First some probability basics
then expected run time of Quicksort.

Geometric random variable:
Flip a coin that's heads with probability p
& tails with prob. $1-p$.

Let X = # of flips until a heads occurs.
(including flip with the heads)

$X \sim \text{Geometric}(p)$.

Let $\mu = E[X]$
What's μ?
Should be $\mu = E[X] = 1/p$.
Why?

$$E[X] = \sum_{j=1}^{\infty} j(1-p)^{j-1} p$$

can try to simplify...
or easier approach:

Look at the 1st flip, (+1 flip)
if it's heads we're done (0 more flips)
if not we repeat the experiment again
(μ more flips)

Hence,

$$\mu = 1 + 0 \times p + \mu \times (1-p)$$

$$\mu = 1 + \mu - \mu p$$

$$\mu p = 1$$

$$\mu = 1/p.$$

$\boxtimes$

Coupon Collector's problem:

There's an urn with n distinct coupons.
In every step we choose a random coupon.
Then we put it back in & repeat.
How many steps until we see every coupon at least once?

Let $X = \#$ of steps in total.

$X_1 = \#$ of ~~boxes~~ steps to get the 1st coupon

$X_2 = \#$ of steps after seeing the 1st till we get the 2nd one.

& $X_j = \#$ of steps after seeing the $(j-1)^{st}$ coupon till we see the j^{th} coupon.

Thus,
$$X = \sum_{j=1}^{n} X_j$$

and
$$E[X] = E\left[\sum_{j=1}^{n} X_j\right] = \sum_{j=1}^{n} E[X_j].$$

Let $\mu_j = E[X_j]$. What's μ_j?

We've seen $(j-1)$ coupons & if we see any of $n-(j-1)$ then we've got a new one.

Prob. of seeing a new one in one step is

$$P_j = \frac{n-(j-1)}{n} = 1 - \frac{j-1}{n}$$

This X_j is a geometric random variable, hence

$$\mu_j = \frac{1}{P_j} = \frac{n}{n-(j-1)} = \frac{n}{n-j+1}$$

Therefore,

$$\begin{aligned} E[X] &= \sum_{j=1}^{n} \mu_j \\ &= \sum_{j=1}^{n} \frac{n}{n-j+1} \\ &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} \\ &= n \sum_{j=1}^{n} \frac{1}{j} \end{aligned}$$

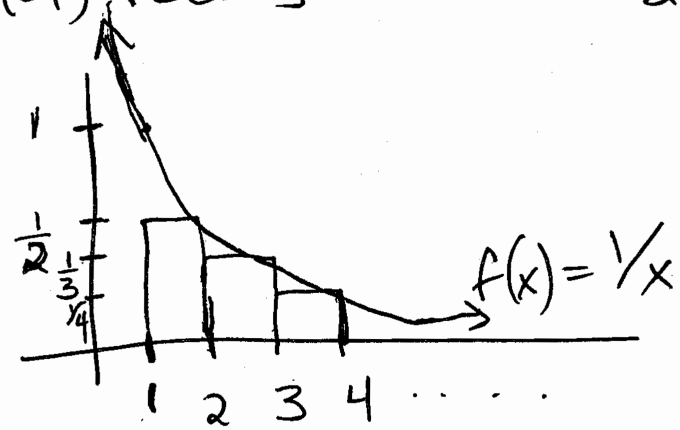
We'll see that: $\sum_{j=1}^{n} \frac{1}{j} \leq 1 + \ln n$

Hence, $E[X] = n \ln n + \Theta(n)$.

To bound $\sum_{j=1}^{n} \frac{1}{j}$:

look at $\sum_{j=2}^{n} \frac{1}{j} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$

draw $(n-1)$ rectangles of area $\frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$:



Notice that:

$$\sum_{j=2}^{n} \frac{1}{j} \leq \int_{x=1}^{n} \frac{1}{x}\,dx$$

and $$\int_{x=1}^{n} \frac{1}{x}\,dx = \ln x \Big|_{x=1}^{n} = \ln n$$

Thus, $$\sum_{j=2}^{n} \frac{1}{j} \leq \ln n$$

and $$\sum_{j=1}^{n} \frac{1}{j} \leq 1 + \ln n.$$

QuickSort:

input: array $A=[a_1, \ldots, a_n]$ of n numbers

output: A in sorted order

if $n=1$, return(A)

Choose an element of A as a pivot p — How?

Partition A into $A_{<p}, A_{=p}, A_{>p}$

Recursively sort $A_{<p}$ & $A_{>p}$

Return$(A_{<p}, A_{=p}, A_{>p})$

In the worst case, Quicksort takes $\Omega(n^2)$ time.

If p was the median, then we get running time:

$$T(n) = 2T\left(\frac{n}{2}\right) + O(n)$$
$$= O(n \log n).$$

What if p is chosen at random from A?

This is <u>Randomized QuickSort</u>.

We'll analyze the <u>expected</u> run time of randomized QuickSort.

Look at # of comparisons:

a comparison takes a pair a_i, a_j and checks whether $a_i < a_j$ or $a_i = a_j$ or $a_i > a_j$?

Lemma: for randomized QuickSort, the expected # of comparisons is $\leq 2n \ln n$.

Proof:

Say the initial input is $A = [a_1, a_2, \ldots, a_n]$ & let the sorted version of A be

$$s_1 \leq s_2 \leq \cdots \leq s_n$$

Note: a pair s_i & s_j are compared at most once.

Why? The first time they are compared either s_i or s_j are the pivot P at that step.

Say $p = s_i$. Then s_i is put in $A_{=p}$ & s_j is put in $A_{<p}$ or $A_{>p}$.

So they are not in the same subproblem again.

For $i<j$,

$$\text{let } X_{ij} = \begin{cases} 1 & \text{if } s_i \,\&\, s_j \text{ are compared by the algorithm} \\ 0 & \text{if } s_i \,\&\, s_j \text{ are not compared} \end{cases}$$

Let X = total # of comparisons

Then, $$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$$

Our goal is to compute $E[X]$.

$$E[X] = E\left[\sum_i \sum_j X_{ij}\right] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[X_{ij}]$$

To compute $E[X_{ij}]$:

$$E[X_{ij}] = 1 \times \Pr(X_{ij}=1) + 0 \times \Pr(X_{ij}=0)$$
$$= \Pr(X_{ij}=1)$$

When is $X_{ij}=1$?

Need that s_i or s_j is selected as a pivot before s_i & s_j are put in different subproblems — to get in diff. subproblems we need a pivot that splits them

So we need a s_ℓ where

$$s_i \leq s_\ell \leq s_j \quad \cancel{\text{or } s_j \leq s_\ell \leq s_i}.$$

(we know $s_i \leq s_j$ since s is sorted)

Look at the set

$$s_i, s_{i+1}, \ldots, s_{j-1}, s_j$$

Which of these $(j-i+1)$ numbers is the first one selected as a pivot?

If it's s_i or s_j then $X_{ij}=1$

If not then s_i & s_j are split so $X_{ij}=0$.

One of these is first, so

$$\Pr(X_{ij}=1) = \frac{2}{j-i+1} = E[X_{ij}].$$

Thus,

$$E[X] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1}$$

$$= 2 \sum_{i=1}^{n} \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n-i+1} \right)$$

$$\leq 2 \sum_{i=1}^{n} \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \right)$$

$$\leq 2 \ln n \qquad \text{since} \quad \sum_{j=2}^{n} \frac{1}{j} \leq \ln n$$

Therefore,

$$E[X] \leq 2 \ln n.$$