

UNIVERZITA KARLOVA

Přírodovědecká fakulta

Studijní obor: Sociální geografie a geoinformatika



Ester Kaliská

Úvod do programování

Skúškové zadanie 2

Praha 2026

1. Zadanie

Cieľom úlohy je overiť normalitu dátového súboru obsahujúceho 100 hodnôt, ktoré sú načítané z externého textového súboru. Na testovanie normality som použila Shapiro-Wilkov test, keďže je vhodný pre menšie a stredne veľké výbery a je bežne používaný v praxi.

2. Popis a rozbor problému

Overenie normality rozdelenia je dôležitým krokom v štatistickej analýze, pretože veľa parametrických metód predpokladá, že dáta majú normálne rozdelenie. Ak tento predpoklad nie je splnený, výsledky ďalších testov môžu byť skreslené alebo nesprávne.

V tejto úlohe som sa zamerala na Shapiro-Wilkov test, ktorý porovnáva usporiadané hodnoty výberu s hodnotami, ktoré by sa očakávali pri normálnom rozdelení. Výsledkom testu je testovacia štatistika a p-hodnota, na základe ktorej sa rozhoduje o platnosti nulovej hypotézy.

Ak je p-hodnota väčšia ako 0,05, nulovú hypotézu o normálnom rozdelení dát nezamietame. V opačnom prípade možno predpokladať, že dáta normálne rozdelenie nemajú.

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

V praxi je potrebné brať do úvahy, že samotný test normality nemusí vždy poskytnúť jednoznačný záver. Pri väčších výberoch môže byť Shapiro-Wilkov test citlivý aj na malé odchýlky od normálneho rozdelenia, ktoré nemusia mať významný vplyv na ďalšiu analýzu. Z tohto dôvodu sa testovanie normality často kombinuje s grafickými metódami, ako sú histogramy alebo Q-Q grafy, ktoré umožňujú vizuálne posúdenie tvaru rozdelenia dát.

3. Použitý algoritmus

Program je rozdelený do niekoľkých logických krokov. Najskôr sa rieši správne načítanie vstupného súboru, následne samotná štatistická analýza a nakoniec vyhodnotenie výsledkov.

Postup algoritmu je nasledovný:

- program zistí absolútnu cestu k priečinku, v ktorom sa nachádza skript, aby sa predišlo problémom s relatívnymi cestami
- používateľ zadá názov vstupného súboru
- súbor sa otvorí a číta po jednotlivých riadkoch
- každý riadok sa očistí od bielych znakov a prevedie na typ float

- v prípade chyby (neexistujúci súbor alebo nečíselná hodnota) sa vypíše chybové hlásenie
- po načítaní dát sa skontroluje, či výber obsahuje aspoň tri hodnoty
- následne sa pomocou funkcie shapiro() z knižnice scipy.stats vypočíta p-hodnota
- výsledok sa porovná s hladinou významnosti 0,05 a vypíše sa slovný záver

Zvolený postup algoritmu je navrhnutý tak, aby bolo možné jednotlivé kroky jednoducho kontrolovať a prípadne upravovať. Oddelenie načítania dát, ich kontroly a samotnej štatistickej analýzy zvyšuje prehľadnosť kódu a uľahčuje jeho údržbu. Tento prístup je vhodný najmä pri práci s externými dátami, kde môže dôjsť k rôznym neočakávaným chybám vo vstupoch.

4. Ošetrovanie problematických situácií

Pri tvorbe programu som sa snažila myslieť aj na situácie, ktoré by mohli viesť k chybe počas behu programu.

Konkrétne ide o:

- nesprávny pracovný adresár, ktorý je riešený pomocou `os.path.abspath(__file__)`
- neexistujúci vstupný súbor, zachytený výnimkou `FileNotFoundError`
- nečíselné alebo prázdne riadky v súbore, ktoré sú ošetrené pomocou `ValueError` a podmienky na preskočenie prázdnych riadkov
- nedostatočný počet dát, keďže Shapiro-Wilkov test vyžaduje minimálne tri hodnoty

Vďaka týmto kontrolám program nespadne, ale používateľ dostane zrozumiteľnú informáciu o probléme.

5. Vstupné dáta

Vstupom je externý textový súbor s príponou `.txt`, ktorý obsahuje jednu číselnú hodnotu na každom riadku. Desatinné čísla musia byť zapísané pomocou bodky. Odporúčaný počet hodnôt je približne 100, čo je postačujúca veľkosť výberu na testovanie normality.

Pre potreby zadania bol vygenerovaný pomocou umelej inteligencie súbor so 100 náhodnými hodnotami, nad ktorým bol program testovaný. Súbor `data.txt` je uložený v priečinku spolu s kódom.

6. Výstupné dáta

Výstupom programu je:

1. informácia o tom, či sa podarilo dáta úspešne načítať
2. vypočítaná p-hodnota zaokrúhlená na štyri desatinné miesta
3. slovná interpretácia výsledku, teda či dáta majú alebo nemajú normálne rozdelenie

Okrem samotnej p-hodnoty je dôležité správne interpretovať výsledok v kontexte riešeného problému. Ak program vyhodnotí, že dáta majú normálne rozdelenie, je možné v ďalších krokoch použiť parametrické štatistické metódy, ako napríklad t-test alebo analýzu rozptylu. V prípade, že normalita potvrdená nie je, je vhodné zvážiť použitie neparametrických metód alebo transformáciu dát. Výstup programu tak slúži ako podklad pre ďalšie štatistické spracovanie dát a nie ako konečný cieľ analýzy.

7. Programová dokumentácia

Program je implementovaný ako trieda "AnalyzaDat", ktorá zabezpečuje prehľadnosť a zapuzdrenie dát. Na štatistické výpočty je použitá knižnica "scipy.stats" a odtiaľ funkcia shapiro. Knižnica "os" slúži na prácu so súborovým systémom. Pri testovaní bol problém s nájdením súboru a cesty k súboru data.txt a využitie knižnice "os" tento problém efektívne rieši za pomoci funkcie "os.path.dirname" a "os.path.join". Spracovanie výnimiek pomocou blokov "try-except" zvyšuje stabilitu programu a zabraňuje jeho pádu pri chybnom vstupe. Ošetruje chybu hodnoty (ValueError), nenájdenný súbor (FileNotFoundError) a všeobecnú chybu (Exception). V bloku with open je taktiež ošetrená chyba, ak nie je v riadku žiadna hodnota. V tomto prípade sa riadok preskočí a program nespadne.

8. Možné rozšírenia riešenia

Program spĺňa zadanie, avšak do budúcnosti by bolo možné ho ďalej rozšíriť. Medzi možné vylepšenia patrí:

- doplnenie vizuálnej analýzy (histogram, Q-Q graf),
- zavedenie logovania chýb do externého súboru,
- podpora vstupných súborov vo formáte CSV,
- možnosť voľby medzi viacerými testami normality.

Takto navrhnutý program je možné využiť aj v iných úlohách, kde je potrebné rýchlo overiť základné štatistické vlastnosti dát. Môže slúžiť napríklad ako podporný nástroj pri spracovaní výsledkov meraní, dotazníkových prieskumov alebo experimentálnych dát. Vďaka modulárnej štruktúre je možné program jednoducho prispôbiť aj iným typom vstupných dát.