Úlohy na cvičenia pre 3. týždeň semestra

1. Na rozohriatie: **Hod mincou**

Budeme hádzať mincou pomocou príkazu ***rbinom()*.** Testujte postupne situáciu pri 10 hodoch, pri 100, 1000, 10 000, 100 000 hodoch, aká je pravdepodobnosť, že hodíte „hlavu“.

Fyzické sčítanie pozitívnych výsledkov realizujte minimálne 5-timi rôznymi spôsobmi (aspoň jeden cyklus, aspoň dva spôsoby cez funkčné programovanie).[1]

Sledujte aká časová a pamätová náročnosť pri každom spôsobe, ktorý ste naprogramovali je potrebná. Sledujte ako sa jednotlivé programátorské štýly správajú v tomto a aj v nasledujúcich algortimoch. Sledujte, kde sú ich slabé a silné stránky.

2. **„Priemer“ kladných hodnôt**

Vytvorte vektor náhodne generovaných **celých čísel z intervalu 0 až 10 s normálnym rozdelením** (napr. ***rnorm()***)– opäť postupne s dĺžkou 100 – 1 000 000 čísel. Vypočítajte aritmetický, geometrický, harmonický a kvadratický priemer z týchto čísel. Vymyslite aspoň 5 spôsobov, ako túto úlohu zrealizovať (napr. ***mean()***, cyklus, funkčné programovanie) pre každý z priemerov. Najmä pri väčšom dátovom objeme sledujte, či je z časového hľadiska lepšie vygenerovať celú množinu, uložiť ju, alebo ju generovať a počítať po dávkach – samozrejme s ohľadom na použitý algoritmus. Identifikujte slabé a silné stránky všetkých postupov.[2]

Aritmetický priemer- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

Geometrický priemer $\overline{x}_G = \sqrt[n]{x_1 . x_2 . \ldots . x_n}$

Je zrejmé, že geometrický priemer má zmysel iba pre dáta, v ktorých sú všetky hodnoty kladné čísla. Geometrický priemer sa na rozdiel od aritmetického priemeru používa na koeficienty, napr. na výpočet priemerného rastu: ak rast cien bol postupne 20 %, 10 %, potom 15 % pokles a 10 % rast, tak priemerný rast sa rovná ($\sqrt[4]{1,20 \cdot 1,10 \cdot 0,85 \cdot 1,10} \cong 1,054$ čiže priemerný rast je približne 5,4 %. Toto číslo vyjadruje, že výsledná cena by bola taká istá aj v prípade, ak by rast bol konštantný, každý rok 5,4 %

[1] Pre úplnosť – termín „rôzny spôsob“ sa chápe ako myšlienkovo úplne iný prístup ku danému problému. Do tejto kategórie nespadá postup, že použijem inú knižnicu a mierne iné príkazy na zrealizovanie toho istého zadania. Chápe sa tým procedurálne programovanie, objektové programovanie, funkcionállny prístup, conditional functions, pure functions, anonymous functions ....

[2] Zdroj http://math.ku.sk/data/portal/data/zbornik2007/Articles/Kulcar_Ladislav.pdf

Asi to veľmi rýchlo zistíte, že väčšej dátovej množine rýchlo narazíte na limity aritmetických operácií, takže jeden z možných nápadov pre veľké dátové množiny - pri použití logaritmov možno súčiny zmeniť na súčty a umocňovanie na súčin, $\exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right]$

Harmonický priemer $\overline{x}_H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}$

Harmonický priemer sa používa na určenie stredu variability takého znaku, ktorý cez určitú konštantnú hodnotu je v nepriamom vzťahu s iným znakom. Súčet takýchto hodnôt znaku nedáva logický zmysel. Harmonický priemer sa používa na charakterizovanie hodnôt, ktoré predstavujú napríklad výkonové limity – teda dosiahnuť u každej osoby ten istý výkon pri rôznom čase alebo rôzny výkon za jednotku času (1. osoba urobí prácu za hod, teda jej hodinový výkon je, …,atď.) V prípade rôznych vzdialeností a rovnakých časov sa však musí použiť aritmetický priemer.

