Discrete Random Variables

Chapter 3 – Lecture 10

Yiren Ding

Shanghai Qibao Dwight High School

March 27, 2016

1. The best-choice problem
   - Solution
   - General solution
   - Simulation approach
   - Harmonic Sum
   - Euler-Mascheroni Constant
   - Harmonic Approximation Theorem
   - Proof of Theorem
   - Complete solution

Example 1 (The best-choice problem).

Your friend proposes the following wager: twenty people are requested, independently of one another, to write a number on a piece of paper. They may write any number they like, no matter how high. You fold up the twenty pieces of paper and place them randomly onto a tabletop. Your friend opens the papers one by one. Each time he opens one, he must decide whether to stop with that one or go on to open another one. Your friend's task is to single out the paper displaying the highest number. Once a paper is opened, your friend cannot go back to any of the previously opened papers. He pays you one dollar if he does not identify the paper with the highest number on it, otherwise you pay him five dollars. Do you take the wager? If your answer is no, what would you say to a similar wager where 100 people were asked to write a number on a piece of paper and the stakes were one dollar from your friend for an incorrect guess against ten dollars from you if he guesses correctly?

The best-choice problem

- Assume that your friend will use the the strategy of allowing the first half of the papers to pass through his hands, but keeping a mental note of the highest number. Then he stops at the next number that is larger than the one he took note of earlier.
- Let p represent the probability that your friend will win under this strategy. And suppose you will lose \$5 to your friend if he wins and otherwise you receive \$1. The *expected value* of your net win is

$$(1-p)\times 1 - p \times 5 = 1 - 6p.$$

- Hence the contest is unfavorable to you if $p > \frac{1}{6}$.
- Let A be the event that the second highest number is among the first 10 papers, but that the highest number is not.

The best-choice problem

- In this case, $p > P(A)$. By counting, we find that

$$P(A) = \frac{10 \times 10 \times 18!}{20!} = \frac{100}{19 \times 20} = 0.263$$

- So p is indeed greater than $1/6$; hence, your net payoff is negative and the bet is unfavorable to you.
- When the number of people is 100, we have

$$P(A) = \frac{50 \times 50 \times 98!}{100!} = \frac{2,500}{99 \times 100} = 0.253.$$

- In fact, we will show that if the number n of people is sufficiently large, then the maximum probability of winning is approximately $\frac{1}{e} = 0.3679$, irrespective of the value of n.

The best-choice problem

- In real life we often need to know the "optimal stopping." The best-choice problem is one of the most famous examples.
- Suppose there are n items ranked in a random order. The observer is looking for the largest item by going through them one by one. Once he decides to stop, he cannot go back to the previous items.
- A reasonable strategy is to reject the first s items and mentally taking a note of the largest item, and then choose the next item (among the $n-s$ remaining) which has the highest rank.
- We let $p(s,n)$ denote the probability of getting the item of the highest rank when using this strategy.
- This probability $p(s,n)$ can be easily obtained by simulation. We find that $P(10,20)$ is approximately equal to 0.359.

```
import random
def P(s, n):
    count = 0
    numTrials = 100000
    a = list(range(n))
    for num in range(numTrials):
        b = random.sample(a, len(a))
        b1 = b[0:s]
        b2 = b[s: len(a)]
        b1Max = max(b1)
        b2Max = max(b2)
        for i in range(len(b2)):
            if b2[i] > b1Max:
                if b2[i] >= b2Max:
                    count += 1
                break
    prob = count / numTrials
    return prob
print('The probability of winning is approximately', P(10, 20),'.')
```

The best-choice problem

- To find the analytic solution of $p(s, n)$, let A_k be the event that the kth item in the sequence has the highest rank of all n items, and let B_k be the event that the item having the highest rank among the first $k-1$ items appears in the first s items, for $s < k$.
- We find that $P(A_k) = \frac{1}{n}$ and $P(B_k \mid A_k) = \frac{s}{k-1}$, for $s < k$, so

$$p(s, n) = \sum_{k=s+1}^{n} P(A_k B_k) = \sum_{k=s+1}^{n} P(B_k \mid A_k) P(A_k)$$
$$= \sum_{k=s+1}^{n} \frac{s}{k-1} \times \frac{1}{n}.$$

- Let s_n^* denote the value of s for which $p(s, n)$ is maximal, and let p_n^* denote that maximal probability. We shall prove (with a lot of fun) that for n sufficiently large $s_n^* \approx \frac{n}{e}$ and $p_n^* \approx \frac{1}{e} = 0.3679$.

Harmonic Sum

- Before we prove the result, let's talk about the **harmonic sum**, which is the sum of reciprocals of the positive integers,

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

.

