Максимумът на точките, които можете да получите общо от всички домашни е 50, като това кореспондира с бонус от 0.5 към оценката за упражнения. Успех.

Ще считаме, че навсякъде работим върху вероятностно пространство $(\Omega, \mathcal{A}, \mathbb{P})$. *За удобство, дефинираме* $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Задача 1. (15 т., Coupon collector) Да предположим, че номерата на колите в България са случайно разпределени. Искаме да оценим колко коли приблизително трябва да видим, за да сме видели всички номера от 0000 до 9999.

Да дефинираме случайните величини

- X = "броят коли, които е трябвало да видим, за да видим всички номера";
- за $i \in \mathbb{N}$, X_i = "броят коли, които е трябвало да видим, за да видим i-тия различен номер, след като сме видели $(i-1)$ различни номера.

Пример: Виждаме на улицата номерата 0231, 2313, 4441, 1234, 2313, 1244, 0231, 1444. Още е прекалено рано да определим X. $X_1 = 1$ - трябва да видим 1 кола, за да видим 1 различен номер. $X_2 = 1, X_3 = 1, X_4 = 1$, но $X_5 = 2$, тъй като е трябвало да видим 2 коли след като сме "събрали" 4 различни номера, за да видим 5-ти различен. Аналогично $X_6 = 2$.

1. Какво е разпределението на сл. вел. X_i за $i \in \mathbb{N}$? Намерете очакването и дисперсията ѝ.
2. Каква е връзката между X и $X_1, \ldots, X_{10000}$? Намерете $\mathbb{E}X$ и DX.
3. Ако нарисувате графиката на функцията $f(x) = 1/x$, може да се опитате да забележите, че $1+1/2+\cdots+1/n$ е близко до $\int_1^n \ln x\, dx = \ln n$. Така може да дадете приближена стойност на $\mathbb{E}X$. Използвайте компютър, за да пресметнете колко е грешката в това приближение. *(За сравнение, аз получавам грешка около 5770).*
4. * Всъщност, вярно е, че
$$1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n \xrightarrow[n\to\infty]{} \gamma \approx 0.5772\,,$$
където γ е константата на Ойлер-Маскерони. Ако използваме този резултат, бихме могли да оценим $\mathbb{E}X \approx 10^4(\ln 10^4 + \gamma) \approx 97876$, което е с грешка по-малка от 0.5 от точната стойност на $\mathbb{E}X$!
5. Получихме, че $\mathbb{E}X$ е грубо около 100000 , но това не значи, че ако видим 100000 коли ще сме сигурни с "голяма" вероятност, че сме видели всички номера. Тук (по най-лесен, но не много точен начин) може да ни помогне неравенството на Чебишов:
$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{DX}{a^2}$$
за $a > 0$. Да кажем, че малка вероятност значи 1/100 (или изберете друга стойност сами). Ако изберем a, така че $\mathbb{P}(|X-\mathbb{E}X| \geq a) \leq 1/100$, то в такъв случай, ако видим $\mathbb{E}X+a$ коли, ще сме видели всички номера с вероятност поне 99%. Използвайте компютър, за да сметнете DX и съответно a. (*за сравнение, аз получавам* $DX \approx 1.64386 * 10^8, a \approx 128213$, *като второто е грубо. В крайна сметка, ако чакаме около 230000 коли, с вероятност 99% ще сме видели всички номера. По 2 коли на 10 секунди, това са около 80 дни.)*

Задача 2 (20 т., залагане на ези/тура)**.** Нека $n \geq 3$ и p са естествени числа. Разглеждаме следната игра с участници $A_1, \ldots, A_n$:

- Всеки играч написва предвижданията си за изходите от $(2p+1)$ хвърляния на стандартна монета (1/2 вероятност за ези и тура), като пише E за ези и T за тура;
- Монетата се хвърля $(2p+1)$ пъти и играчите с най-много познати си разделят награда от $n!$ лева.

Пример: $n = 4, p = 1$, предвиждания: ETE, EEE, TTT, EET. При изход TTE, играчи 1 и 3 имат по 2 познати и си разделят 4!, т.е. всеки взима по 12 лева.

$n = 4, p = 1$, предвиждания: ETE, EEE, TTT, EET. При изход TTT, играч 3 има 3 познати и печели $4! = 24$ лева.

За $i \in [n]$ да дефинираме

- X_i = "броят познати от A_i изходи";
- G_i = "печалбата на играч A_i";

- $q_k = \mathbb{P}(X_i = k), r_k = \mathbb{P}(X_i \leq k)$.

