Homework5

Qinyun Song

1 Markov's inequality

For a random variable X and $X = -9$ with probability 0.5 while $X = 11$ with probability 0.5. Then we can see that

$$E[X] = 0.5 \times -9 + 0.5 \times 11 = 1 \tag{1}$$

And the probability

$$Pr\left[X > 10\right] = 0.5 = \frac{1}{2} \tag{2}$$

2 Comparing concentration inequalities

1. The expectation of X is

$$E\left[X\right] = 1 \times \frac{1}{3} + 0 \times \frac{2}{3} = \frac{N}{3} \tag{3}$$

   Then from the *Markov's inequality*, we can see that

$$Pr\left[X > 2N/3\right] < \frac{E\left[X\right]}{2N/3} \tag{4}$$

   From above we already know the value of $E\left[X\right]$. So we know that

$$Pr\left[X > 2N/3\right] < \frac{N/3}{2N/3} = \frac{1}{2} \tag{5}$$

2. We can calculate the variance of X as

$$var(X) = N \times p \times (1-p) = N \times \frac{1}{3} \times \frac{2}{3} = \frac{2N}{9} \tag{6}$$

3. If we want $X > 2N/3$, when we need

$$|X - E\left[X\right]| = |X - N/3| > |2N/3 - N/3| = N/3 \tag{7}$$

   The *Chebychev's inequality* states that

$$Pr\left[|X - E[X]| > t\right] \leq \frac{var(X)}{t^2} \tag{8}$$

Then by using *Chebychev's inequality*, we now that

$$Pr\left[|X - E[X]| > N/3\right] < var(X)/(N/3)^2 = \frac{var(X) \times 9}{N^2} \tag{9}$$

From above, we know that $var(X) = \frac{2N}{9}$, so we can get

$$Pr\left[|X - E[X]| > N/3\right] < \frac{9 \times 2N}{N^2 \times 9} = \frac{2}{N} \tag{10}$$

4. The *Chernoff inequality* is

$$Pr\left[X \geq cE[X]\right] \leq e^{-(c \ln c - c + 1)E[X]} \tag{11}$$

when $c \geq 1$. Then we know that

$$Pr\left[X > 2N/3\right] = Pr\left[X > 2E[X]\right] < e^{-(2 \ln 2 - 2 + 1)\frac{N}{3}} = e^{-0.129N} \tag{12}$$

3 Hashing

1. For any bin, if it is empty, it means that all objects are matched to the remaining $N - 1$ bins. So the probability that this bin is empty is

$$Pr\left[\text{One bin is empty}\right] = (1 - 1/N)^M = (\frac{N-1}{N})^M \tag{13}$$

2. Define variable X meaning that whether the $i - th$ bin is empty or not. Then from above, we know that

$$E\left[X\right] = N \times (\frac{N-1}{N})^M \tag{14}$$

By applying the equation

$$1 - x \leq e^{-x} \tag{15}$$

We know that

$$E\left[X\right] = N \times (1 - 1/N)^M \leq N \times e^{-M/N} \tag{16}$$

Since $M = 4N \log N$, we finally get that

$$E\left[X\right] \leq N \times e^{-\frac{4N \log N}{N}} = N \times (e^{\log N})^{-4} = N^{-3} \tag{17}$$

Then if we want to know the probability that exist at least one empty bin, we need to calculate

$$Pr\left[X \geq 1\right] \leq E\left[X\right]/1 = N^{-3} < N^{-1} \tag{18}$$

So we know that the probability that exist an empty bin is less than $\frac{1}{N}$.

3. From the above, we know that $(1 - 1/N)^M \leq e^{-M/N}$, in this case, we know that $M = N$. So we know that

$$E\left[X\right] = n \times (1 - 1/N)^M \leq n \times e^{-M/N} = N/e \tag{19}$$

Then by using the *Markov's inequality*, we know that

$$Pr\left[X \geq 0.8N\right] \leq E\left[X\right]/0.8N = fracNe/0.8N = \frac{1}{0.8e} \approx 0.46 < \frac{1}{2} \tag{20}$$

So we show that the probability that 80% of the bins are empty is less than $\frac{1}{2}$.

