COMPSCI 514: Problem Set 1

Due: Friday 2/20 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- Each group should **submit a single solution set**: one member should upload a pdf to Gradescope, marking the other members as part of their group in Gradescope.
- The problem set is meant for your own practice. We strongly discourage the use of LLMs to directly solve problems.
- Each problem will be graded on the following scale:
  - ✓+: (2 points) Submitted work demonstrates a full understanding of the problem. There may be some errors, omissions, or unclear steps, but overall, a reader would be able to understand how to solve the problem by looking at the submitted work.
  - ✓-: (1 point) Submitted work demonstrates partial understanding of the concepts, but contains significant omissions or errors.
  - X: (0 points) Not completed or submitted work doesn't not provide enough information to determine whether there is understanding of the problem.
- The 'Challenge Problems' are **completely optional** and not worth any extra credit. You may want to complete them e.g., if you are interested in further advanced study or research in the area, or if you just find the problem interesting.

1. Concentration Bound Practice

1. On a given day, 1000 users visit your website. User i makes a purchase with probability p_i. You are given that $\sum_{i=1}^{1000} p_i = 100$. Let $\mathbf{S}$ be the total number of purchases that are made. What is $\mathbb{E}[\mathbf{S}]$?
2. Use Markov's inequality to give an upper bound on the probability that at least 500 purchases are made.
3. Let $\mathbf{S}_i$ be an indicator random variable which is 1 if user i makes a purchase and 0 otherwise. Show that $\text{Var}[\mathbf{S}_i] \leq p_i$.
4. Assume the events that each user makes a purchase are independent of each other. Use this assumption and part (3) to give an upper bound on $\text{Var}[\mathbf{S}]$.
5. Use part (4) and Chebyshev's inequality to give a tighter upper bound (as compared to part (2)) on the probability that at least 500 purchases are made.
6. If we could not guarantee that the purchases were independent events, what is the largest $\text{Var}[\mathbf{S}]$ could be? If we used this value in Chebyshev's inequality to bound the probability of 500 purchases, what would we get?

2. Probability and Expectation Practice

1. Prove that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random variables, $\mathbb{E}[\mathbf{X} \cdot \mathbf{Y}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]$.

2. Let $\mathbf{Z}$ be a random variable, equal to 1 with probability 1/2 and 2 with probability 1/2. Let $\mathbf{X}$ and $\mathbf{Y}$ both be independent random variables, equal to 1 with probability 1/2 and -1 with probability 1/2. Assume $\mathbf{Z}, \mathbf{X}$, and $\mathbf{Y}$ are all independent. Are $\mathbf{X} \cdot \mathbf{Z}$ and $\mathbf{Y} \cdot \mathbf{Z}$ independent? Are they uncorrelated? I.e., we do have $\mathbb{E}[(\mathbf{XZ})(\mathbf{YZ})] = \mathbb{E}[\mathbf{XZ}]\mathbb{E}[\mathbf{YZ}]$?

3. Design two random variables $\mathbf{X}$ and $\mathbf{Y}$ that satisfy $\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{Y}) > \mathrm{Var}(\mathbf{X} + \mathbf{Y})$. Design two different random variables that satisfy $\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{Y}) < \mathrm{Var}(\mathbf{X} + \mathbf{Y})$

4. For a random variable $\mathbf{X}$, let $D[\mathbf{X}] = \mathbb{E}[|\mathbf{X} - \mathbb{E}[\mathbf{X}]|]$ denote the mean absolute deviation of $\mathbf{X}$ from its mean (this is like the variance but without squaring the deviation). True or False: if $\mathbf{X}, \mathbf{Y}$ are independent then $D[\mathbf{X} + \mathbf{Y}] = D[\mathbf{X}] + D[\mathbf{Y}]$. If true, prove it. If false, give a counterexample.

5. Consider storing n items in a hash table with $m = n$ buckets, using a fully random hash function $\mathbf{h} : [n] \to [n]$ (i.e., each item is assigned independently to a uniform random bucket). What is the probability that a given item lands in its own bucket (i.e., that it does not collide with any other items)? What is the limit of this probability as $n \to \infty$?

3. Population Size Estimation via Mark-and-Recapture

You want to estimate the number of individuals in a large population (e.g., a population of animals in some area), by randomly capturing individuals from the population, tagging them, and observing if you re-capture them in the future. The less re-captures you see, the higher your estimate for the population size will be. This idea is widely employed in ecology for population size estimation, and is similar to the CAPTCHA database example discussed in class.

