**VŠB – Technická univerzita Ostrava**

**Fakulta elektrotechniky a informatiky**

**pravděpodobnost a statistika**

**Domácí úkoly – Zadání 12**

|  |  |  |  |
| --- | --- | --- | --- |
| 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 2017/2018**

**Popis datového souboru**

V datovém souboru *pripojeni\_12.xlsx* najdete výsledky měření rychlosti internetového připojení u zákazníků, kteří si vyžádali servisní prohlídku. U každého servisního zásahu jsou uvedeny rychlosti připojení (Mb/s) před příchodem technika a rychlosti připojení po servisním zásahu. Také je uvedeno o jaký typ připojení se jednalo (ADSL/KABEL/OPTIKA) a geografické zařazení (město-centrum/město-okraj/venkov).

**Obecné pokyny:**

* Portfolio domácích úkolů budete odevzdávat postupně. Tj. nejdříve odevzdáte titulní stránku + úkol 1, následně doplníte úkol 2, atd.
* Jednotlivé domácí úkoly odevzdávejte vždy v termínu, který určil váš cvičící.
* Domácí úkoly zpracujte dle obecně známých typografických pravidel.
* **Všechny** tabulky i obrázky musí být opatřeny titulkem.
* Do domácích 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.

**Úkol 1**

1. Prezentujte strukturu datového souboru, tj. strukturu zákazníků, kteří si vyžádali servisní prohlídku dle typu připojení (ADSL/KABEL/OPTIKA) a geografické polohy (město-centrum/město-okraj/venkov). Použijte tabulku sdružených četností a výsledky vhodným způsobem vizualizujte.
2. Srovnejte zvýšení rychlosti připojení po zásahu servisního technika pro připojení typu KABEL a OPTIKA na základě explorační analýzy, data graficky prezentujte (histogram, vícenásobný krabicový graf) a doplňte následující tabulku a text.

Tab. 1: Výběrové charakteristiky zvýšení rychlosti připojení po zásahu servisního technika pro připojení typu KABEL a OPTIKA

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Zvýšení rychlosti připojení po zásahu servisního technika (Mb/s)** | | | |  | **Po odstranění odlehlých pozorování** | | |
|  | **KABEL** | **OPTIKA** | | **KABEL** | **OPTIKA** | |
| **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** |  |  | |  |  |  |  |

**Grafická prezentace sledovaného statistického znaku (vícenásobný krabicový graf + histogramy, resp. odhady hustoty pravděpodobnosti, q-q grafy):**

**Zvýšení rychlosti připojení po zásahu servisního technika pro připojení typu KABEL**

Byly analyzovány záznamy servisního technika a z nich určeno zvýšení rychlosti připojení po jeho zásahu u …………. zákazníků s připojením typu KABEL (viz Tab. 1). Hodnoty zvýšení rychlosti připojení se pohybovaly v rozmezí …………… až …………………. Mb/s. Hodnoty zvýšení rychlosti připojení ležící mimo interval ……………… až ………………….. Mb/s byly identifikovány jako odlehlá pozorování a příslušná měření byla z dalšího zpracování vyřazena. / Žádné z měření nebylo identifikováno jako odlehlé pozorování. Níže uvedené výsledky pocházejí z analýzy datového souboru o rozsahu ………………. zákazníků.

Průměrné pozorované zvýšení rychlosti připojení bylo …………….. Mb/s, směrodatná odchylka ………………….. Mb/s. Polovině zákazníků bylo zjištěno zvýšení rychlosti připojení nižší než …………….. Mb/s. (Podrobněji: U čtvrtiny zákazníků s připojením KABEL bylo zjištěno zvýšení rychlosti připojení nižší než ………………… Mb/s, u čtvrtiny zákazníků bylo zjištěno zvýšení rychlosti připojení vyšší než ………………… Mb/s.) Vzhledem k hodnotě variačního koeficientu (……….%) lze / nelze analyzovaný soubor považovat za homogenní.

**Ověření normality zvýšení rychlosti připojení po zásahu servisního technika pro připojení typu KABEL 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 ) lze / nelze předpokládat, že zvýšení rychlosti připojení zákazníků s připojením KABEL má normální rozdělení. Dle pravidla 3 / Čebyševovy nerovnosti lze tedy očekávat, že přibližně u 95% / více než 75% zákazníků s připojením KABEL lze očekávat zvýšení rychlosti připojení v rozmezí …………..……………. až ……………..……… Mb/s.

Obdobně lze popsat výsledky analýzy rychlosti připojení zákazníků s připojením typu OPTIKA.

**Úkol 2**

Porovnejte zvýšení rychlosti připojení po zásahu servisního technika u zákazníků s připojením typu KABEL a OPTIKA. 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.

