Vysoká škola báňská - Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky

Základy statistiky

Domácí úkol č. 2

Vypracoval: David Ulčák

Osobní číslo: ULC0011

Datum: 24. března 2020

Úvod

Cílem tohoto domácího úkolu je pomocí základních statistických prostředků analyzovat vybranou kvantitativní proměnnou z datového souboru `dotaznik_pp.xlsx` a srovnat ji pro uživatele výhradně legálního software (A) a ostatní (N). V tomto konkrétním provedení jsme si vybrali proměnnou `naklady`, popisující průměrné měsíční náklady respondentů v Kč. Hodnoty této proměnné předpokládáme jako nezáporné.

Uvedená data byla sbírána formou dotazníkového šetření v časovém rozmezí 10. 2. - 23. 2. 2020, respondenty pak byli studenti kurzu Základy statistiky. Pro vlastní analýzu jsou použity pouze tzv. úplné záznamy, tj. záznamy respondentů, kteří odpověděli na všechny otázky (s výjimkou odpovědí na otázku „S jakými technikami sociálního inženýrství jste se setkal(a)?“.

Vypracování

Než se podíváme na základní výběrové charakteristiky a grafické výstupy, učiňme úvahu, jejíž platnost si budeme chtít ověřit: Dá se předpokládat, že uživatelé legálního softwaru budou pravděpodobně za některé služby odvádět určité poplatky, které se mohou promítnout v jejich životních nákladech. Jelikož respondenty dotazníkového šetření byli studenti bakalářského studia, dá se nicméně předpokládat, že tyto poplatky nebudou pro většinu respondentů výše než ve stokorunách. Dá se tedy tušit, že rozdíl mezi průměrnými měsíčními životními náklady studentů ze skupin (A) a (N) nebude nijak zvlášť výrazný.

Tabulka 1: Průměrné měsíční náklady respondentů dle výhradního užívání legílního software (výběrové charakteristiky)

Charakteristika	Průměrné měsíční náklady (Kč)			
	Původní data		Po odstranění odlehlých pozorování	
	A	N	A	N
rozsah	38	55	37	55
minimum	400	1	400	1
dolní kvartil	2 000	2 000	2 000	2 000
medián	3 000	3 500	3 000	3 500
průměr	3 570	4 220	3 390	4 220
horní kvartil	5 000	6 000	5 000	6 000
maximum	10 000	12 000	8 000	12 000
směrodatná odchylka	2 220	2 700	1 970	2 700
variační koeficient (%)	62,2	64,1	58,0	64,1
šikmost	0,7	0,8	0,3	0,8
špičatost	0,3	0,2	-0,7	0,2
Vnitřní hradby pro identifikaci odlehlých pozorování				
dolní mez	-2 500	-4 000		
horní mez	9 500	12 000		

Pojďme si stručně popsat, co jsme se dozvěděli o respondentech, kteří používají výhradně legální software (anebo to alespoň tvrdí):
Za uživatele výhradně legálního software se označilo 38 respondentů. Průměrné měsíční náklady této skupiny studentů se pohybovaly od 400 do 10 000 Kč. Odpovědi, ležící mimo interval $\langle -2\,500; 9\,500 \rangle$ Kč byly vyhodnoceny jakožto odlehlá pozorování (dolní mez vnitřních hradeb se přitom vzhledem k nezápornosti sledované proměnné jeví jako irelevantní). Další komentář se bude týkat pouze záznamů po odstranění odlehlých pozorování. V našem případě byly udané náklady označeny za odlehlé pozorování u jediného respondenta, dále tedy budeme mluvit o průměrných měsíčních nákladech 37 respondentů.