Целта ни ще е да пресметнем очакваната печалба на A_1, т.е. $\mathbb{E}(G_1)$ при две различни стратегии.

Част 1. Начални съображения.

1. * Както винаги работим върху вероятностно пространство $(\Omega, \mathcal{F}, \mathbb{P})$. Кое множество бихте предложили за Ω?

2. Какво е разпределението на сл. вел. X_i?

3. Пресметнете r_p.

4. Напомняме, че с $X(\Omega)$ означаваме всички образи на елементи от Ω под функцията X. Например, $X_1(\Omega) = \{0, 1, \ldots, 2p+1\}$, т.к. играчът A_1 може да уцели от 0 до $2p+1$ изхода.

   Ако $\mathbb{P}(A) > 0$, дефинираме условно очакване на сл. вел. X при условие A като

   $$\mathbb{E}(X|A) = \sum_{k \in X(\Omega)} k\mathbb{P}(X = k|A)\,.$$

   Докажете, че ако $B_1, \ldots, B_s$ е пълна система от събития с ненулева вероятност, то

   $$\mathbb{E}X = \sum_{k=1}^{s} \mathbb{E}(X|B_k)\mathbb{P}(B_k)\,.$$

   (Може да разпишете $\mathbb{E}X$ по дефиниция, да приложите формулата за пълната вероятност и да разменените реда на сумиране.)

Част 2. Всички играчи играят случайно. В тази част приемаме, че всички играчи избират ези/тура случайно с вероятност $1/2$ и независимо един от друг.

1. На какво е равно е $G_1(\Omega)$, т.е. какви са възможните печалби на играча A_1?

2. За всяко $k \in X_1(\Omega), j \in [n]$, докажете, че

   $$\mathbb{P}\left(G_1 = \frac{n!}{j} \;\middle|\; X_1 = k\right) = \binom{n-1}{j-1}(r_k - r_{k-1})^{j-1} r_{k-1}^{n-j}\,.$$

3. Докажете, че за $k > 0$

   $$\mathbb{E}(G_1|X_1 = k) = \frac{(n-1)!(r_k^n - r_{k-1}^n)}{q_k}\,.$$

   Забележете, че горната формула е вярна и за $k = 0$, ако дефинираме $r_{-1} = 0$.

   (Разпишете с помощта на предишната подточка и използвайте бинома на Нютон)

4. Заключете, че $\mathbb{E}(G_1) = (n-1)!$.

   (Разгледайте пълната система от събития $\{X_1 = k\}, k \in \{0, 1, \ldots, 2p+1\}$)

5. Очаквано ли е заключението на горната подточка? Можехме да достигнем по него и по друг начин: можем ли да пресметнем $G_1 + \cdots + G_n$? Използвайте линейността на очакването и симетричността, за да получите (отново) $\mathbb{E}(G_1)$.

Част 3. A_1 и A_2 играят заедно, а всички други играят случайно. В тази част, да приемем, че всички освен A_2 играят случайно. A_2 от своя страна се познава с A_1 и залага на обратното на A_1. Например, ако A_1 залага на EET, то A_2 ще заложи на TTE. Ако някой от двамата спечели, си делят печалбата поравно.

Да означим

- с G', G'_1, G'_2 респективно общата печалба на A_1 и A_2, и печалбите им поотделено; (В такъв случай $G' = G'_1 + G'_2$.)

- $Y =$ "по-големият брой познати изходи от A_1 и A_2"(при EET, TTE и изход от хвърлянията EEE, $Y = 2$);

- както преди, за $k \in \{1, 3, 4, \ldots, n\}$, $q_k = \mathbb{P}(X_i = k), r_k = \mathbb{P}(X_i \leq k)$.

1. Докажете, че $Y(\Omega) = \{p+1, \ldots, 2p+1\}$.

2. Докажете, че за $k \in \{p+1, \ldots, 2p+1\}, \mathbb{P}(Y = k) = 2q_k$.

3. Докажете, че за $k \in \{p+1, \ldots, 2p+1\}, j \in \{0, \ldots, n-2\}$

$$\mathbb{P}\left(G' = \frac{n!}{j+1} \;\middle|\; Y = k\right) = \binom{n-2}{j}(r_k - r_{k-1})^j r_{k-1}^{n-2-j}$$

и като следствие (използвайки бинома на Нютон), че

$$\mathbb{E}(G'|Y=k) = n(n-2)!\frac{r_k^{n-1} - r_{k-1}^{n-1}}{q_k}.$$

4. Заключете, че

$$\mathbb{E}(G') = 2n(n-2)!\left(1 - \frac{1}{2^{n-1}}\right)$$

и проверете, че тази стратегия е по-добра за A_1 и A_2 от тази в част 2.