Kvadratický priemer $\overline{x}_K = \sqrt{\frac{x_1^{\,2}+x_2^{\,2}+\ldots+x_n^{\,2}}{n}}$. Kvadratický priemer sa obyčajne používa vo fyzike, kde sa často označuje ako efektívna hodnota.

3. Zopakujte úlohu 2 **pre vektor náhodne generovaných reálnych čísel z intervalu** $(0,1)$**.** Kedy a v ktorých prípadoch budete pozorovať výraznú zmenu oproti predchádzajúcemu prípadu (spotreba pamäte, čas potrebný na výpočet...) Ako overíte, či sa sa na vypočítaný výsledok môžete spoľahnúť?

4. **RMSD – root mean square deviation s podmienkou**

Opäť vytvorte **vektor náhodne generovaných reálnych čísel z intervalu** $(-1,1)$ **s normálnym rozdelením** – opäť postupne s dĺžkou 100 – 1 000 000 čísel. ***Naučte sa v tomto príklade používať funkcionálnu podmienku!*** To znamená, že vymyslíte aspoň 5 spôsobom (jeden cyklus, aspoň dva rôzne funkcionálne prístupy) ako z vami vygenerovanej dátovej množiny vyberiete len kladné čísla (len záporné čísla, len čísla z nejakého intervalu ...) a pre tento výber vypočítate aká bude RMSD

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}}$$

Pre každý prístup (každý naprogramovaný spôsob) a každú dĺžku vektora testujte aj pamätové aj časové nároky a overte hraničné limity použitia jednotlivých programátorských štýlov.

5. Trochu genomiky - (práca s typom char)

Stiahnite si z drivu kompletnú genetickú informáciu pre covid 19 (NCBI Reference Sequence: NC_045512.2)[3] a urobte jeho analýzu pre jednoduché základné úlohy. Pre každú úlohu (ako obvykle) aspoň 5 rôznych spôsobov ako sa to dá naprogramovať a minimálne dva (tri?) funkcionálne

- Zistite, koľko obsahuje písmeno A, C, T, G.
- Zistite koľko krát sa v reťazci vyskytujú všetky základné aminokyseliny (Jednotlivé dusikaté bázy – písmená – navzájom utvárajú trojice. Každá trojica predstavuje kodón – jednu aminokyselinu. Spájaním kodónov, aminokyselín, vznikajú kódy - gény - pre funkčné bielkoviny.

Prvý nukleotid	Druhý nukleotid: U	Druhý nukleotid: C	Druhý nukleotid: A	Druhý nukleotid: G	Tretí nukleotid
U	fenylalanín	serín	tyrozín	cysteín	U
	fenylalanín	serín	tyrozín	cysteín	C
	leucín	serín	"koniec reťazca"	"koniec reťazca"	A
	leucín	serín	"zač. reťazca"	tryptofán	G
C	leucín	prolín	histidín	arginín	U
	leucín	prolín	histidín	arginín	C
	leucín	prolín	glutamín	arginín	A
	leucín	prolín	glutamín	arginín	G
A	izoleucín	treonín	asparagán	serín	U
	izoleucín	treonín	asparagán	serín	C
	izoleucín	treonín	lyzín	arginín	A
	metionín "zač.reť."	treonín	lyzín	arginín	G
G	valín	alanín	kys. asparágová	glycín	U
	valín	alanín	kys. asparágová	glycín	C
	valín	alanín	kys. glutámová	glycín	A
	valín "zač. reťazca"	alanín	kys. glutámová	glycín	G
	U	C	A	G	

Druhý nukleotid

[3] Ak vás zaujíma iná (nie základná mutácia, stiahnite si dáta z databázy https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049 , klik na meno, potom vľavo hore formát FASTA a vpravo hore sent to file