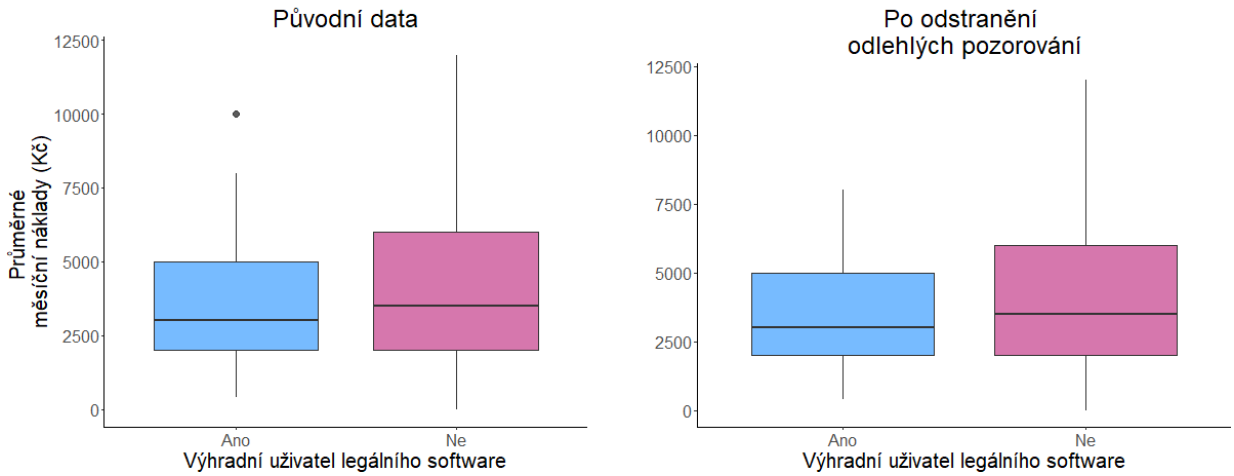
Průměrné měsíční náklady těchto respondentů jsou 3 390 Kč, směrodatná odchylka pak 1 970 Kč. Polovina respondentů uvedla, že jejich průměrné měsíční náklady nepřekračují 3 000 Kč. Čtvrtina respondentů uvedla náklady nejvýše 2 000 Kč, další čtvrtina respondentů zase vyšší než

5 000 Kč.
Grafy na Obrázcích 1 - 3 dávají, na rozdíl od výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží v intervalu $\langle -2; 2 \rangle$), tušit, že průměrné měsíční náklady ani jedné ze skupin nemají normální rozdělení. Můžeme tak usuzovat jak z určité asymetrie krabicového grafu, tak z podoby histogramů a empirických hustot pravděpodobnosti. Pro odhad průměrných měsíčních nákladů studentů druhého ročníku elektrooborů FEI proto musíme použít Čebyševovu nerovnost: Lze očekávat, že alespoň 75 % studentů druhého ročníku elektrooborů FEI má průměrné měsíční náklady v rozmezí $\langle 0; 7\ 330 \rangle$ Kč. (Poznámka: Interval jsme získali kombinaci aplikace Čebyševovy nerovnosti ($\langle -550; 7\ 330 \rangle$ Kč) a znalosti o nezápornosti průměrných nákladů.)

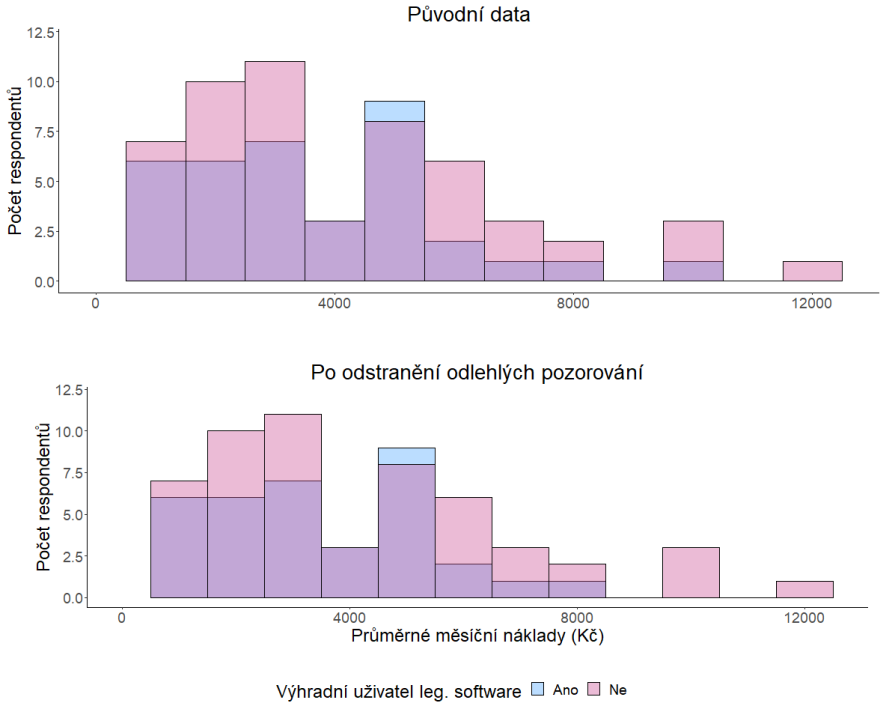
Obdobným způsobem by bylo možno okomentovat také průměrné měsíční náklady studentů, kteří nejsou výhradními uživateli leg. software (viz Tabulka 1).