5. Пресметнете $\mathbb{E}(G_i)$ за $i \in \{3, \ldots, n\}$.

Задача 3. (15 т., Симулации на случаен граф) Случайните графи са област на вероятностите, която привлича особен интерес през последните години, намирайки връзки със статистическата механика, моделирането на големи мрежи (Facebook, Google Searches, DNA). Проблемът в директното разглеждане на графа на познанствата във Фейсбук например е, че дори да имаме достъп до него, матрицата му на инцидентност би прекалено голяма, за да можем да я манипулираме с днешния хардуер. Затова, ако имаме добър вероятностен модел, това може да ни позволи да правим наблюдения без проверки (например да докажем, че в при всеки граф, който симулираме ще можем да стигнем от всеки до всеки връх с най-много X стъпки - така наречения small world network)

Тук ще се запознаем емпирично с най-лесния модел на случаен граф - този на Ердьош и Рени (1960).

Нека n е броят на върховете на граф G и $p \in [0,1]$ е параметър. За всяка двойка от върхове, хвърляме монета с вероятност за ези p. Ако се падне ези, свързваме тези върхове, а в противен случай - не.

Целта ни ще е да анализираме вероятността графът да е свързан.

1. Параметрите, които ще меним са n, p и N - броят симулации. При всеки избор на n и p ще трябва да симулирате голям брой графи ($N = 10000, 1000, 100$? изборът на N и n ще зависи от това колко симулации са постижими в обозримо време). При фиксирани n и p, това може да се направи като итерирате по всички двойки върхове и избирате всеки път $Ber(p)$, например чрез *scipy.stats.bernoulli*.

   След като сте избрали кои от $\binom{n}{2}$-те двойки върхове са свързани с ребро, то проверете дали графа е свързан. Ако в M от N-те симулации графът е свързан, оценката ни за вероятността това да е така в общия случай ще е M/N.

   Направете графика на частта свързани графи (M/N) към p - на оста x ще бъде параметъра p, а по y - получената пропорция. Обяснете защо функцията трябва да е растяща.

   Пробвайте с $n = 100, 1000, 5000, 10000, 100000$? $N = 10000$? и разгледайте за p стойностите от 0 до 1 през някаква стъпка (например 0.01).

   Можете да пресметнете броя операции при фиксирани n, p и N, за да знаете какви параметри можете да избирате.

   Забелязвате ли нещо? Разгледайте стойности за p, които са $o(\ln n/n)$, например $\ln(\ln(n))/n$ и такива, които са $\omega(\ln n/n)$, например $(n/2)/n = 1/2$.

2. * Направете същото, но не оценявайте пропорцията на свързаните графи, а големината на най-голямата свързана компонента. В първата част, разглеждахме само дали големината на най-голямата свързана компонента е n или не. Резултат още от първите статии на Ердьош и Рени гласи, че, ако n е голямо, то при $p = o(1/n)$ нямаме свързана компонента с линейна (по n) големина, а ако $p = \omega(1/n)$, т.е. $1/n = o(p)$, то имаме такава компонента. Можете ли да забележите това емпирично?

Задача 4. (15 т., Неравенство на Марков, Граници на Чернов) Нека X е положителна случайна величина с крайно очакване и $a > 0$.

1. Като използвате, че $\mathbb{E}X = \mathbb{P}(X \geq a)\mathbb{E}(X|X \geq a) + \mathbb{P}(X < a)\mathbb{E}(X|X < A)$, докажете неравенството на Марков

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

   Докажете чрез него неравенството на Чебишов

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{DX}{a^2}.$$

2. Докажете неравенството на Чернов: за всяко $t > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\,(e^{tx})}{e^{ta}}\,.$$

(Каква е връзката между $\{X \geq a\}$ и $\{e^{tX} \geq e^{ta}\}$?)

3. Нека $X_1, \ldots, X_n \sim Ber(1/2)$ са независими и $X = X_1 + \cdots + X_n$. Какво е разпределението, очакването и дисперсията на X? Пресметнете $\mathbb{E}e^{tX}$.

