one place, we suppose the first time it comes is in position r, r=1, 2…k. Then we must have that the first r-1 comparisons did not yield the searched element and that the r-th comparison yielded it.

Evidently, this probability is $P(r) = C\{1-f(y, n, p)\}^{r-1}f(y, n, p)$, r=1, 2…k………(1)
where y is the searched element.

This is because the k array elements are independently filled with Binomial (n, p) variates, so that the probability of any array element to be y is P(X=y)=f (say) and not to be y is 1-f. C is a normalization factor to ensure ΣP(r) =1, r = 1, 2…k It can be shown that the expected number of comparisons
= E(r) = Σ rP(r), the summation over r is from1 to k, = C f S where C = 1/ [1- P(0)-P(k+1)-P(k+2)……]

$$f = {}^{n}C_{y}\, p^{x}(1-p)^{n-y}$$

and, $$S = [1- \{1- f\}^{k}]/ f^{2} - k\{1- f\}^{k}/ f$$

Remark: The random variable r follows a *doubly truncated Geometric distribution* since r cannot take the value 0, nor can it take a value higher than k.

The expression for the expected number of comparisons, however, does not establish the significance of the interaction effect n*p and hence we resorted to factorial experiments. Our results confirm the interaction effect, besides the main effects, is highly significant. Further, figures 1-8 suggest an O(n) complexity for fixed p and k and O(p) complexity for fixed n and k.

4. Conclusion

Using 3^2 factorial experiment, it is observed that not only the main effects n and p but even the interaction effects n*p are highly significant in influencing the number of comparisons in linear search for Binomial (n, p) input. However, it is also observed that the main effects n, p and the interaction effects n*p are comparatively less significant in influencing the number of comparisons in binary search for Binomial (n, p) input.. Moreover, the mean comparisons seems to depend linearly on n and p for fixed k. Interestingly, this is true for both linear and binary search. The results clearly suggest why, apart from the size of the input, the parameters of the input distribution need also be taken into account to explain the behavior of certain algorithms. The role of factorial experiments is firmly established in parameterized complexity analysis in such algorithms.

To the question which algorithms are better suited for such studies in parameterized complexity, the answer is that those in which fixing the input parameter characterizing the array size (k in our case) does not fix all the computing operations. The sorting and searching algorithms fall into this category. Future work includes similar interesting case studies.

TABLES AND FIGURES

Table 1: Mean number of comparisons in linear search for fixed array size (2000) and varying N and P.

First set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	119.92	515.02	967.79
0.5	112.47	581.42	1185.53
0.8	108.03	535.30	1022.51

Second set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	118.03	516.09	967.95
0.5	112.09	582.50	1184.36
0.8	107.56	533.40	1020.31

Third set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	119.06	514.86	968.03
0.5	111.43	581.89	1186.46
0.8	108.98	536.20	1021.56

Table 2: Mean number of comparisons in binary search for fixed array size (2000) and varying N and P.

First set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	1001.33	4019.86	7059.48
0.5	994.98	3934.99	6934.33
0.8	1034.68	4086.44	7029.36

Second set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	975.69	4046.29	7054.04
0.5	998.20	3966.21	6952.95
0.8	982.34	4002.58	7058.36

Third set of observations

P	**N=3000**	**N=6000**	**N=9000**
0.2	1014.41	4032.38	7019.08

0.5	1005.05	3966.83	6989.9
0.8	988.91	3991.40	7039.2

Table 3: Analysis of Variance for y, using Adjusted SS for Tests (linear search)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
N	2	4030817	4030817	2015409	2659942.28	0.000
p	2	42272	42272	21136	27895.62	0.000
N*p	4	42014	42014	10504	13862.63	0.000
Error	18	14	14	1		
Total	26	4115118				

S = 0.870453   R-Sq = 100.00%   R-Sq(adj) = 100.00%

MINITAB version 15 was used to yield the results of the factorial experiments of Linear search:-

Multilevel Factorial Design

Factors: 2 Replicates: 3
Base runs: 9 Total runs: 27
Base blocks: 1 Total blocks: 1

Number of levels: 3, 3

General Linear Model: y versus N, p

Factor	Type	Levels	Values
N	fixed	3	1, 2, 3
p	fixed	3	1, 2, 3

Table 4: Analysis of Variance for y, using Adjusted SS for Tests (binary search)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
N	2	162847793	162847793	81423897	123575.44	0.000
P	2	16681	16681	8340	12.66	0.000
N*P	4	8489	8489	2122	3.22	0.037
Error	18	11860	11860	659		
Total	26	162884823				

```
S = 25.6691   R-Sq = 99.99%   R-Sq(adj) = 99.99%
```

