Math 30530, spring 2019

The secretary problem

Notes by David Galvin, last updated February 5, 2019

Abstract

The "secretary problem" asks: how can a company maximize the probability of hiring the best candidate for a job, if the candidates are interviewed in random order, and if after each interview the company has to make an irrevocable "hire or not" decision? It turns out that there is a simple strategy that gives the company about a $1/e \approx 0.368$ chance of getting the best person.

The problem can also be framed in terms of dating: if you date serially, and can't ever return to a previous partner, how do you maximize the probability of settling down with the best possible partner? Again, there's a strategy that gives you about a 37% chance.

Remember the following problem from Homework 1:

> Four envelopes contain four different amounts of money. You are allowed to open them one by one, each time deciding whether to keep the amount or discard it and open another envelope. Once an amount is discarded, you are not allowed to go back and get it later. Compute the probability that you get the largest amount under the following different strategies:
>
> 1. You take the first envelope.
> 2. You open the first envelope, note that it contains the amount x, discard it and take the next amount which is larger than x (if no such amount shows up, you must take the last envelope).
> 3. You open the first two envelopes, call the amounts x and y, and discard both and take the next amount that is larger than both x and y.

This is an instance of a famous problem, called the *secretary problem*. Here's how the problem goes: a company is interviewing for an open position. There are n applicants. HR decides to interview the candidates one after another. After each new candidate is interviewed, (s)he is ranked relative to all the previous interviewees. (So, for example, after 5 of 20 candidates have been interviewed, the company has a ranking of those 5, but no information about where those 5 stack up relative to the remaining 15.) Once this ranking is done, the company can choose to either:

- hire the person just interviewed, in which case the process stops, or

- not hire the person just interviewed, and interview the next person. In this case the process continues, BUT (and here's the key rule in the process) the company can never go back to the person just interviewed (or, by extension, to any person interviewed earlier), no matter what they see in subsequent interviews.

If, by the time the last person comes to be interviewed and no-one has been hired yet, then the nth person must be hired.

The company's goal, is, of course,

to hire the best of the n candidates.

This goal can't be achieved with certainty: maybe the best candidate is interviewed first, but the company doesn't realize how good (s)he is until after rejecting him/her, and going on to the second candidate. So the real goal of the company is

to come up with a strategy that maximizes the probability of hiring the best person.

The randomness in this problem, that makes it a probability problem, is that when the company decides on the order in which it is going to interview the n candidates, it is essentially putting a random order, from among $n!$ possible orders, on n numbers — the n "quality scores" of the n candidates. As the interviews progress, the quality scores are revealed one by one, allowing the company to rank the candidates seen so far; but until all the candidates are interviewed, the company has no idea what the full set of quality scores are, so has no idea which is the best candidate.

Notice that this is *exactly* the same as the envelopes problem from the homework. The four envelopes are the four candidates; the (unknown until observed) amount of money in each envelope is the quality score of each candidate; and the rules

- each time decide whether to keep the amount revealed in an envelope, or discard it and open another envelope, and

- once an amount is discarded, you are not allowed to go back and get it later

correspond exactly to the interview rules imposed by HR.[1]

The homework problem suggest one line of strategies: in advance of the process, fix a number k. Interview the first k people, and commit to discarding them. Then keep interviewing people, until the first time that you come to someone who is better than all of the first k, and hire that person. If you come to the end of the process without ever

[1]The secretary problem can also be framed in terms of relationship choices. Suppose you know that in your early adulthood, you have the potential for n relatively serious dating partners. You can imagine these n partners coming at you in random order, one after the other (I'm assuming you are a serial dater, not a multi-dater). As you are with each partner, you have to make a choice: is (s)he the one, or should you move on? If (s)he's the one, you've found your life-partner; if not, you move on to the next relatively serious dating partner, and, whether they are better or worse than previous partners, you can't go back to any previous one (that would be awkward). If your goal is to maximize the probability of ending up with the best possible partner, this is exactly the secretary/envelope problem.

seeing anyone who is better than all of the the first k, hire the last person. (In the envelope problem, you were asked to evaluate the probability of hiring the best person under this strategy, when $n = 4$, for each of the values $k = 0, 1, 2$. It turned out that $k = 1$ gave the best probability, 11/24).

It is possible to systematically analyze all of these strategies, for general n and k. For convenience, denote by

$$p(n, k)$$

the probability that the best person is selected under the "discard the first k" strategy.