*(*Може да направите най-добра оценка за $\mathbb{P}(X \geq a)$ в неравенството на Чернов като минимизирате $\mathbb{E}(e^{tx})/e^{ta}$ по $t > 0$.)*

4. Да приемем без доказателство следното следствие от неравенството на Чернов: за $\delta \in (0; 1)$ и дефинираното по-горе X

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq \delta n/2\right) \leq e^{-\frac{n\delta^2}{6}}\,.$$

Колко трябва да изберем δ, така че да сме сигурни, че $X \in [\mathbb{E}X - \delta n/2; \mathbb{E}X + \delta n/2]$ с вероятност най-много $4/n$.

А колко трябва да изберем a в неравенството на Чебишов за да сме сигурни, че $X \in [\mathbb{E}X - a; \mathbb{E}X + a]$ с вероятност най-много $4/n$. Забележете, че неравенството на Чернов дава интервал, който е по порядък по-малък от получения чрез неравенството Чебишов. Сравнете двата интервала за $n = 10000$.

Задача 5. (15 т., Bogosort) Както знаете, асимптотично, скоростта на сортиращ алгоритъм, основан на сравнения, е най-малко $O(n \log n)$ (може да го докажете директно с формулата на Стирлинг). Quicksort например има скорост $O(n^2)$ в най-лошия случай, но в средния случай, т.е. очакването на броя операции при случайно подреден масив, е $O(n \log n)$. Тъй като последното е класически резултат, ще разгледаме друг алгоритъм, като илюстрация на вероятностен анализ на алгоритми.

Bogosort е класически пример за изключително неефективен алгоритъм: за да сортираме масив A, пермутираме случайно елементите му и проверяваме дали полученият масив е сортиран.

За улеснение ще разглеждаме масиви с различни елементи. Ще считаме, че сортираме в **нарастващ** ред.

Algorithm 1: Bogosort

```
input  : array A [1 ... n]
output: sorted A
while While A[1...n] is not sorted (algorithm 2) do
 | randomly permute A (algorithm 3)
end
```

Algorithm 2: Check if sorted

```
input  : array A [1 ... n]
output: true if if A is sorted, false otherwise
for i = 1 to n- 1 do
 | if A[i] > A[i+1] then
 |  | return false
 | end
end
return true
```

Algorithm 3: Randomly permute

```
input  : array A [1 ... n]
output: a random permutation of A
for i = 1 to n- 1 do
 | j := random[i, i+1, ... n]
 | swap A[i] and A[j]
end
```

1. Докажете, че алгоритъм 3 наистина продуцира равномерно случайна пермутация на елементите на даден масив.

2. Колко сравнения са нужни, за да разбере алгоритъм 2, че масивът не е сортиран, ако първият елемент, който не е на мястото си, е на k-та позиция.

3. Нека σ е случайна пермутация на $[n] = \{1, \ldots, n\}$. Дефинираме случайните величини I_k за $k \in [n-1]$ като 1, ако поне първите k елемента на σ са подредени и 0 иначе. Например за $\sigma = (1243)$, $I_1 = 1, I_2 = 1, I_3 = 0, I_4 = 0$. Какво е разпределението и очакването на I_k?

4. Нека C е броят сравнения, нужни за определянето дали σ е сортирана. Изразете $\mathbb{P}(C \geq k)$ чрез I_k и като използвате (докажете, ако искате), че $\mathbb{E}(C) = \sum_{k>0} \mathbb{P}(C \geq k)$, докажете, че

$$\mathbb{E}(C) = \sum_{i=1}^{n-1} \frac{1}{i!} = e - 1 + O\left(\frac{1}{n!}\right) \sim e - 1\,.$$

(Ако имате проблем с доказването на грешката от $O(1/n!)$, докажете само еквивалентността.)

5. Нека I е случайната величина, равна на броя пермутации, които ще пробваме, преди да попаднем на правилната. Какво е разпределението и очакването на I?

6. Колко е очакваният брой на размените, направени за разместването на елементите? А в най-добрия/лошия случай?

(Ако за една итерация са нужни k размествания, общо ще са нужни $I \cdot k$)

7. Нека $C_n =$"броят сравнения направени от Bogosort, при вход с дължина n". Докажете, че

$$\mathbb{E}(C) = \begin{cases} n-1 & \text{в най-добрия случай;} \\ (e-1)n! + n + O(1) & \text{в най-лошия случай;} \\ (e-1)n! + O(1) & \text{в средния случай, т.е. при случайно подреден вход.} \end{cases}$$

(Разсъждавайте като в горната подточка. Ако се чудите дали това е напълно коректно, вижте "Wald's equation".)

В крайна сметка каква е сложността на Bogosort?