4. The variables X_i are not independent. Suppose we are thinking about one specific bin b_i, if it is empty, it shows that all the objects are hashed to other bins. Thus will effect the empty status of other bins. So they are related to each other.

4 Election prediction

1. Define a variable x_i meaning that whether the $i-th$ person vote 1 or 0. Then we know that

$$E[x_i] = p \times 1 + (1-p) \times 0 = p \tag{21}$$

Define variable $y_i = x_i - p$. Then we know that

$$E[y_i] = 0 \tag{22}$$

And we can see that

$$y_i = \begin{cases} 1-p & w.p. \quad p \\ -p & w.p. \quad 1-p \end{cases} \tag{23}$$

Then we can see that

$$E\left[y_i^2\right] = (1-p)^2 \times p + (-p)^2 \times (1-p) = p(1-p) \tag{24}$$

So $E\left[Y^2\right] = Np(1-p)$. Then apply the *Chernoff bound*, we know that

$$Pr\left[\sum_i^N x_i > t + np\right] = Pr\left[\sum_i^N y_i > t\right] \leq exp(\frac{-t^2/2}{np(1-p) + 1/3t}) \tag{25}$$

Since we want to predict the majority, we need to see the probability of $X > \frac{1}{2}$. So we need

$$t + np = \frac{N}{2} \tag{26}$$

And we can know that

$$t = n(\frac{1}{2} - p) \tag{27}$$

So we know that

$$Pr\left[X > \frac{N}{2}\right] = Pr\left[Y > \frac{N}{2} - np\right] \leq exp(\frac{-(n(\frac{1}{2}-p))^2/2}{np(1-p) + \frac{1}{3}n(\frac{1}{2}-p)}) \tag{28}$$

Since $p = 0.75$, we know that

$$Pr\left[X > \frac{N}{2}\right] \leq e^{\frac{-(n \times 0.25)^2/2}{0.75N \times 0.25 + \frac{1}{3} \times N \times (-0.25)}} = e^{-0.3N} \tag{29}$$

Since we want to have confidence 99%, we need to set

$$e^{-0.3N} = 0.01 \tag{30}$$

So we know that

$$N \approx 16 \tag{31}$$

2. From the above answer, we know that

$$Pr\left[X > \frac{N}{2}\right] = Pr\left[Y > \frac{N}{2} - np\right] \leq exp(\frac{-(n(\frac{1}{2}-p))^2/2}{np(1-p)+\frac{1}{3}n(\frac{1}{2}-p)}) \tag{32}$$

By applying $p = 0.501$, we know that

$$Pr\left[X > \frac{N}{2}\right] \leq exp(\frac{3N}{2(10^3 - 3 \times 501 \times 499)}) \tag{33}$$

So if we want to have confidence at least 99%, we need that

$$exp(\frac{3N}{2(10^3 - 3 \times 501 \times 499)}) = 0.01 \tag{34}$$

So we know that

$$N \approx 2296925 \tag{35}$$

5 Estimating mean and median

1. For the case $n = 3$ and $k = 2$, define the empirical average as

$$u'_{i_1,i_2} = \frac{1}{2}(A[i_1] + A[i_2]) \tag{36}$$

So we have

$$E[u'] = \frac{1}{3}u'_{1,2} + \frac{1}{3}u'_{2,3} + \frac{1}{3}u'_{1,3} \tag{37}$$

$$= \frac{1}{3}\sum_{i=1}^{3} A[i] \tag{38}$$

$$= u \tag{39}$$

Compare with the method without replacement, we can see that the expectation of mean using replacement is more close to the actual mean. The expectation with replacement can be the same as the original mean. The variance can now be calculated as

$$var = E\left[(u - E[u])^2\right] \tag{40}$$

$$= \sum_{i=1}^{3}(Pr[i] \times (i - E[u_i])) \tag{41}$$

$$= \frac{1}{18}(u_1^2 + u_2^2 + u_3^2 - u_1u_2 - u_1u_3 - u_2u_3) \tag{42}$$

Compare with the replacement case, the variance without replacement is $(\frac{1}{9})(2u_1^2 + 2u_2^2 + 2u_3^2 + u_1u_2 + u_1u_3 + u_2u_3)$.