$k = 0$ Here, you discard no-one, so the first person you interview will be better than everyone you discarded, and you will hire him/her. There are $(n - 1)!$ orderings of the n candidates in which the first one is the best, so the probability under this strategy that the best person is hired is $(n - 1)!/n!$. That is,

$$p(n, 0) = \frac{1}{n}.$$

$k \geq 0$ In the rest of the discussion, we imagine that the candidates have names "1" through "n", with "'1" the weakest candidate, and "n" the strongest candidate. Suppose you interview k people, for some k between 1 and $n - 1$, and discard them all. Our goal is to count how many of the $n!$ orderings of the candidates does this result in hiring the best candidate, that is, candidate "n"? Well, the first thing that has to happen is that best candidate, "n", *cannot* be among the first k interviewed (otherwise, you definitely don't hire "n" in the end). So, among the first k candidates that you see, the best of those k (i.e., the one with the highest "name") must have a "name" somewhere between "k" (the smallest possible "name" that is highest among a set of k) and "$(n - 1)$".

Suppose that among the first k candidates that you see, the best of those k is named "j", for some j between k and $n-1$. Here is a way to count the number of orderings that result in hiring the best candidate in this case, that uses the multiplication principle:

- first, choose a location in the first k interview positions for "j" to go — k options;
- next, candidates named "$j + 1$" through "n" must appear in the last $n - k$ spots (otherwise, "j" would not be the maximum in the first k), so choose $n-j$ locations in the last $n - k$ interview positions for these candidates — $\binom{n-k}{n-j}$ options;
- next, within the $n-j$ positions identified for "$j+1$" through "n", "n" must appear in the first of those positions; because the interview rule is exactly that the first person seen among "$j + 1$" through "n" is the one to get hired; so place "n" in that position, and arrange the rest any way you want — $(n - j - 1)!$ options; and
- finally, there are $j - 1$ candidates as-yet unplaced (those named "1" through "$j-1$"), and they can be placed anywhere in the remaining unfilled $j-1$ interview positions — $(j - 1)!$ options.

By the multiplication principle, the number of orderings that result in hiring the best candidate in the case where the best of the first k interviewed is named "j", for some

j between k and $n-1$, is

$$k\binom{n-k}{n-j}(n-j-1)!(j-1)!$$

and so the the number of orderings that result in hiring the best candidate, under the "discard the first k" rule, is

$$\sum_{j=k}^{n-1} k\binom{n-k}{n-j}(n-j-1)!(j-1)!.$$

Finally, this leads to a precise expression for the probability that the best candidate is selected, under this strategy:

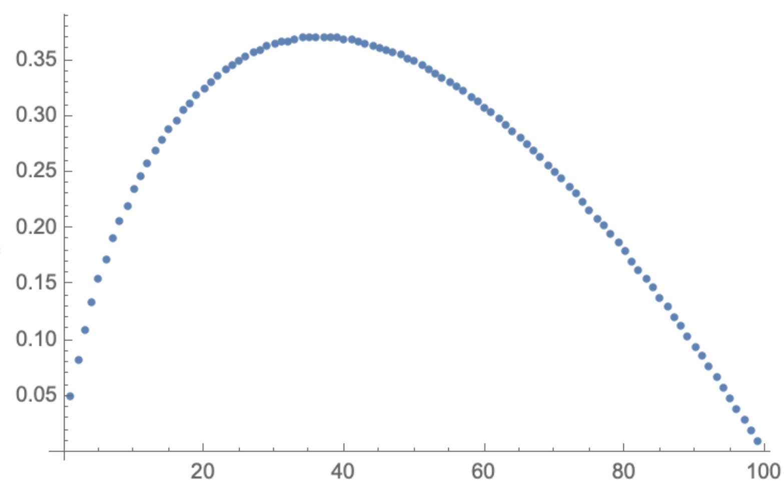
$$p(n,k) = \frac{1}{n!}\sum_{j=k}^{n-1} k\binom{n-k}{n-j}(n-j-1)!(j-1)!.$$

(Check, when $n = 4$ and $k = 1$ and 2, that this agrees with the solution to the envelope problem!)

