**VŠB – Technická univerzita Ostrava**

**Fakulta elektrotechniky a informatiky**

**Biostatistika**

**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 2018/2019**

**Popis datového souboru**

Kachexie označuje stav silné celkové sešlosti jedince spojené s úbytkem váhy a výraznou slabostí. Většinou je následkem vážných nádorových onemocnění nebo těžkých infekcí. V tomto kontextu vědci často zkoumají složení krve v závislosti na onemocnění pacienta a na výskytu kachexie.

V datovém souboru *kachexie\_2.xlsx* najdete údaje o onkologických pacientech s rakovinou plic s kachexií a bez projevů kachexie. Zároveň jsou v datech údaje i o třetí skupině – a to kontrolní skupině zdravých pacientů, tj. bez rakoviny i bez kachexie. U všech pacientů byla měřena hladina hormonu rezistinu (ng/ml).

U onkologických pacientů byla hladina rezistinu změřena při vstupní prohlídce a poté 3 měsíce po chemoterapii. U pacientů z kontrolní skupiny byl rezistin změřen pouze při vstupní prohlídce.

Skupina onkologických pacientů s kachexií je v datech kódována jako „RK“. Skupina onkologických pacientů bez kachexie je kódována „RBK“ a kontrolní skupina jako „KONT“. U onkologických pacientů je navíc uvedeno stádium rakoviny (na škále 0, I, II, III, IV).

**Obecné pokyny:**

* Domácí úkoly odevzdávejte vždy v termínu, který určil váš cvičící.
* Portfolio domácích úkolů budete odevzdávat postupně. Tj. nejdříve odevzdáte titulní stránku + úkol 1, následně doplníte úkol 2, atd.
* 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 pacientů dle jejich zdravotního stavu (onkologičtí pacienti s kachexií, onkologičtí pacienti bez kachexie, kontrolní skupina). Použijte tabulku četností a výsledky vhodným způsobem vizualizujte.
2. Na základě explorační analýzy srovnejte hodnoty vzrůstu rezistinu během doby mezi vstupní prohlídkou a prohlídkou po 3 měsících od chemoterapie u onkologických pacientů s kachexií a bez kachexie. Data graficky prezentujte (histogram, vícenásobný krabicový graf) a doplňte následující tabulky a text.

Tab. 1: Výběrové charakteristiky vzrůstu rezistinu během doby mezi vstupní prohlídkou a prohlídkou po 3 měsících od chemoterapie u onkologických pacientů s kachexií a bez projevů kachexie

**o**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Vzrůst rezistinu (ng/ml)** | | | |  | **Po odstranění odlehlých pozorování** | | |
|  | **onkologičtí pacienti**  **s kachexií** | **onkologičtí pacienti bez projevů kachexie** | | **onkologičtí pacienti**  **s kachexií** | **onkologičtí pacienti bez projevů kachexie** | |
| **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 (histogramy, vícenásobný krabicový graf):**

**Onkologičtí pacienti s kachexií**

Byly analyzovány záznamy o vzrůstu rezistinu …………. onkologických pacientů, u nichž se projevila kachexie (viz Tab. 1). Hodnoty vzrůstu rezistinu se pohybovaly v rozmezí …………… až …………… ng/ml. Hodnoty vzrůstu hladiny rezistinu ležící mimo interval ……………… až ………………….. ng/ml byly identifikovány jako odlehlá pozorování a příslušní pacienti byli z dalšího zpracování vyřazeni. / Žádné z měření nebylo identifikováno jako odlehlé pozorování. Níže uvedené výsledky pocházejí z analýzy datového souboru o rozsahu ………………. pacientů.

Průměrný vzrůst rezistinu byl …………….. ng/ml, směrodatná odchylka ………………….. ng/ml. U poloviny pacientů byl pozorován vzrůst rezistinu o více než …………….. ng/ml a u čtvrtiny pacientů byl pozorován vzrůst rezistinu o více než ………………… ng/ml. Vzhledem k hodnotě variačního koeficientu (……….%) lze / nelze hodnoty rezistinu v této skupině pacientů považovat za homogenní.

Obdobně lze popsat výsledky analýzy vzrůstu hladiny rezistinu u onkologických pacientů bez projevů kachexie.

**Ověření normality vzrůstu hladiny rezistinu u onkologických pacientů s projevy kachexie 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 vzrůst rezistinu u onkologických pacientů s kachexií má normální rozdělení. Dle pravidla 3 / Čebyševovy nerovnosti lze tedy očekávat, že u 95% / více než 75% onkologických pacientů s kachexií bude zaznamenán vzrůst hladiny rezistinu v rozmezí …………..……………. až ……………..……… ng/ml.