MINITAB version 15 was used to yield the results of the factorial experiments of binary search:-

Multilevel Factorial Design

```
Factors:       2     Replicates:     3
Base runs:     9     Total runs:    27
Base blocks:   1     Total blocks:   1
Number of levels: 3, 3
```

General Linear Model: y versus N, P

```
Factor  Type   Levels  Values
N       fixed     3    1, 2, 3

P       fixed     3    1, 2, 3
```

Table 5: Mean and SD of no. of comparisons for fixed p=0.2 and N varying from 3000 to 9000

LINEAR SEARCH BINARY SEARCH

	Linear			Binary	
N	MEAN	SD		MEAN	SD
3000	95.42429	5.099625		971.1229	21.92183
4000	211.8057	8.217798		1928.013	42.26884
5000	340.4814	10.77158		2893.254	76.00657
6000	483.5814	15.46534		3880.792	1111.529
7000	632.7686	19.9731		4879.703	148.2382
8000	820.7614	25.74744		5854.994	185.6159
9000	971.8386	31.84307		6849.33	222.3924

Table 6: Mean and SD of no. of comparisons for fixed p=0.5 and N varying from 3000 to 9000

LINEAR SEARCH BINARY SEARCH

	Linear			Binary	
N	MEAN	SD		MEAN	SD
3000	130.1014	7.00651		972.08	21.59409
4000	264.1129	8.887761		1980.031	42.77784
5000	415.3614	13.31927		2928.61	77.99048
6000	588.0328	18.06139		3899.347	112.9273
7000	756.0186	24.04293		4844.13	148.9676
8000	924.8	26.66672		5850.78	185.8966
9000	1121.763	36.24661		6819.733	222.2228

Table 7: Mean and SD of no. of comparisons for fixed p=0.8 and N varying from 3000 to 9000

LINEAR SEARCH BINARY SEARCH

	Linear			Binary	
N	MEAN	SD		MEAN	SD
3000	115.6329	6.243703		1018.141	21.48802
4000	245.1214	8.52181		1985.369	44.04666
5000	393.5871	12.7249		2975.52	78.13666
6000	550.3929	17.45485		3982.537	114.5793
7000	704.8671	22.51765		4990.81	152.1838
8000	856.3285	27.90663		5981.506	189.8659
9000	1041.273	33.69814		7011.628	227.1954

Table 8: Mean and SD of no. of comparisons for fixed N=10,000 and p varying from 0.1 to 0.9

LINEAR SEARCH BINARY SEARCH

	Linear			Binary	
P	MEAN	SD		MEAN	SD
0.1	146.4986	8.256072		1007.9	21.96101
0.2	312.5357	9.884373		2030.541	43.94901
0.3	488.8714	14.8072		3028.646	79.62766
0.4	683.6443	20.83033		4020.094	116.2303
0.5	887.1243	27.49452		5019.079	153.6259
0.6	1076.687	34.59078		6050.274	190.9012
0.7	1289.791	42.12352		7076.607	229.6028
0.8	1460.953	49.30895		8077.373	268.2914
0.9	1601.191	55.57834		9095.618	305.9862