The expression for $p(n,k)$ looks unwieldy, but in fact it is not; using the factorial formula for the binomial coefficients, and a little algebra, it can be simplified to

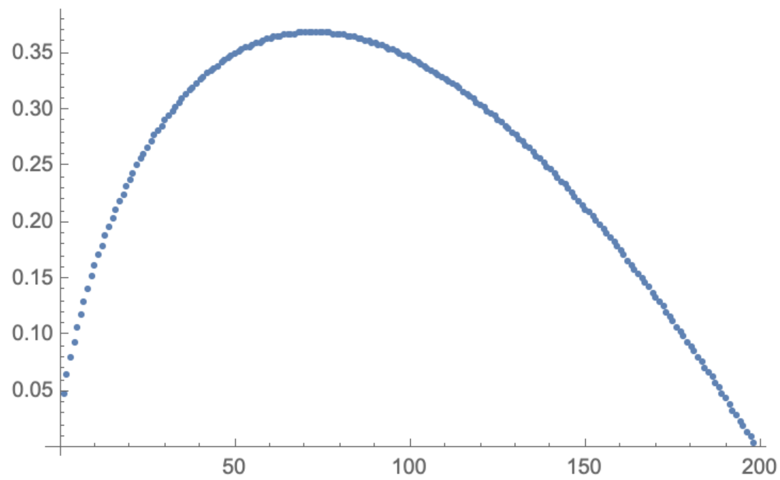
$$p(n,k) = \frac{k}{n}\sum_{j=k}^{n-1}\frac{1}{j}.$$

This is still not trivial to analyze exactly; but it is a snap for a computer to evaluate it, even for fairly large values of n and k. Here is a plot, produced by `Mathematica`, of $p[100,k]$ as k ranges from 1 to 99:



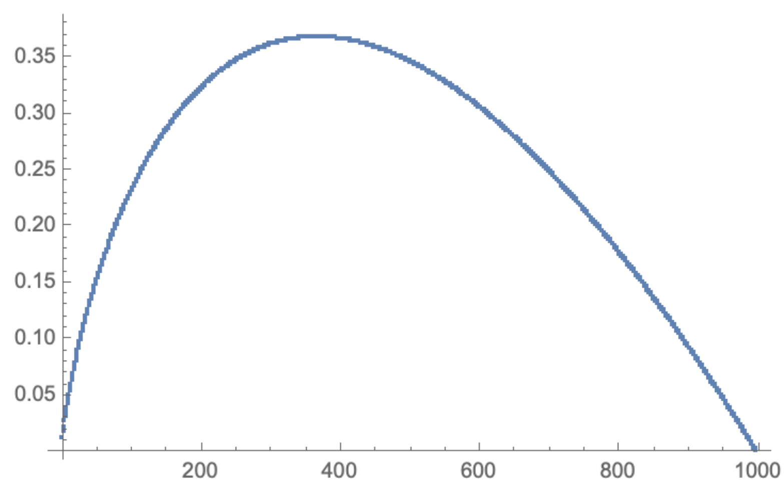
Notice that the plot rises to a unique peak at around $k = 37$, where it takes value around 0.37. This suggests that when interviewing 100 candidates, and using the "discard the first k" strategy, it is best to interview the first 37, and then hire the first person better than all of the first 37 — that gives around a 37% chance of nabbing the best candidate.

Here's a plot of $p[200,k]$ as k ranges from 1 to 199:



It's the same graph, essentially! Now it peaks at around $k = 74$ (37% of 200), and the peak probability is still around 37%.

Finally, here's a plot of $p[1000, k]$ as k ranges from 1 to 999:



Again, the same graph — now it peaks at around $k = 368$ (around 37% of 1000), and the peak probability is still around 37%.

It *is* possible, with some mild calculus, to precisely analyze the expression $p(n, k)$, and with that analysis one can reach the following conclusion, consistent with the three specific examples above:

> When interviewing n candidates, and using the "discard the first k" strategy, it is best to interview the first around $n/e \approx 0.368n$ candidates, and then hire the first person who is better than all of those — this gives around a $1/e \approx 36.8\%$ chance of hiring the best candidate.

Of course, the strategy "discard the first k" isn't the only type of strategy one can consider for this problem. But it turns out that there is no other strategy that does better than the simple one of "discard the first n/e". Explaining why that is so is a little beyond the scope of this note. To learn more, look at the wikipedia page on the problem, and the references given there: `https://en.wikipedia.org/wiki/Secretary_problem`.