1. Consider capturing m individuals, which you assume are drawn independently and uniformly at random with replacement from a population of size n. Let $\mathbf{C}$ denote the number of pairs of captured individuals that are the same (which you can count by observing if you have already tagged any captured individuals). Prove that $\mathbb{E}[\mathbf{C}] = \frac{\binom{m}{2}}{n}$.

2. For any $i < j$, let $\mathbf{C}_{i,j}$ be a random variable, which is 1 if the i^{th} and j^{th} captured individuals are the same, and 0 otherwise. Observe that the $\mathbf{C}_{i,j}$ random variables are pairwise independent. I.e., for any two pairs (i, j) and (k, ℓ) that differ in at least one element, $\mathbf{C}_{i,j}$ and $\mathbf{C}_{k,\ell}$ are independent.

   Prove that for any set of pairwise independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_z$,

   $$\mathrm{Var}\left[\sum_{i=1}^{z} \mathbf{X}_i\right] = \sum_{i=1}^{z} \mathrm{Var}[\mathbf{X}_i].$$

   This is, pairwise independence suffices for linearity of variance to hold. **Hint:** Use that $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$ and that when $\mathbf{X}, \mathbf{Y}$ are independent, $\mathbb{E}[\mathbf{X} \cdot \mathbf{Y}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]$.

3. Use the above fact to show that $\mathrm{Var}[\mathbf{C}] \leq \frac{\binom{m}{2}}{n}$. **Hint:** First compute $\mathrm{Var}[\mathbf{C}_{i,j}]$.

4. Prove that for any $\epsilon, \delta \in (0,1)$ if we set $m \geq \frac{2\sqrt{n}}{\epsilon\sqrt{\delta}}$ then with probability at least $1-\delta$, $|\mathbf{C} - \mathbb{E}[\mathbf{C}]| \leq \epsilon\mathbb{E}[\mathbf{C}]$. **Hint:** I used that fact that for $m \geq 2$, $\binom{m}{2} \geq \frac{m^2}{4}$ to simplify my calculations. Do not worry about getting the absolute tightest bound here. If you can show the above bound for $m \geq \frac{c\sqrt{n}}{\epsilon\sqrt{\delta}}$ for any fixed constant c, that will be considered correct.

5. Consider estimating the population size as $\tilde{n} = \frac{\binom{m}{2}}{\mathbf{C}}$. Prove that if $|\mathbf{C} - \mathbb{E}[\mathbf{C}]| \leq \frac{\epsilon}{2} \cdot \mathbb{E}[\mathbf{C}]$ for some $\epsilon \in (0,1)$, then $|\tilde{n} - n| \leq \epsilon n$. **Hint:** Use that for any $x \in (0, 1/2)$, $\frac{1}{1-x} \leq 1 + 2x$ and that for any $x \in (0,1)$, $\frac{1}{1+x} \geq 1 - x$.

6. Conclude that for any $\epsilon, \delta \in (0,1)$, setting $m \geq \frac{4\sqrt{n}}{\epsilon\sqrt{\delta}}$ suffices to estimate the population size to error ϵn with probability at least $1 - \delta$.

4. Implementing Mark-and-Recapture[1]

The link https://en.wikipedia.org/wiki/Special:Random will bring you to a random article on English language Wikipedia.

1. Use this link to implement the mark-and-recapture algorithm from Problem 3 and evaluate the claim that English language Wikipedia has 7.1 million unique articles (see e.g., https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.)

2. Include your code, all relevant results and calculations, and a discussion of how accurate you think your estimate is. We should be able to understand your approach and results without looking at the code. The code is just as a sanity check that you completed the problem.

3. Describe your methodology for choosing the number of random articles sampled to make your estimate. Did you choose as single value up front? Did you have to adjust it at all?

4. How do you think the results are impacted by the fact that the random article feature doesn't return a truly uniformly random articles (see discussion at https://en.wikipedia.org/wiki/Wikipedia:FAQ/Technical#random)? Do you think this biases your estimate to be too low or too high and why? You might want to look at Optional Challenge Problem C1 when thinking about this.

Hint: In Python, you can obtain a random url by running:

```
 import requests
headers = {'User-Agent': 'wikiDuplicateScript (fakeEmail@gmail.com)'}
response = requests.head("https://en.wikipedia.org/wiki/Special:Random",allow_redirects=True,headers=headers)
random_url = response.url
```

In my experiments, downloading 5000 articles took around 30 minutes, so your code might take a bit to run. But compare this to scanning all possible articles to check the claim, which would take roughly 25 days (if your IP isn 't blocked for scrapping).