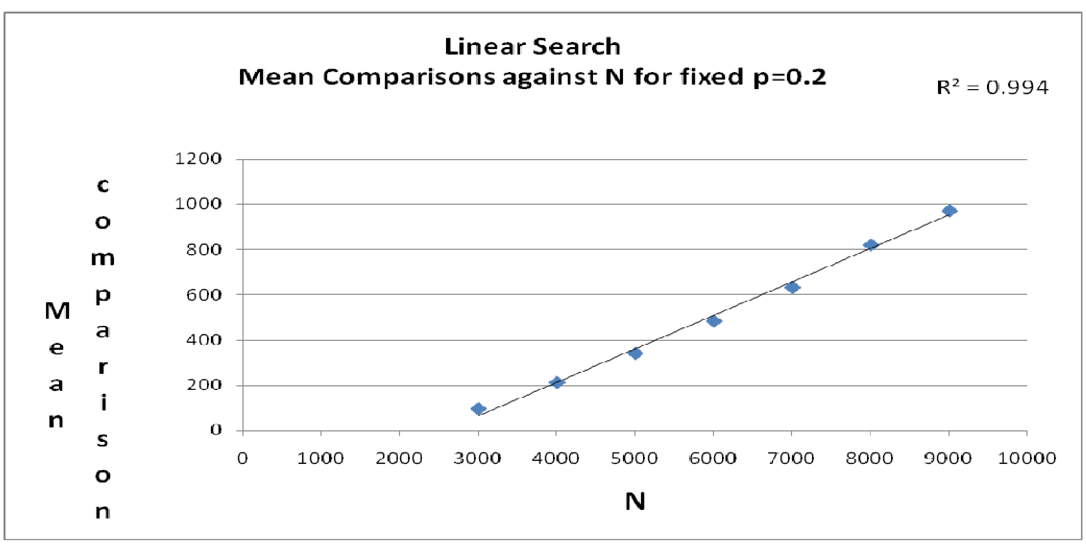

Fig. 1 Plot between N and Mean Comparisons for linear search (p=0.2)

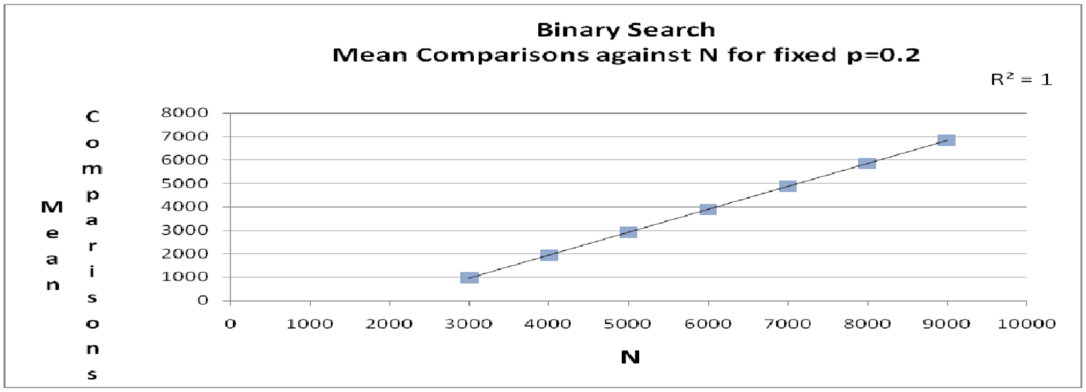

Fig. 2 Plot between N and Mean Comparisons for Binary Search (p=0.2)

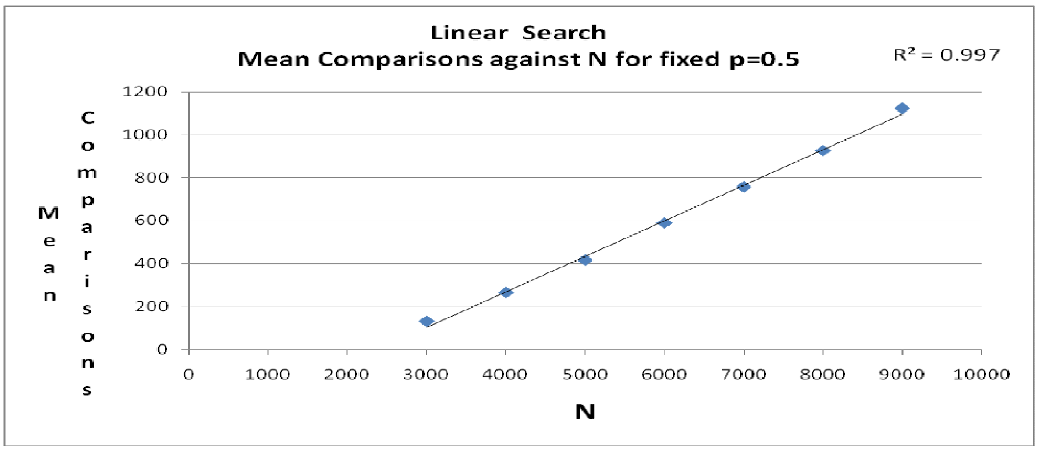

Fig. 3 Plot between N and mean comparisons for linear search (p=0.5)

