OPONENTSKÝ POSUDEK NA BAKALÁŘSKOU PRÁCI

Název: Hardyho-Weinbergova rovnováha
Autor: Katarína Vlčková

Shrnutí:

Bakalářská práce Kataríny Vlčkové popisuje testy na Hardyho-Weinbergovu rovnováhu, tj. určitou kombinaci pravděpodobností výskytu genotypů či alel, k níž za jistých předpokladů vede náhodné mísení genetické informace v populaci.

Po krátkém úvodu jsou uvedeny základní vlastnosti multinomického rozdělení, ve čtyřech kapitolách jsou popsány a částečně odvozeny čtyři testy Hardyho-Weinbergovy rovnováhy, na závěr je zařazeno jejich stručné porovnání pomocí simulací.

Práce je po grafické stránce zpracována pečlivě, má hezkou úpravu. Svým obsahem zcela naplňuje zadané téma. Má slabiny v zavádění pojmů a značení, v uvádění dílčích faktů do kontextu a v matematické preciznosti. V práci je relativně dost překlepů, chyb a nepřesností. Naopak závěrečná část s porovnáním testů je na bakalářskou práci udělána celkem pěkně.

Mezi nejdůležitějšími připomínkami bych zmínil:

1. Není vysvětlena ani zdůvodněna Hardyho-Weinbergova rovnováha. Proč náhodné mísení genů v populaci musí vést k určité struktuře v pravděpodobnostech výskytu genotypů a proč má tato struktura zrovna danou podobu? (Popis na straně 5 nic z toho nevysvětluje.)
2. Chybí některé důležité definice a zavedení symbolů. Co to je např. „počet alel“ X_A, X_a (str. 9)? Má to nějaký vztah k počtu jedinců s daným genotypem?
3. U každého testu v kapitolách 2 – 5 mělo být uvedeno: (a) co jsou pozorovaná data a z jakého rozdělení mají pocházet; (b) co je testová statistika; (c) co je kritický obor. Teprve jsou-li tyto tři věci dané, mohou se zkoumat vlastnosti testu, tj. dokazovat, že dodržuje stanovenou hladinu.
4. Nejslabší částí celé práce je kapitola 2 (exaktní test). Exaktní test není zaveden (chybí testová statistika i kritický obor), přesto se poněkud zmateným postupem zkoumá výpočet jeho p-hodnoty.
5. Druhá problematická část je kapitola 4 (modifikovaný χ^2 test). Testová statistika je zavedena až na konci (s opravou na spojitost) a její vlastnosti jsou zdůvodněny poznámkou nejasného původu s neověřenými předpoklady. Odhad $\tilde{\theta}$ nejspíš není asymptoticky ekvivalentní maximálně věrohodnému odhadu ani není „asymptoticky efektívny“.

Přestože práce Kataríny Vlčkové má jisté nedostatky, zadání podle mého splňuje. Považuji za přijatelnou a doporučuji ji uznat jako práci bakalářskou.

Konkrétní připomínky:

- 8_3 Z čeho plyne, že var $(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j)$?
- 9_8 $p_{aa} = p_a^2(1-\theta)^2$ Tato rovnost neplatí.
- 9_3 $H_0 : p_{Aa}^2 = 4p_{AA}p_{aa}$. To je stejná hypotéza jako předchozí H_0 nebo je to jiná hypotéza? Odkud pochází?
- 10^3 *„alela A sa za n pokusov vyskytne vo výbere x_A-krát“*. O jaké pokusy a o jaký výběr se jedná?
- 10^{10} *„zo vzťahu $x_A = 2x_{AA} + x_{Aa}$.“* Odkud pochází tento vztah? V práci je na tomto místě uveden poprvé.
- 12_4 θ_m místo θ_n.
- 13^9 Ztratily se kvadráty v čitatelích.
- 13_5 *„multinomické . . . rozdelenie aproximujeme rozdelením χ_1^2. . . “* Nejedná se o multinomické rozdělení.
- 14_{11} Co znamená *„sčítavame cez $x_{Aa} - 1$“*? Sčítat se musí přes index s jasně danými mezemi, toto vyjádření není matematicky korektní. Ve výrazech $*$ nevím, co se míní *„sčítaniem cez všetky možné pravdepodobnosti trinomického rozdelenia“* a netuším, co vede k tomu, že se tyto součty vysčítají do 1.
- 14_5 *„vyžadujeme aby $(\theta^*)^2 = nE[X_{AA} \mid X_A = x_A, X_a = x_a]$.“* Proč se vyžaduje zrovna toto a ne něco jiného? Na pravé straně je chyba, takhle by ten odhad nefungoval.

- 15^3 Jak by se zdůvodnilo, že $E(\theta^2 - \frac{\theta X_A}{n} + \frac{X_A^2}{(2n)^2}) \longrightarrow 0$?
- 15^5 To, že dva odhady konvergují k témuž parametru, neznamená, že jsou asymptoticky ekvivalentní.
- 17, věta 5.2. $L(\alpha)$ a $J(\alpha)$ nebyly definovány.

doc. Mgr. Michal Kulich, PhD.
KPMS MFF UK
12. června 2012