[1] This problem is taken from Chris Musco's NYU Course: *CS-GY 6763: Algorithmic Machine Learning and Data Science.*

Challenge Problems (OPTIONAL)

C1. Analyzing Mark-and-Recapture for Wikipedia 🌶🌶🌶

In Problem 4, you should see consistent inaccurate estimates for the article count using the mark-and-recapture method. We suspect that this is due to the fact that the random article feature doesn't return truly uniformly random articles (see https://en.wikipedia.org/wiki/Wikipedia:FAQ/Technical#random). Here, we will analyze carefully how the distribution of the random article generated affects our results.

We will model the random article generator as follows: there are n articles. Each is independently assigned a uniformly random real number $r_i \in [0, 1]$, which serves as its id. To randomly sample an article, we sample a uniformly random real number $x \in [0, 1]$. We then return the article with the lowest id greater than x. If there is such no article (i.e., all articles have ids $< x$), then we return the article with the lowest id.

Hint: It maybe be helpful to visualize the process by considering a circle with circumference 1. The articles are each placed uniformly at random at n points along this circle according to their ids. An article is returned by the random article generator if the random point x lies in between that article and the article directly adjacent to it, counter-clockwise.

1. Let $\mathbf{p}_i$ denote the probability that article i is returned by the generator. Show that the expected number of collisions amongst m random articles is $\mathbb{E}[\mathbf{C}] = \binom{m}{2} \cdot \sum_{i=1}^{n} \mathbf{p}_i^2$.

2. Note that each $\mathbf{p}_i$ is a random variable that depends on $r_1, \ldots, r_n$. What is $\mathbb{E}[\mathbf{p}_i]$? **Hint:** Use linearity of expectation and symmetry.

3. For any t, what is $\Pr[\mathbf{p}_i \geq t]$? **Hint:** What sequence of $n-1$ events needs to happen for the event $\mathbf{p}_i \geq t$ to occur?

4. What is $\mathbb{E}[\mathbf{p}_i^2]$? What is $\mathbb{E}[\sum_{i=1}^{n} \mathbf{p}_i^2]$? **Hint:** Use part (2) along with the fact that for a non-negative random variable $\mathbf{X}$, $\mathbb{E}[\mathbf{X}] = \int_0^\infty \Pr[\mathbf{X} \geq t]\, dt$.

5. Use parts (1) and (4) to explain why our Wikipedia article count estimates were roughly half of what we expected.

C2. Testing Uniform Samples with Duplicates 🌶🌶

Let P be a distribution over $[n]$ that places probability p_i on outcome i. We would like to take a small number of samples from P and determine if P is close to uniform – i.e., if it is close to placing probability $1/n$ on each outcome. We let $\Delta(P) \stackrel{\text{def}}{=} \sum_{i=1}^{n} |p_i - 1/n|$ denote P's distance to uniformity.

1. Consider taking m independent samples from P. Let $\mathbf{D}$ be the number of pairwise duplicate samples we observe. Prove that $\mathbb{E}[\mathbf{D}] = \binom{m}{2} \cdot \sum_{i=1}^{n} p_i^2$ (same as C1 part (1)).

2. Prove that $\mathrm{Var}[\mathbf{D}] \leq \binom{m}{2} \cdot \sum_{i=1}^{n} p_i^2 + 6\binom{m}{3} \cdot \sum_{i=1}^{n} p_i^3$.

3. Argue that for any $\gamma \in (0, 1)$, if we take $m = \frac{c\sqrt{n}}{\gamma^2}$ samples for a large enough constant c, then, with probability at least $9/10$, $|\mathbf{D} - \mathbb{E}[\mathbf{D}]| \leq \gamma \mathbb{E}[\mathbf{D}]$.

4. Let $\Delta_2(P) = \sum_{i=1}^{n} (p_i - 1/n)^2$. Prove that $\sum_{i=1}^{n} p_i^2 = 1/n + \Delta_2(P)$. In turn, prove that $\sum_{i=1}^{n} p_i^2 \geq \frac{1+\Delta(P)^2}{n}$.

5. Describe and analyze an algorithm that takes $O(\sqrt{n}/\epsilon^4)$ samples from P and satisfies: a) if P is in fact the uniform distribution, the algorithm outputs YES with probability at least 9/10; b) if $\Delta(P) \geq \epsilon$, the algorithm outputs NO with probability at least 9/10.