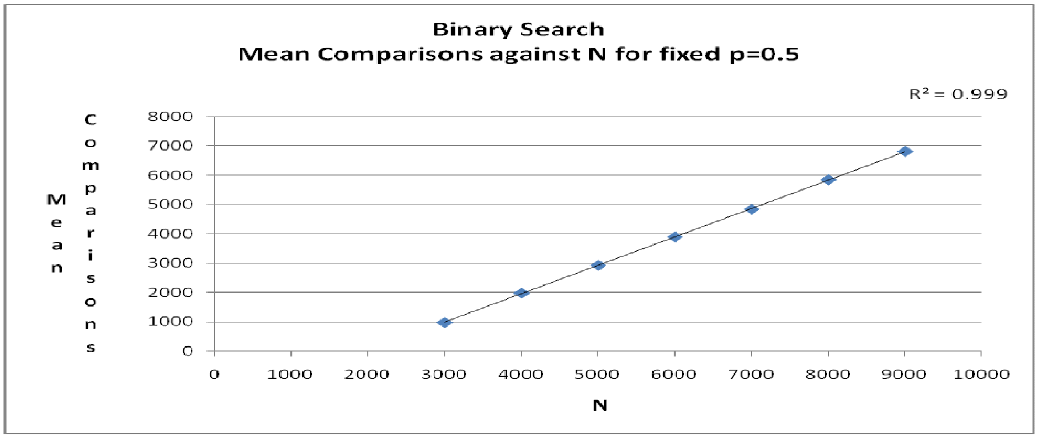

Fig. 4: Plot between N and mean comparisons for binary search (p=0.5)

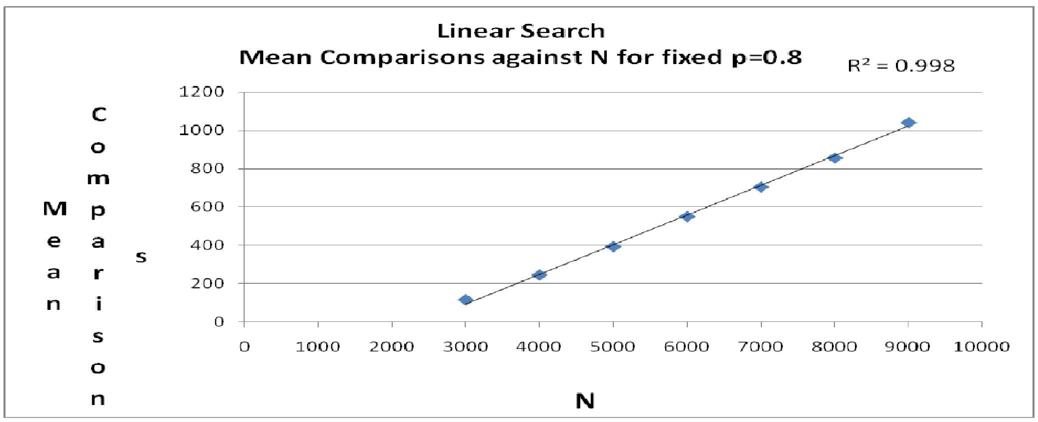

Fig. 5: Plot between N and mean comparisons for linear search (p=0.8)

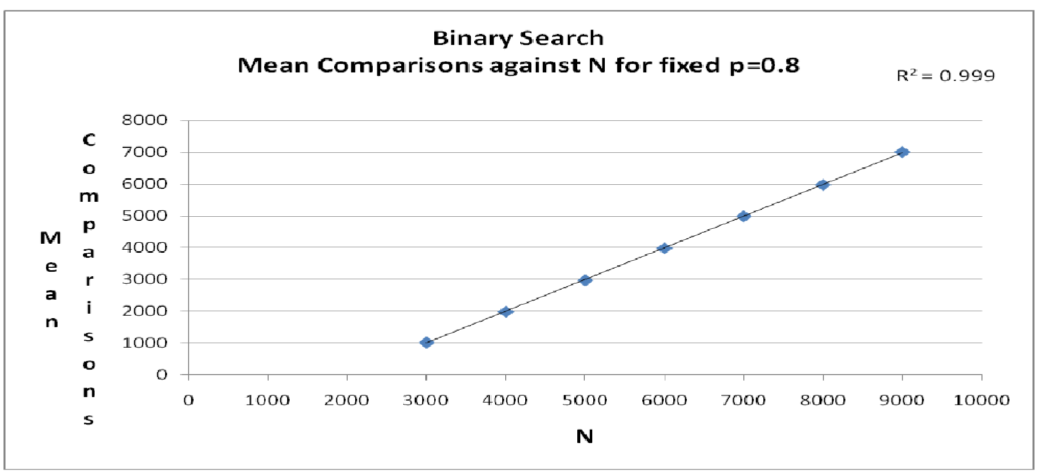

Fig. 6 Plot between N and mean comparisons for binary search (p=0.8)