- It is well known from calculus that this sum diverges. In addition, we can find a nice approximation by noting that

$$\frac{1}{n} < \int_{n-1}^{n} \frac{dt}{t}, \; n \geq 2 \tag{1}$$

$$\frac{1}{n} > \int_{n}^{n+1} \frac{dt}{t}, \; n \geq 1. \tag{2}$$

Harmonic Sum

- Adding inequalities (1) and (2) for n up to N, we will get

$$\int_1^{N+1} \frac{dt}{t} < \sum_{n=1}^{N} \frac{1}{n} \leq 1 + \int_1^{N} \frac{dt}{t}.$$

- By *The Fundamental Theorem of Calculus*, we have

$$\ln(N+1) < \sum_{n=1}^{N} \frac{1}{n} \leq 1 + \ln N.$$

- This means that the harmonic sum can be approximated by $\ln N$ with an error equal to at most 1:

$$\sum_{n=1}^{N} \frac{1}{n} = \ln N + O(1).$$

Harmonic Sum – Euler-Mascheroni Constant

- This is wonderful, but can we find a more precise approximation of the error $\sum_{n=1}^{N} \frac{1}{n} - \ln N$?
- Let's consider the difference

$$a_n = \frac{1}{n} - \int_n^{n+1} \frac{dt}{t} = \frac{1}{n} - \ln\left(1 + \frac{1}{n}\right).$$

- It is well-known from calculus (will prove in options calculus), that for $-1 < x \leq 1$, $\ln(1+x)$ has its *Taylor series* expansion,

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots$$

- Hence we have the wonderful result

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

Harmonic Sum – Euler-Mascheroni Constant

- This is an alternating series, it is easy to see that

$$0 < a_n < \frac{1}{2n^2}.$$

- A consequence of this inequality is that the sum

$$\sum_{n=1}^{\infty} a_n = \gamma \approx 0.5772156649,$$

  converges to the *Euler-Mascheroni constant.*
- It is not known whether γ is rational or irrational!
- With this in mind, we are ready to state the next theorem.

Theorem 1 (Harmonic Approximation).

For $N \geq 1$, we have

$$\sum_{n=1}^{N} \frac{1}{n} = \ln N + \gamma + O\left(\frac{1}{N}\right).$$

where $O(x)$ represent the error term that is at most as large as cx for some fixed constant c.

- *Proof.* Assume that N is an integer. We have

$$\sum_{n=1}^{N} \frac{1}{n} = \sum_{n=1}^{N} \left(a_n + \int_{n}^{n+1} \frac{dt}{t}\right) = \sum_{n=1}^{N} a_n + \int_{1}^{N+1} \frac{dt}{t}$$
$$= \sum_{n=1}^{\infty} a_n - \sum_{n=N+1}^{\infty} a_n + \ln(N+1) = \ln(N+1) + \gamma - \sum_{n=N+1}^{\infty} a_n.$$

Theorem 1 Proof

- Using the inequality $0 < a_n < \frac{1}{2n^2}$, we have

$$0 < \sum_{n=N+1}^{\infty} a_n < \sum_{n=N+1}^{\infty} \frac{1}{2n^2} < \sum_{n=N+1}^{\infty} \frac{1}{2n(n-1)}$$
$$= \sum_{n=N+1}^{\infty} \frac{1}{2}\left(\frac{1}{n-1} - \frac{1}{n}.\right) = \frac{1}{2N} = O\left(\frac{1}{N}\right).$$

- It remains to show that $\ln(N+1) = \ln N + O\left(\frac{1}{N}\right)$. Again this can be easily verified by its Taylor series expansion:

$$0 < \ln(N+1) - \ln(N) = \ln\left(\frac{N+1}{N}\right) = \ln\left(1 + \frac{1}{N}\right)$$
$$= \frac{1}{N} - \frac{1}{2N^2} + \frac{1}{3N^3} \cdots \leq \frac{1}{N} = O\left(\frac{1}{N}\right). \quad \square$$

The best-choice problem

- Empowered with Theorem 1, recall that probability of winning is,

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

- For a fixed n, let's now define $\Delta_s = p(s,n)-p(s-1,n)$. Hence,

$$\Delta_s=\frac{1}{n}\left(\sum_{k=s+1}^{n}\frac{s}{k-1}-\sum_{k=s}^{n}\frac{s-1}{k-1}\right)=\frac{1}{n}\left(\sum_{j=s}^{n-1}\frac{1}{j}-1\right)$$

- Since Δ_s is decreasing in s, the value s_n^* that maximizes $p(s,n)$ must be the largest value s for which

$$\frac{1}{s}+\cdots+\frac{1}{n-1}-1\geq 0.$$

The best-choice problem

- So we set $\Delta_s = 0$, and use the approximation $\sum_{j=1}^{k-1} \frac{1}{j} \approx \ln k + \gamma$ for sufficiently large k (Theorem 1) to obtain

$$1 = \sum_{j=s_n^*}^{n-1} \frac{1}{j} = \sum_{j=1}^{n-1} \frac{1}{j} - \sum_{j=1}^{s_n^*-1} \frac{1}{j} \approx (\ln n + \gamma) - (\ln s_n^* + \gamma)$$

- This implies that $\ln n - \ln s_n^* \approx 1$ for n large, so we have $\ln(n/s_n^*) \approx 1$, and so $n/s_n^* \approx e$. This proves that $s_n^* \approx \frac{s}{e}$.
- Substitution the value $s_n^* = \frac{s}{e}$ into $p(s, n)$, we obtain that maximal probability of winning tends to

$$p_n^* \approx \frac{1}{e}.$$

- You must be crying right now at the amazing power of calculus.