*Posudek oponenta na disertační práci*

# *Michaela Šedová*: Odhad parametru při dvoufázovém stratifikovaném a skupinovém výběru

Předložená práce je věnována problematice odhadování parametru při stratifikovaném a skupinovém výběru. Tato tzv. složitější uspořádání výběru jsou předmětem zájmu teorie výběrových šetření již několik desetiletí. Výsledky výzkumu v této oblasti však byly převážně motivována cílem popsat danou konečnou populaci. Tento přístup vyhovuje především tzv. oficiální statistice, která si právě klade za cíl popis populace.

Mnohé vědní obory však mají cíle analytické, tj. pomocí statistického modelu zkoumat vztahy mezi veličinami, odhadovat parametry, dělat testy. K tomuto účelu se hodí představa, že daná konečná populace byla vygenerována z modelu nazývaného „superpopulace". O tomto modelu pak chceme provádět statistickou inferenci. Je však nutné se vypořádat s tím, že kvůli uspořádání výběru nelze považovat data za kolekci nezávislých stejně rozdělených náhodných veličin.

Je záhodno říci, že přestože existuje rozsáhlá literatura o teorii výběrových šetření, je relativně málo prací, které se zabývají využitím dat z výběrových šetření pro analytické účely. Předložené práce je cenným příspěvkem k této oblasti. Oceňuji zejména, že práce nevznikla naprosto disjunktně ke klasické teorii výběrů z konečných souborů. Práce si všímá souvislosti postupů využívaných např. při studiích v biostatistice s postupy při výběrech z konečných populací. Je tak spojovníkem mezi „klasickou" teorií výběrových šetření a „klasickou" matematickou statistikou založenou na modelech.

Práce se soustřeďuje na stratifikovaný a skupinový výběr, přičemž uvažuje i skupinový výběr uvnitř strat. Pro každý z těchto výběrů se zabývá odhadem střední hodnoty a také využitím pomocných veličin k vylepšení statistických vlastností tohoto odhadu. V poslední kapitole pak uvádí, jak tyto výsledky aplikovat na odhad parametrů v zobecněném lineárním modelu.

Následující připomínky míním spíše jako upozornění pro případné publikování částí této práce v odborné literatuře, než jako závažné nedostatky. Obecně lze říct, že na některých místech by čitelnosti práce pomohla trocha více názornosti při odvozování. Některé důkazy by také dle mého názoru měly být provedeny precizněji.

Chapter 2 „Two-phase Stratified Sampling".

(1) Strana 20: Čtenáři, který se v dané oblasti nepohybuje nemusí být ihned jasné, že odezvou v popisovaném logistickém modelu jsou indikátory zahrnutí $\{\xi_i,\, i = 1, \ldots, N\}$.
(2) Strana 21: Asymptotická reprezentace (2.13) je zřejmě standardní výsledek asymptotické linearity parametrů pro logistickou regresi. Přesto by čtenáři pomohl nějaký vhodný odkaz na literaturu.
(3) Strana 21: Rovnice na řádcích 10 a 11. Obávám se, že tento rozvoj by měl být podrobněji zdůvodněn. Zbytkový člen při Taylorově rozvoji $\frac{1}{\pi_i(\tilde{\gamma}^\top \mathbf{Z}_i)}$ obecně závisí na indexu $i$ a to že, zbytkový člen je $o_P(\frac{1}{\sqrt{N}})$ stejnoměrně v $i$ není zřejmé. Podobně dále v důkazech Věty 6 a Věty 7.
(4) Strana 23 dole: Dá se nějak intuitivně zdůvodnit, proč je lépe použít místo $\mathbf{X}_i$ raději $\frac{\mathbf{X}_i}{\pi_i}$?

(5) Strana $23_3$: Nevidím, proč $\mathrm{cor}_k(X, Y) = 1$ implikuje, že $c^\mathsf{T} V^{-1} c = \sum_k p_k \frac{1-\pi_k}{\pi_k} \mathrm{var}_k Y_i$. Není k tomu zapotřebí, že $\mathrm{var}_k X_i$ nezávisí na $k$?

Chapter 3 „Cluster sampling“.

(1) Strana 29: Nevidím, jak (3.6) plyne z (3.2).
(2) Strana 31: Ve Větě 4 nerozumím předpokladu nezávislosti náhodných vektoru $(Y_j, \xi_j, M_j)$ pro $j = 1, \ldots, N$. Jestli tomu správně rozumím, pak všechny jednotky, které patří do dané domácnosti mají společné $M_j$.
(3) Strana $32^{15}$: Není asi příliš šikovné, že $n$ zůstalo na pravé straně limity. Pokud $N \to \infty$ tak také $n \to \infty$.
(4) Strana $32_3$: Zajímalo by mě, které statistické metody vyvinuté pro Bernoulliho výběr jsou v pořádku pro skupinový výběr o rozsahu jedna uvnitř skupinky.
(5) Strana 33: Při vysčítávání jednotlivých skupinek je na několika místech uvedena $\sum_{i=1}^{N} M_i$, ale správně je asi $\sum_{i=1}^{n} M_i$.
(6) Strana $33_6$: V rovnici pro $E_I(\mathrm{var}_{II}(\hat{\theta}))$ je asi pouze aproximativní rovnost.

Chapter 4 „Stratified Cluster Sampling“.

(1) Strana 38: V rovnicích (4.7) a (4.8) by asi mělo být $\tilde{\pi}_i$ místo $\tilde{\pi}_k$.

Chapter 5 „Use of Auxiliary Variables in Cluster Sampling“.

(1) Strana $64_{10}$: Dle Tabulky 5.9 je odhad sexuálních partnerů při zohlednění velikosti domácnosti a pohlaví 4.7655 (a ne 4.7782).

Chapter 6 „Extension to the Regression Problem“.

(1) Strana 68: Nejdříve se v (6.2) definuje $f(y|\mathbf{x})$, potom se v (6.3) používá $f_i(\boldsymbol{\theta}|y_i)$, což bude asi $\log f(y_i|\mathbf{x}_i)$.
(2) Strana $71_6$: Není zřejmé, proč by měl být zbytkový člen v rozvoji $\mathbf{V}(\hat{\boldsymbol{\theta}}) - \mathbf{V}(\boldsymbol{\theta})$ řádu $o_p(\frac{1}{\sqrt{N}})$. Zdá se, že autorka předpokládá, že $\boldsymbol{\theta}$ je $\sqrt{N}$-konzistentní odhad $\hat{\boldsymbol{\theta}}$. Ale i tak pokud danou rovnici nevydělíme $N$, pak zbytkový člen bude řádu $o_p(\sqrt{N})$.
(3) Strana 81 Tabulka 6.2.: Ve sloupečku „Average estimate“ je hodnota „$< 0.001$“. Má tam být „$-0.001$“?

Chapter 7 „Summary and Conclusion“.

(1) Strana $84_1$: ... Domnívám se, že před slovem „rozptyl“ by měl být přívlastek „asymptotický“.

Práce má jednotnou a přehlednou koncepci a velmi zajímavě přispívá k dané problematice. Navíc je napsána s velkou pečlivostí (s minimem překlepů) a také (nakolik dokáži posoudit) velmi pěknou angličtinou.

Dle mého názoru práce splňuje předpoklady kladené na doktorskou disertační práci. Doporučují proto, aby na jejím základě byl autorce přiznán titul Ph.D. v oboru matematika

30. března 2011

Ing. Marek Omelka, Ph.D.
KPMS MFF UK