VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky

PRAVDĚPODOBNOST A STATISTIKA

Zadání 22

JMÉNO STUDENTKY/STUDENTA:

OSOBNÍ ČÍSLO:

JMÉNO CVIČÍCÍ/CVIČÍCÍHO:

	DATUM ODEVZDÁNÍ	HODNOCENÍ
DOMÁCÍ ÚKOL 1:		
DOMÁCÍ ÚKOL 2:		
DOMÁCÍ ÚKOL 3:		
DOMÁCÍ ÚKOL 4:		
CELKEM:	----------	

Ostrava, AR 2018/2019

Jméno: Číslo zadání: 22

Popis datového souboru

V datovém souboru *ukol_22.xlsx* jsou zaznamenány výpočetní časy (ms) třídících algoritmů Quicksort, Mergesort, Heapsort a Shellsort. Algoritmy byly opakovaně testovány na určitém počtu (ne nutně stejném pro všechny algoritmy) náhodně přeuspořádaných číselných řad (polí) délky 10^6 a to vždy nejdříve na počítači s méně výkonným procesorem a poté na počítači s výkonnějším procesorem. Vaším úkolem je porovnat mezi sebou výpočetní časy třídících algoritmů.

Obecné pokyny:

- Úkoly zpracujte dle obecně známých typografických pravidel.
- **Všechny** tabulky i obrázky musí být opatřeny titulkem.
- Do úkolů nevkládejte tabulky a obrázky, na něž se v doprovodném textu nebudete odkazovat.
- Bude-li to potřeba, citujte zdroje dle mezinárodně platné citační normy ČSN ISO 690.

Jméno: Číslo zadání: 22

Úkol 1

a) Popište strukturu datového souboru, tj. určete počty testovacích polí dle použitého třídícího algoritmu. Použijte tabulku četností a výsledky vhodným způsobem vizualizujte.

Pomocí nástrojů explorační analýzy srovnejte výpočetní časy zjištěné na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort. Data vhodně graficky prezentujte (krabicový graf, histogram, q-q graf) a doplňte následující tabulky a text.

Tab. 1: Výběrové charakteristiky výpočetních časů zjištěných na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort

Výpočetní čas (ms) na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort – výběrové charakteristiky			Po odstranění odlehlých pozorování	
	Algoritmus Quicksort	Algoritmus Heapsort	Algoritmus Quicksort	Algoritmus Heapsort
rozsah souboru				
Míry polohy				
minimum				
dolní kvartil				
medián				
průměr				
horní kvartil				
maximum				
Míry variability				
směrodatná odchylka				
variační koeficient (%)				
Míry šikmosti a špičatosti				
šikmost				
špičatost				

Identifikace odlehlých pozorování – vnitřní hradby		
dolní mez		
horní mez		

Jméno: Číslo zadání: 22

Grafická prezentace (krabicový graf, histogram, q-q graf):

Jméno: Číslo zadání: 22

Analýza výpočetního času testovaného na počítači s výkonnějším procesorem pro algoritmus Quicksort

Během testu byl na počítači s výkonnějším procesorem změřen výpočetní čas třídícího algoritmu Quicksort pro ..... číselných řad. Změřený výpočetní čas se pohyboval v rozmezí …….. až ……. ms. Hodnoty výpočetního času ležící mimo interval ..... až ..... ms (vnitřní hradby) byly identifikovány jako odlehlá pozorování a nebudou zahrnuty do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: ..... / Žádné z měření nebylo identifikováno jako odlehlé pozorování. Dále uvedené výsledky tedy pocházejí z analýzy výpočetních časů pro …… polí. Průměrný výpočetní čas byl …….. ms, směrodatná odchylka pak …… ms. U poloviny z polí výpočetní čas nepřekročil …….. ms. V polovině měření se výpočetní čas pohyboval v rozmezí …… až ..... ms. Vzhledem k hodnotě variačního koeficientu (…….%) lze / nelze analyzovaný soubor považovat za homogenní.

Analýza výpočetního času testovaného na počítači s výkonnějším procesorem pro algoritmus Heapsort

Během testu byl na počítači s výkonnějším procesorem změřen výpočetní čas třídícího algoritmu Heapsort pro ..... číselných řad. Změřený výpočetní čas se pohyboval v rozmezí …….. až ……. ms. Hodnoty výpočetního času ležící mimo interval ..... až ..... ms (vnitřní hradby) byly identifikovány jako odlehlá pozorování a nebudou zahrnuty do dalšího zpracování. Možné příčiny vzniku odlehlých pozorování jsou: ..... / Žádné z měření nebylo identifikováno jako odlehlé pozorování. Dále uvedené výsledky tedy pocházejí z analýzy výpočetních časů pro …… polí. Průměrný výpočetní čas byl …….. ms, směrodatná odchylka pak …… ms. U poloviny z polí výpočetní čas nepřekročil …….. ms. V polovině měření se výpočetní čas pohyboval v rozmezí …… až ..... ms. Vzhledem k hodnotě variačního koeficientu (…….%) lze / nelze analyzovaný soubor považovat za homogenní.