Na základě výběrových charakteristik uvedených v Tabulce 1 se můžeme pokusit srovnat obě skupiny studentů. Předně si všimněme, že pouze záznamy o nákladech výhradních uživatelů leg. software obsahovaly odlehlá pozorování, konkrétně jedno jediné, které bylo „příliš vysoké“. Dolní meze vnitřních hradeb vyšly pro náklady obou skupin studentů záporné, „nízká“ odlehlá pozorování tedy ani nebylo možno očekávat, s ohledem na předpokládanou nezápornost dat. Vidíme, že i toto jediné odlehlé pozorování celkový obraz značně ovlivnilo, neboť jeho odstraněním u skupiny A poměrně výrazně klesla jak míra variability, tak výběrová špičatost, částečně i šikmost průměrných měsíčních nákladů. Toto odlehlé pozorování nicméně nejspíš bylo způsobeno jednoduše odlišnými poměry daného respondenta, neboť daná hodnota nevypadá nijak „neuvěřitelně“, velmi pravděpodobně tedy nejde o odpověď z recese. Paradoxem je, že ač nízká odlehlá pozorování v datových souborech nemáme, minimální hodnoty 400, resp. 1 Kč naopak nepůsobí příliš věrohodně. Přesto jsme se tato pozorování rozhodli v analýze ponechat.

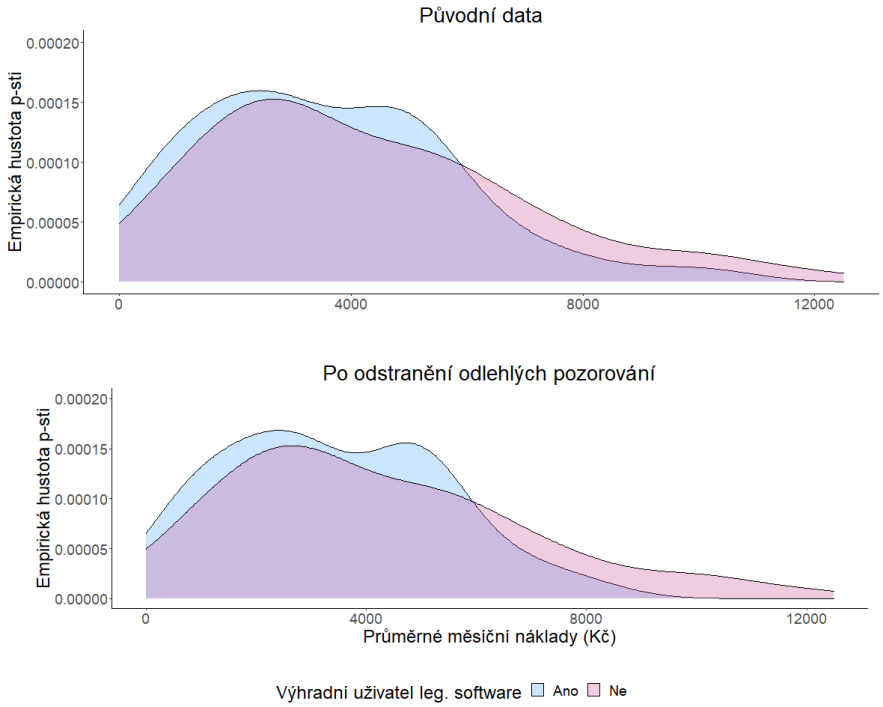
Co nás však může zaujmout především, navzdory naší úvodní domněnce vidíme, že charakteristiky poukazují na o něco vyšší náklady skupiny N (medián, průměr). Tento rozdíl je nicméně poměrně nevýrazný, čímž se nám částečně potvrzuje domněnka, že náklady na software tvoří pouze malou část celkových měsíčních nákladů studentů.



Obrázek 1: Průměrné měsíční náklady respondentů (krabicové grafy)



Obrázek 2: Průměrné měsíční náklady respondentů (histogramy)



Obrázek 3: Průměrné měsíční náklady respondentů (empirické hustoty p-sti)

Grafy na Obrázcích 1 - 3 opět poukazují na o něco vyšší náklady pro studenty ze skupiny N, tento rozdíl se však nadále nejeví jako příliš výrazný. Také rozložením průměrných nákladů jsou si obě skupiny velice podobné. Samozřejmě je otázka, jaké výsledky bychom dostali při větším rozsahu výběrového souboru, ale lze tušit, že fakt, zda člověk užívá výhradně legální software nikterak významným způsobem do celkových nákladů nepromlouvá.