Posudek oponenta bakalářké práce

Název: Korelačné koeficienty a ich použitie
Autor: Roman Lacuš

Shrnutí obsahu práce

Bakalárska práca študenta Romana Lacuša pojednáva o troch typoch korelačných koeficientoch a ich praktickom využití pri testovaní nezávislosti dvoch náhodných veličín. V úvode autor stručne popisuje základnú matematickú/štatistickú teóriu nutnú k definovaniu Pearsonovho, Spearmanovho, alebo Kendallovho korelačného koeficientu a uvádza niektoré základné štatistické vlastnosti príslušných empirických odhadov potrebných pre konštrukciu štatistického testu.

Jedná sa už o druhú revíziu pôvodnej bakalárskej práce (s identickým názvom), ktorú autor neúspešne obhajoval v septembri 2019 a následne v septembri 2020. Aktuálne predložená práca sa ale žial len minimálne líši od predchádzajúcich dvoch verzií. Autor zapracoval len niektoré z explicitne uvedených pripomienok v posudkoch, pričom hlavným autorovým nástrojom pri zapracovaní jednotlivých pripomienok bolo v zásade len odstránenie problematickej vety, alebo pasáže, ale samotný problém zostal naďalej zachovaný. Autor nedokázal úspešne zapracovať ani jednoduchú pripomienku ohľadom formálneho zarovnania textu do blokov, alebo opakujúcu sa poznámku o nekonzistentnom značení—napríklad v prípade strednej hodnoty (E vs. $\boldsymbol{E}$).

Z teoretického hľadiska v práci pretrváva základný problém správne rozlíšiť medzi teoretickým modelom a empirickými pozorovaniami, medzi náhodným výberom, náhodným vektorom a náhodnou veličinou, alebo medzi náhodnou a nenáhodnou veličinou, prípadne všeobecným až neurčitým pojmom "data". V práci sa objavuje pomerne značné množstvo nesprávných a nezmyselných formulácii (z logického aj gramatického pohľadu) a na mnohých miestach pôsobí práca dojmom, že autor netuší, o čo sa vpodstate jedná.

Z formálneho hľadiska je práca hlboko podpriemerná. Okrem nezarovnaného textu a nekonzistentného značenia je v práci množstvo preklepov a gramatických chýb a paradoxne aj takých, ktoré sa v prvej, či druhej verzii práce vôbec nevyskytovali.

Celkovo považujem prácu za veľmi slabú a osobne nerozumiem autorovmu prístupu a vpodstate nulovému prejavenému záujmu o pochopenie nedostatkov explicitne a opakovane formulovaných v predchádzajúcich dvoch posudkoch. Predložená práca na mňa pôsobí skôr ako draft bakalárskej práce pred finalizáciou, rozhodne nie ako finalizovaná práca po druhej revízii.

Predloženú prácu preto nedoporučujem komisii uznať ako bakalársku prácu na MFF UK.

Praha, 24.01.2022

Matúš Maciak
maciak@karlin.mff.cuni.cz

- **Nesprávne/nelogické formulácie:** "*analogicky rovnosť platí práve vtedy*" (str.3); "*kovariácia spĺňa defníciu*" (str.4); "*hodnotu $r_p = 1$ nadobúda práve vtedy, keď ... leží na priamke*" (str.5); Veta 1.2 nedáva zmysel ako celok (str.6); "*dve regresívne rovnice priamok*" (str.6); "*tangens ∞ je rovný 90°*" (str.7); "*realizácia stredných hodnôt*" (str.7); "*náhodný výber X_n*" (str.8); "*u je kvantilová funkcia*" (str.9); "*poradia náhodného výberu $(X_1, \ldots, X_n)$ a $(Y_1, \ldots, Y_n)$*" (str.12); "*Nech $R_1, \ldots, R_n$ je usporiadané poradie...*" (str.12); "*korelácia v náhodnom výbere bude vysoká*" (str.14); "*aplikovanie odhadov korelačných koeficientov na rovnaký typ dát*" (str.16); "*pridáme 5 hodnôt, ktoré budú $90, 92, 95, 97$ a $99\,\%$ kvantilom danho dvojrozmerného normálneho rozdelenia*" (str.18); "*tak, aby $P[X < x_\alpha, Y < y_\alpha] \doteq 1 - \alpha$*" (str.18); "*po pridaní odľahlých koeficientov*" (str.20); "*data sú nezávislé*" (str.23);

- **Náhodné vs. nenáhodné veličiny:** Napr. na str.5 sú uvedené vzťahy (1.18) a (1.19). Koeficienty β_0 a β_1 sú síce neznáme, ale taktiež nenáhodné kvantity (parametre). Následujú ale rovnosti, za ktorými sa objavujú v sumách náhodne veličiny—resp. realizácie náhodnej veličiny Y (hoci hodnoty $y_1, \ldots, y_n$ nie sú formálne nikde zavedené). V tejto súvislosti nie je jasné, či hodnoty $x_1, \ldots, x_n$ sú myslené ako náhodné, alebo nenáhodné veličiny. Na pravej strane rovnosti (1.18) a (1.19) sa ale opäť objavujú teoretické charakteristiky—t.j., nenáhodné veličiny, takže uvedené rovnosti nedávajú zmysel (ani jedná zo štyroch).

  Z matematického hľadiska vôbec nie je jasné, čo predstavujú hodnoty $y_1, \ldots, y_n$, $x_1, \ldots, x_n$, y x, prípadne μ_x a μ_y. Čo je náhodné a čo je nenáhodné? V tejto náväznosti nedávajú zmysel ani rovnice (1.20), (1.21), (1.25) – (1.29) a ani poznámka o *aplikovaná strednej hodnoty* nedáva význam. Navyše rovnice (1.26) a (1.19) nemôžu platiť obe zároveň, ak $p \neq 1$.

- **Neintuitívne/nekonzistentné značenie:** Napr. na str.5 je zavedená matica *daných čísel* $X_{n\times k}$. Jedná sa o náhodné, alebo nenáhodné veličiny? Matica $X_{n\times k}$ sa ale následne nikde neobjavuje. Správne by sa mala objaviť vo výraze (1.14), kde ale autor používa (nedefinovanú) kvantitu $\boldsymbol{X}$. V následujúcom príklade je uvedená matica *daných čísel* $X_{n\times 2}$, ktorá sa ale opäť nikde následne nevyskytuje—hoci by sa formálne mala objaviť v rovnosti (1.16). Navyše v rovnosti (1.17) je zavedené ďalšie značenie pre tú istú maticu $X_{2\times n}$. V Obrázku 1.1 sú zavedené zase hodnoty m_1 a m_2, ktoré sa taktiež nikde mimo obrázok nevyskytujú. V simuláciach sa zase používa niekedy korelačný koeficient ρ, inokedy zase r.