**Úkol 2**

Porovnejte vzrůst hladiny rezistinu u onkologických pacientů s kachexií a bez projevů kachexie. 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 srovnání vzrůstu rezistinu u onkologických pacientů s kachexií a bez projevů kachexie.
2. Určete bodové a 95% intervalové odhady středního vzrůstu rezistinu (resp. mediánu vzrůstu rezistinu) pro obě skupiny onkologických pacientů.
3. U obou skupin onkologických pacientů ověřte čistým testem významnosti, zda je pozorovaný vzrůst rezistinu statisticky významný na hladině významnosti 5 %.
4. Určete bodový a 95% intervalový odhad rozdílu středních hodnot (resp. mediánů) vzrůstů rezistinu srovnávaných skupin pacientů.
5. Na hladině významnosti 5 % rozhodněte, zda je rozdíl středních hodnot (resp. mediánů) vzrůstů rezistinu u srovnávaných skupin pacientů statisticky významný.

**Úkol 3**

Na hladině významnosti 5 % rozhodněte, zda se statisticky významně liší střední hodnoty (resp. mediány) hladiny rezistinu měřeny při vstupní prohlídce mezi jednotlivými skupinami pacientů. Nyní uvažujte všechny skupiny (tj. onkologické pacienty s kachexií, onkologické pacienty bez kachexie a kontrolní skupinu). Posouzení proveďte nejprve na základě explorační analýzy a následně pomocí vhodného statistického testu, včetně ověření potřebných předpokladů.

1. Daný problém vhodným způsobem graficky prezentujte (vícenásobný krabicový graf, histogramy, q-q grafy).
2. Ověřte normalitu hladin rezistinu všech tří skupin pacientů empiricky i exaktně.
3. Ověřte homoskedasticitu (shodu rozptylů) hladin rezistinu srovnávaných skupin pacientů (empiricky i exaktně).
4. Určete bodové i 95% intervalové odhady středních hodnot (resp. mediánů) hladin rezistinu u všech srovnávaných skupin pacientů. (Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.)
5. Na hladině významnosti 5 % ověřte, zda je pozorovaný rozdíl středních hodnot (resp. mediánů) hladin rezistinu u srovnávaných skupin pacientů statisticky významný. Pokud ano, zjistěte, zda lze některé skupiny pacientů označit (z hlediska jejich hladiny rezistinu) za homogenní. (Nezapomeňte na ověření předpokladů pro použití zvoleného testu.)

**Úkol 4**

Analyzujte asociaci mezi stádiem rakoviny a výskytem kachexie u onkologických pacientů. Uvažujte dvě skupiny pacientů podle stadia rakoviny – skupinu pacientů s nižším stadiem rakoviny (0 až II) a skupinu s vyšším stadiem rakoviny (III a IV). Potřebné údaje získejte ze záznamů o onkologických pacientech s kachexií a bez projevů kachexie.

1. Pomocí vhodné asociační tabulky prezentujte vliv stádia rakoviny na výskyt kachexie. 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 na výskyt kachexie u obou skupin pacientů, tj. u pacientů se stadiem rakoviny 0 až II a pacientů se stadiem rakoviny III až IV. (Komentujte, tj. vysvětlete praktický význam nalezených číselných hodnot šancí.)
5. Určete relativní šanci (včetně 95% intervalového odhadu) na výskyt kachexie u pacientů s vyšším stadiem rakoviny vzhledem k pacientům s nižším stádiem rakoviny. Na základě svého zjištění určete, zda lze vliv stadia rakoviny na výskyt kachexie považovat za statisticky významný na hladině významnosti 5 %.
6. Určete riziko (tj. pravděpodobnost) výskytu kachexie (včetně 95% intervalového odhadu) u obou skupin pacientů, tj. u pacientů se stadiem rakoviny 0 až II a pacientů se stadiem rakoviny III a IV. Nezapomeňte na ověření předpokladů pro použití příslušných intervalových odhadů.
7. Určete relativní riziko (včetně 95% intervalového odhadu) výskytu kachexie u skupiny pacientů s vyšším stadiem rakoviny vzhledem k pacientům s nižším stádiem rakoviny. Na základě svého zjištění určete, zda lze vliv stadia rakoviny na výskyt kachexie považovat za statisticky významný na hladině významnosti 5%.
8. Pomocí Chí-kvadrát testu nezávislosti ověřte na hladině významnosti 5 %, zda existuje statisticky významná závislost mezi stadiem rakoviny a výskytem kachexie. Nezapomeňte ověřit předpoklady pro použití testu.

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

*Empirické posouzení:*

* *použití vnitřních (vnějších) hradeb*
* *vizuální posouzení krabicového grafu.*

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

## Jak ověřit normalitu dat?

*Empirické 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)?

*Empirické 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).*