Ověření normality výpočetního času testovaného na počítači s výkonnějším procesorem pro algoritmus Quicksort (na základě explorační analýzy)

Na základě grafického zobrazení (viz ……..) a výběrové šikmosti a špičatosti (viz Tab. 1, výběrová šikmost i špičatost leží / neleží v intervalu $(-2; 2)$) lze / nelze předpokládat, že výpočetní čas algoritmu Quicksort změřený na počítači s výkonnějším procesorem má normální rozdělení. Dle pravidla 3σ / Čebyševovy nerovnosti lze tedy očekávat, že přibližně pro 95 % / pro více než 75 % testových úloh bude pro algoritmus Quicksort spuštěný na počítači s výkonnějším procesorem zjištěn výpočetní čas v rozmezí …….. až ……..ms.

Ověření normality výpočetního času testovaného na počítači s výkonnějším procesorem pro algoritmus Quicksort (na základě explorační analýzy)

Na základě grafického zobrazení (viz ……..) a výběrové šikmosti a špičatosti (viz Tab. 2, výběrová šikmost i špičatost leží / neleží v intervalu $(-2; 2)$) lze / nelze předpokládat, že výpočetní čas algoritmu Heapsort změřený na počítači s výkonnějším procesorem má normální rozdělení. Dle pravidla 3σ / Čebyševovy nerovnosti lze tedy očekávat, že přibližně pro 95 % / pro více než 75 % testových úloh bude pro algoritmus Heapsort spuštěný na počítači s výkonnějším procesorem zjištěn výpočetní čas v rozmezí …….. až ……..ms.

Jméno: Číslo zadání: 22

Úkol 2

Porovnejte výpočetní časy zjištěné na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort a rozdíly výpočetních časů zjištěné na počítači s výkonnějším procesorem a počítači s méně výkonným procesorem bez ohledu na typ algoritmu. Nezapomeňte, že použité metody mohou vyžadovat splnění určitých předpokladů. Pokud tomu tak bude, okomentujte splnění/nesplnění těchto předpokladů jak na základě explorační analýzy (např. s odkazem na histogram apod.), tak exaktně pomocí metod statistické indukce.

Nejdříve porovnejte **výpočetní časy zjištěné na počítači s výkonnějším procesorem** pro **algoritmy Quicksort a Heapsort**.

a) Vraťte se ke grafické prezentaci z úkolu 1 a vytvořte si úsudek o srovnání výpočetních časů zjištěných na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort.

b) Určete bodové a 95% intervalové odhady pro střední hodnoty (resp. mediány) výpočetních časů zjištěných na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort. Výsledky interpretujte!

c) Určete bodový a 95% intervalový odhad rozdílu středních hodnot (resp. mediánů) výpočetních časů zjištěných na počítači s výkonnějším procesorem pro algoritmy Quicksort a Heapsort. Výsledek interpretujte!

d) Na hladině významnosti 5 % rozhodněte, zda střední hodnota (resp. medián) výpočetního času zjištěného na počítači s výkonnějším procesorem pro algoritmus Quicksort je statisticky významně nižší než u algoritmu Heapsort, tj. zda je algoritmus Quicksort statisticky významně rychlejší než algoritmus Heapsort.

Dále se zaměřme na **srovnání výpočetních časů zjištěných na počítači s výkonnějším procesorem a počítači s méně výkonným procesorem bez ohledu na typ algoritmu**.

e) Vizualizujte srovnání výpočetních časů zjištěných na počítači s výkonnějším procesorem a počítači s méně výkonným procesorem **bez ohledu na typ algoritmu** a vytvořte si úsudek o pozorovaných rozdílech.

f) Určete bodový a 95% intervalový odhad střední hodnoty (resp. mediánu) rozdílů výpočetních časů zjištěných na počítači s výkonnějším procesorem a počítači s méně výkonným procesorem **bez ohledu na typ algoritmu**. Výsledek interpretujte!

g) Na hladině významnosti 5 % rozhodněte, zda jsou výpočetní časy zjištěné na počítači s výkonnějším procesorem statisticky významně nižší než na počítači s méně výkonným procesorem (**bez ohledu na typ algoritmu**).

