Statistika I – ÚKOL na testy

Preferovaný způsob zpracování: R Markdown. Jakýkoliv jiný formát se správnými odpověďmi a naznačeným postupem je také v pořádku.

Úloha A

1.) Načtěte soubor **vokaly.xls** a funkcí **table()** vytvořte kontingenční tabulku počtu zastoupení vokál vs. pohlaví (proměnné Vowel a Sex). Kolikrát je zastoupen každý vokál u mužů a kolikrát u žen?

2.) Například pomocí funkce **filter()** z balíčku dplyr (představené v předchozím úkolu) vyberte dvě skupiny (každou uložte do samostatné proměnné): Vokál [a] muži, vokál [a] ženy.
Sedí počty řádků těchto tabulek s počty zjištěnými v bodě 1?

3.) Proveďte test, zda hodnoty F2_Hz obou skupin pocházejí z normálního rozdělení.

4.) Pomocí vhodně zvoleného testu rozhodněte mezi hypotézami při alfa = 0.05. Vypočtěte i p-hodnotu.
H0: střední hodnoty F2 žen a mužů u tohoto vokálu jsou stejné.
H1: střední hodnoty jsou různé.

Úloha B

V percepčním testu se ti samí lidé po proškolení ve svých výsledcích v 5 případech zhoršili, ve 35 zlepšili.
Je to statisticky významná změna při alfa = 0.05?
H0: není to změna
H1: lidé se zlepšili
(Nejdříve si pořádně promyslete volbu testu, v případě nejistoty mi nejdříve napište).

Úloha C

V roce 2013 nám na ústav přišlo 5 dotazů týkajících se intonace, 15 zaměřených na hlásky a 16 ohledně IPA transkripce.
V roce 2014 to bylo 7 dotazů na intonaci, 7 na hlásky a 10 ohledně IPA.
Je evidentní, že v roce 2014 přišlo celkově méně dotazů, nás ale spíše zajímá poměr mezi četnostmi v jednotlivých třech kategoriích. Jde o to, zda se poměry v jednotlivých dvou rocích liší, neboli zda je nějaká souvislost mezi rokem a typem dotazu.

H0: poměry mezi kategoriemi jsou v obou rocích stejné – rozdíly jsou jen vlivem náhody, neboli nezávisí na konkrétním roce.
H1: poměry se statisticky významně změnily, neboli hodnoty v jednotlivých kategoriích jsou závislé na tom, zda jsou v jednom či druhém roce.
(Opět si nejdříve promyslete volbu testu a případně mi napište.)

Úloha D: BONUS na zamyšlenou (pro významné plus k zápočtu)
Představte si situaci, kdy je prováděn percepční test (položka zní stejně / zní odlišně), a to pro tři nezávislé skupiny slov (každá skupina obsahuje 240 slov).

Byly zjištěny tyto četnosti úspěchů (počet neúspěchů je v rámci každé skupiny vždy zbytek do 240).
skupina 1: 88
skupina 2: 95
skupina 3: 118

Jak vyhodnotit, zda mají jednotlivé skupiny rozdílné poměry úspěšnosti nebo ne?

Nápad A) Nasadit na to test chí-kvadrát s hypotézou H0: jednotlivé skupiny mají stejnou četnost úspěšných rozpoznání, H1: četnosti jsou různé. Samozřejmě, bavíme se o odhadu vlastností základního souboru, jde nám tedy o to, zda pozorované rozdíly v tomto výběru jsou jen dílem náhody nebo zda jsou statisticky významné.

Nápad B) Pro každou kategorii zvlášť vypočítat interval spolehlivosti pro binomické rozdělení (viz kapitola 12. Intervalové odhady). V každé kategorii samostatně takto počítáme poměr úspěšnosti (např. skupina 1: 88/240 = 36.6 %), ale díky intervalu spolehlivosti k tomuto bodovému odhadu přidáme statistické vyhodnocení nejistoty, tedy binom.test(88, 240)$conf.int => poměr základního souboru se na 95 % nachází mezi 30.5 % a 43.1 %.
Takto vypočteme intervaly pro každou kategorii zvlášť a pokud se některé intervaly nebudou překrývat, prohlásíme dané skupiny za významně rozdílné.

Rovnou prozradím, že B je správně a A není dobře. Pokuste se nad tím zamyslet a pokud přijdete na to, v čem je problém, pokuste se co nejstručněji sepsat zdůvodnění, které vysvětlí, proč není přístup A vhodný. Vzpomeňte si na typický příklad „rovnoměrné kostky“, kterou jsme testovali pomocí testu chí-kvadrát. Je to opravdu analogický případ, nebo čím zásadním se liší?