1. Vraťte se ke grafické prezentaci z úkolu 1 a vytvořte si úsudek o nárůstu rychlosti připojení u obou skupin zákazníků.
2. Určete 95% intervalové odhady středního nárůstu rychlosti připojení (resp. mediánu nárůstu rychlosti připojení) pro obě skupiny zákazníků (KABEL a OPTIKA). (Nezapomeňte na empirické i exaktní ověření předpokladů pro použití příslušných intervalových odhadů.)
3. Pro obě skupiny zákazníků (KABEL, OPTIKA) ověřte na hladině významnosti 5 %, zda je pozorovaný nárůst rychlosti připojení statisticky významný. (Nezapomeňte na empirické i exaktní ověření předpokladů vybraného testu.)
4. Pokud je to možno, určete 95% intervalový odhad rozdílu středních hodnot nárůstu rychlosti připojení u zákazníků s připojením typu KABEL a OPTIKA. (Nezapomeňte na empirické i exaktní ověření předpokladů pro použití příslušného intervalového odhadu.)
5. Čistým testem významnosti ověřte, zda je pozorovaný rozdíl průměrných nárůstů rychlosti připojení (resp. mediánů nárůstů rychlosti připojení) u zákazníků s připojením typu KABEL a OPTIKA statisticky významný na hladině významnosti 5%. (Nezapomeňte na empirické i exaktní ověření předpokladů vybraného testu.)

**Úkol 3**

Srovnejte rychlost připojení před zásahem servisního technika v závislosti na typu připojení: ADSL, KABEL a OPTIKA pomocí exaktních metod statistické indukce.

1. Ověřte normalitu rychlosti připojení pro všechny typy připojení (empiricky i exaktně).
2. Ověřte homoskedasticitu (shodu rozptylů) rychlosti připojení pro všechny typy připojení (empiricky i exaktně).
3. Určete 95% intervalové odhady střední hodnoty (resp. mediánu) rychlosti připojení pro zákazníky s připojením typu ADSL, KABEL a OPTIKA. (Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.)
4. Čistým testem významnosti ověřte, zda je rozdíl průměrných rychlostí připojení (resp. mediánů rychlostí připojení) před zásahem servisního technika pro jednotlivé typy připojení statisticky významný na hladině významnosti 5 %. Pokud ano, zjistěte, zda lze některé typy připojení označit (z hlediska jejich rychlosti připojení před zásahem servisního technika) za homogenní. (Nezapomeňte na ověření předpokladů pro použití zvoleného testu.)

**Úkol 4**

Analyzujte vliv městské části (centrum/okraj) na výskyt vysokorychlostního připojení (KABEL/OPTIKA).

1. Pomocí vhodné asociační tabulky prezentujte vliv života zákazníka v určité městské části na typ jeho vysokorychlostního připojení. Tabulku doplňte o řádkové relativní četnosti.
2. Analyzovanou závislost prezentujte pomocí vhodného grafu.
3. Určete alespoň jednu míru kontingence. Na základě posouzení vizualizace analyzované závislosti a nalezené míry kontingence posuďte míru analyzované závislosti.
4. Určete šanci, že náhodně vybraný zákazník s vysokorychlostním internetem z města-centrum, nebo města-okraj má připojení typu OPTIKA, tj. šance, že zákazník s vysokorychlostním internetem z města-centrum má připojení typu OPTIKA a šance, že zákazník s vysokorychlostním internetem z města-okraj má připojení typu OPTIKA. (Komentujte.)
5. Určete relativní šanci, že zákazník s vysokorychlostním internetem z města-centrum má připojení typu OPTIKA (včetně 95% intervalového odhadu) vzhledem k zákazníkovi z okraje města. Na základě svého zjištění určete, zda lze vliv místa připojení považovat za statisticky významný na hladině významnosti 5%.
6. Servisní technik Dan má znalost pouze připojení typu ADSL a KABEL. Určete relativní riziko, že Dan nebude schopen závadu opravit (tedy, že zákazník má připojení typu OPTIKA), bude-li vyslán k zákazníkovi do oblasti město – centrum vzhledem k tomu, bude-li vyslán do oblasti město - okraj. Na základě svého zjištění určete, zda lze vliv bydliště zákazníka (město-centrum/město-okraj) na typ připojení (OPTIKA a ADSL/KABEL) považovat za statisticky významný na hladině významnosti 5 %.
7. Ověřte, zda existuje závislost mezi výskytem typu připojení (ADSL/KABEL/OPTIKA) a bydlištěm zákazníka (město-centrum/město-okraj/venkov) pomocí Chí-kvadrát testu nezávislosti. Nezapomeňte ověřit předpoklady pro použití testu.

## Jak identifikovat, zda jsou v datech odlehlá pozorování?

Emiprické posouzení:

* použití vnitřních (vnějších) hradeb, resp. , resp. ,
* 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).