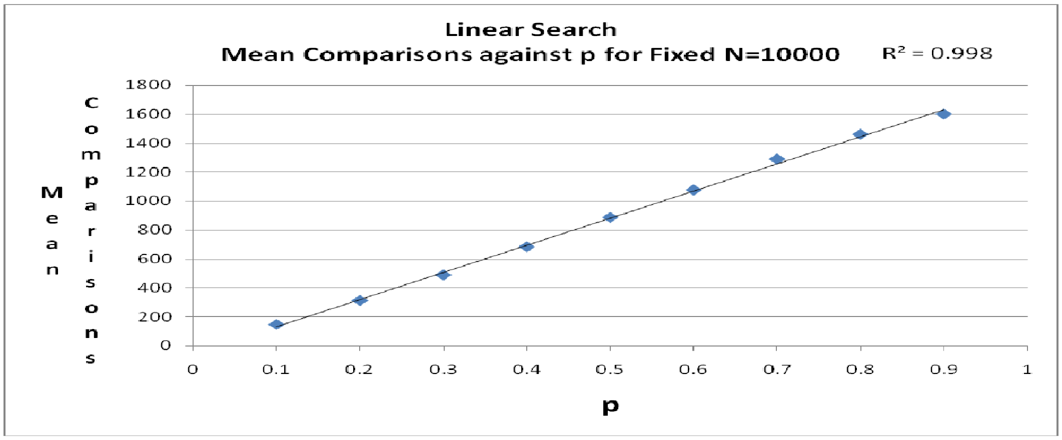

Fig. 7 Plot between p and mean comparisons for linear search (N=10,000)

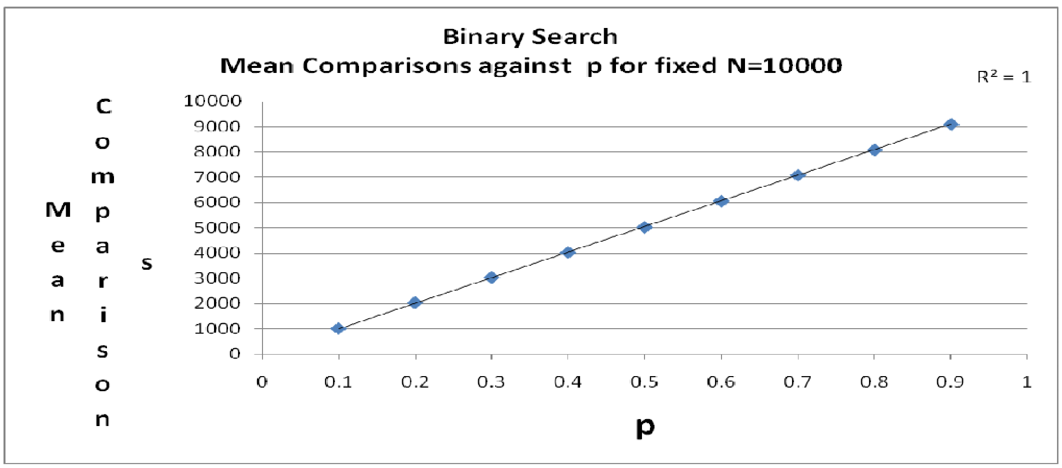

Fig. 8 Plot Between p and corresponding mean for binary search (N=10,000)

References

[1] Knuth, D. E. The Art of Computer Programming, Vol. 3: Sorting and Searching, Addison Wesley, (1997), 3rd ed., 396-408

[2] Kumari , A. and Chakraborty , S., Software Complexity: A Statistical Case Study Through Insertion Sort, Applied Math. and Compu., vol. 190(1), (2007), p. 40-50

[3] Kumari, A. and Chakraborty, S., A Simulation Study on Quick Sort Parameterized Complexity Using Response Surface Design, International Journal of Mathematical Modeling, Simulation and Applications, Vol. 1, No. 4, (2008), 448-458

[4] Downey, Rod G.; Fellows, Michael R. (1999). Parameterized Complexity. Springer.

[5] S. C. Gupta and V. K. Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand and Sons, New Delhi, reprint 2010