2. *Proof.* Let us just consider the most simple case. That is $n = 3$. Suppose the sequence is $[0, 0.4, 1]$. We can see that the median of it is 0.4. When $k = o(n)$, possible value of k can vary from one to three. Consider different cases one by one.

$k = 1$ In this case, possible empirical median value can be $\{0, 0.4, 1\}$ with probability $\frac{1}{3}$. And both $0, 1$ are off by an amount ≥ 0.1.So with probability $\frac{2}{3}$, the empirical median is off by an amount ≥ 0.1 which we can regard as high probability.

$k = 2$ **without replacement** In this case, three possible empirical median are $0.2, 0.5, 0.7$ with same probability $\frac{1}{3}$. So we can see that the empirical median is *always* off by an amount ≥ 0.1.

$k = 2$ **with replacement** In this case, we can check all possible empirical median

	0	0.4	1
0	0	0.2	0.5
0.4	0.2	0.4	0.7
1	0.5	0.7	1

From the table we can see that, with probability $\frac{8}{9}$ the empirical median is at least off by an amount ≥ 0.1. This probability can be regarded as high probability.

$k = 3$ In this case, we didn't try to use less examples to estimate the original median. So this case is disgarded.

So after all, we can show that, there are examples in which the estimate is off by an amount ≥ 0.1. □

6 Randomized Min-Cut

1. *Proof.* Since E' is needed to be cut, it will devide the graph into two parts. And the vertices in one part cannot reach the vertices in another part. Since the each we remove is not in E', they must belong to exactly one part. So when we remove the edge, we will not hurt E' at all. So the character of E' remains the same. So the size of the min cut in the new graph is equal to that in G. □

2. *Proof.* First of all, the total degree of all vertices is $2|E|$. When we pick one vertex at random, the average of its degree is

$$\sum_i p_i \times degree_i = \frac{1}{n}\sum_i degree_i = \frac{1}{n}2|E| = \frac{2|E|}{n} \tag{43}$$

So there must exist one vertex u with $degree_u \leq \frac{2|E|}{n}$. Because otherwise, the total degree of all vertices should be bigger than $2|E|$ which is contradict with the problem. So we can simply choose the vertex u as one side and all other vertices as another side. Then the min cut of this situation cannot be bigger than $\frac{2|E|}{n}$. So if E' is one of the min cut, we have $|E'| \leq \frac{2|E|}{n}$. □

3. *Proof.* Suppose the edge we choose is e. Let E' denotes one of the min cut of graph $G = (V, E)$. Since we need that the min cut value is maintained, from problem 1, we know that we wish $e \notin E'$. The probability that $e \in E'$ can be calculated as

$$Pr\left[e \in E'\right] = |E'|/|E| \leq 2|E|/n|E| = \frac{2}{n} \tag{44}$$

So the probability that the min cut can remain is no less than $(1 - \frac{2}{n})$. □

4. *Proof.* For any min cut, lets say E'_i, the probability that we end up with it

$$Pr\left[\text{min cut is}E'_i\right] = Pr\left[e_1 \notin E'_i\right] \times \cdots \times Pr\left[e_{n-2} \notin E'_i\right] \tag{45}$$

From the result from above problem, we can see that

$$Pr\left[\text{min cut is}E_i'\right] \geq (1-\frac{2}{n}) \times \cdots \times (1-\frac{2}{3}) = \frac{2}{n(n-1)} \geq \frac{2}{n^2} \tag{46}$$

So we know that, for one min cut E_i', we have probability no less than $\frac{2}{n^2}$ such that we end up with this min cut. Suppose there are k possible min cut, the total probability that we end up with one possible min cut is

$$Pr\left[\text{End up with one min cut}\right] \geq \sum_{i=1}^{k} \frac{2}{n^2} \tag{47}$$

And it is obviously that the sum of these probabilities cannot be lager than one. So we know that

$$k \times \frac{2}{n^2} \leq 1 \tag{48}$$

And we can conclude that

$$k \leq \frac{n^2}{2} \tag{49}$$

So the possible number of min cuts is no bigger than $\frac{n^2}{2}$. □