Jméno: Číslo zadání: 22

Úkol 3

Na hladině významnosti 5 % rozhodněte, zda se střední hodnoty (resp. mediány) výpočetního času na počítači s výkonnějším procesorem liší v závislosti na použitém třídícím algoritmu. Uvažujte všechny použité třídící algoritmy. Posouzení proveďte nejprve na základě explorační analýzy a následně pomocí vhodného statistického testu včetně ověření potřebných předpokladů. V případě, že se střední hodnoty (resp. mediány) výpočetního času jednotlivých třídících algoritmů statisticky významně liší, určete, které výpočetní časy se statisticky významně odlišují od ostatních, tj. určete homogenní podskupiny třídících algoritmů.

a) Daný problém vhodným způsobem graficky prezentujte (vícenásobný krabicový graf, histogramy, q-q grafy).

b) Pro všechny čtyři třídící algoritmy ověřte normalitu výpočetních časů na počítači s výkonnějším procesorem (empiricky i exaktně).

c) Ověřte homoskedasticitu (shodu rozptylů) výpočetních časů na počítači s výkonnějším procesorem jednotlivých třídících algoritmů (empiricky i exaktně).

d) Určete bodové a 95% intervalové odhady střední hodnoty výpočetního času (resp. mediánu výpočetního času) pro všechny srovnávané třídící algoritmy. (Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.)

e) Čistým testem významnosti ověřte, zda existuje statisticky významný rozdíl ve výpočetním času testovaných třídících algoritmů. Pro posouzení srovnejte střední hodnoty výpočetních časů (resp. mediány výpočetních časů) pro počítač s výkonnějším procesorem pro všechny třídící algoritmy. Pokud existuje statisticky významný rozdíl (na hladině významnosti 5 %), zjistěte, zda lze některé skupiny třídících algoritmů označit (z daného hlediska) za homogenní, tj. stanovte pořadí třídících algoritmů dle výpočetních časů. Nezapomeňte na ověření předpokladů pro použití zvoleného testu.

Úkol 4

V tomto úkolu opět analyzujte pouze výpočetní časy změřené na počítači s výkonnějším procesorem. Výpočetní časy kategorizujte do dvou variant. Jako „rychlé“ označte časy kratší než 500 ms, ostatní výpočetní časy označte jako „pomalé“.

a) Zjistěte, zda délka výpočetního času („rychlé“ vs. „pomalé“ časy) závisí na typu použitého algoritmu. Výsledky prezentujte pomocí kontingenční tabulky, vhodného grafu a vhodné míry kontingence. Vytvořte úsudek o pozorované závislosti.

b) Pomocí chí-kvadrát testu nezávislosti rozhodněte na 5% hladině významnosti, jestli to, zda výpočetní čas patří do skupiny „rychlé“ nebo „pomalé“ závisí na volbě třídícího algoritmu. (Nezapomeňte na ověření předpokladů testu.)

Dále se zabývejte pouze srovnáním algoritmů Quicksort a Heapsort.

c) Určete bodový a 95% intervalový odhad rizika (tj. pravděpodobnosti), že použití algoritmu Heapsort si vyžádá „pomalý“ výpočetní čas. Nezapomeňte na ověření předpokladu pro použití příslušného intervalového odhadu. Totéž určete i pro algoritmus Quicksort.

d) Určete bodový a 95% intervalový odhad relativního rizika, „pomalého“ výpočetního času u algoritmu Heapsort vůči algoritmu Quicksort. Výsledky slovně interpretujte.

e) Určete bodový odhad šance, že použití algoritmu Heapsort si vyžádá „pomalý“ výpočetní čas. Totéž určete i pro algoritmus Quicksort. Výsledky slovně interpretujte.

f) Určete bodový a 95% intervalový odhad poměru šancí, „pomalého“ výpočetního času u algoritmu Heapsort vůči algoritmu Quicksort. Výsledky slovně interpretujte.

Jak identifikovat, zda jsou v datech odlehlá pozorování?

Emiprické posouzení:

- *použití vnitřních (vnějších) hradeb, resp. $z-$ souřadnice, resp. mediánová suřadnice,*
- *vizuální posouzení krabicového grafu.*

Exaktní posouzení:

- *Grubbsův test (parametrický test - vyžaduje normalitu dat)*
- *Deanův - Dixonův test (neparametrický test)*

Jak naložit s odlehlými hodnotami by měl definovat hlavně zadavatel analýzy (expert na danou problematiku).

Jak ověřit normalitu dat?

Emiprické posouzení:

- *vizuální posouzení histogramu,*
- *vizuální posouzení grafu odhadu hustoty pravděpodobnosti,*
- *Q-Q graf,*
- *P-P graf,*
- *posouzení výběrové šikmosti a výběrové špičatosti.*

Exaktní posouzení:

- *testy normality (např. Shapirův – Wilkův test, Andersonův-Darlingův test, Lillieforsův test, …)*

Jak ověřit homoskedasticitu (shodu rozptylů)?

Emiprické posouzení:

- *poměr největší a nejmenší směrodatné odchylky,*
- *vizuální posouzení krabicového grafu.*

Exaktní posouzení:

- *F – test (parametrický dvouvýběrový test),*
- *Bartlettův test (parametrický vícevýběrový test),*
- *Leveneův test (neparametrický test).*