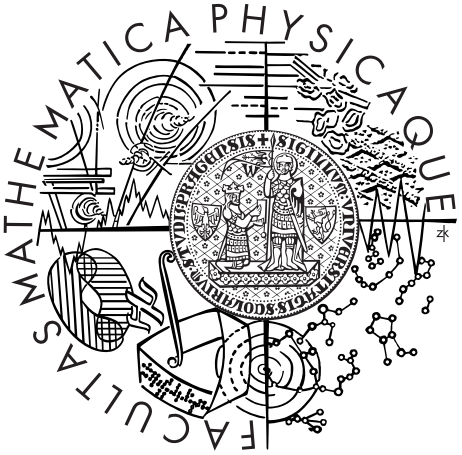
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Petr Hanek

Výběrové metody v lesnictví

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2013

Je mou milou povinností poděkovat RNDr. Zbyňku Pawlasovi, Ph.D., za odborné a svědomité vedení při psaní práce, za podnětné připomínky, za cenné rady, trpělivost, ochotu a čas, který mi věnoval při konzultacích. Dále děkuji svým kolegům, ať už z pracovního kolektivu nebo z akademické obce za jejich podporu v průběhu psaní diplomové práce.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Výběrové metody v lesnictví
Autor: Petr Hanek
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.

Abstrakt:
Tato diplomová práce se zabývá výběrovými strategiemi v lesnictví. Popisuje jejich teoretické aspekty i aplikace na reálné krajině. Výběrové metody v lesnictví mají velký význam především při inventarizaci lesů. Cílem výběrových metod je odhadnout charakteristiky populace na základě znalosti výběrového souboru. Rozlišují se dva základní přístupy k populaci podle její velikosti, mluvíme o diskrétní nebo spojité populaci. V textu jsou popsány různé způsoby volby výběrového souboru a konstrukce příslušných odhadů sledovaných charakteristik pro oba přístupy. Kromě odhadů populačního úhrnu nebo průměru jsou zmíněny vzorce pro výpočet rozptylu těchto odhadů a metody jejich odhadu při různých výběrových plánech. Součástí práce je porovnání studovaných metod na základě počítačových simulací.

Klíčová slova:
Horvitzův-Thompsonův odhad, inventarizace lesů, lineární odhad, výběrový plán

Title: Sampling methods in forestry
Author: Petr Hanek
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Zbyněk Pawlas, Ph.D.

Abstract:
This diploma thesis is devoted to the sampling strategies in forestry. It describes their theoretical aspects and their applications on a real landscape. The sampling methods in forestry are of particular importance in forest inventory. The aim of sampling methods is to estimate population characteristics based on the knowledge of sample. Two basic approaches can be distinguished according to the size of population, we speak about discrete or continuous population. Several types of sampling designs and corresponding estimators of target values are described for both approaches. Besides estimates of population total or average, we mention the formulas for computing variance of these estimates and the methods for their estimation for different sampling designs. The thesis also contains the comparison of studied methods based on computer simulations.

Keywords:
Horvitz-Thompson estimator, forest inventory, linear estimator, sampling design

Obsah

Kapitola 1

Úvod

Cílem této práce je přehledně shrnout výběrové metody uplatnitelné v lesnictví. Jednotlivé metody budou popsány a zároveň si u nich uvedeme základní charakteristiky včetně jejich odhadů. Přestože je práce primárně určena pro čtenáře s pokročilejší znalostí statistiky, pozorný a vnímavý student se základním kurzem statistiky by neměl mít problém s pochopením základní problematiky. Nejprve řekněme něco o potřebě, dokonce nutnosti, vzniku výběrových strategií. V literatuře se můžeme setkat také s pojmem výběrové metody nebo metody náhodného výběru. Všechna tato slovní spojení znamenají to samé. Motivace pro vznik a použití je velmi intuitivní. Přináší nám odpověď na jednu ze základních otázek kladených člověkem: „Kolik?“ Budeme-li uvažovat přírodní zdroje, pak v historii tuto otázku kladli nejčastěji panovníci, sedláci či rolníci. Tato otázka neztrácí na významu ani v současné době moderních technologií. Jediné, co se mění, jsou metody zjišťování informací a technika zpracovávající tyto informace. Pro získání námi požadované informace je třeba mít vhodnou výběrovou strategii. Podstata výběrových metod tkví v usuzování z části na celek. Vybereme tedy část prvků rozsáhlého souboru a provádíme odhady charakteristik celého souboru. Výběrová strategie je jednoduše řečeno kombinace výběru menšího vzorku s odhadem jedné či více charakteristik celé populace.

Zaměřme se na výběrové strategie uplatnitelné v přírodních či biologických směrech. Téměř ve všech případech je výběr do jisté míry ovlivněný navrhovatelem nebo tím, kdo plán realizuje. Můžeme tedy výběr považovat za řízený. Jedním z důvodů vzniku výběrových strategií pro odhad vlastnosti celé množiny je to, že ve většině situací je totiž z důvodů nedostatku času nebo prostředků nemožné uskutečnit kompletní šetření. Výjimku zde tvoří snad jen sčítání lidu. Dalším důvodem je to, že při některých biologických šetřeních může docházet při zkoumání ke zničení vzorku. A je zřejmé, že vykácet všechny stromy v lese, abychom zjistili průměrné stáří stromů, není úplně ta správná cesta. Takové metody jsou nazývány obecným názvem destruktivní. V neposlední řadě je důležité zmínit, že plné vyšetření populace by bylo nákladné nejen časově, ale i finančně. Co si představit pod dobrým výběrovým

plánem? Obecně dobrý výběrový plán je takový plán, který nám dává dostatečně kvalitní alternativu ke kompletnímu šetření populace v tom smyslu, že získaná informace je dostatečně hodnotná s ohledem na zdroje informace. V jednoduchosti řečeno výběrový plán nám řekne, jak vybrat nějaký vzorek a jak z tohoto vzorku odhadnout vlastnost celé populace. Existuje mnoho způsobů, jak vzorek můžeme vybrat, ale bohužel neexistuje žádná univerzální optimální cesta k výběru toho nejlepšího vzorku. Tedy důležitá otázka je způsob, jakým vzorek vybíráme, nebo otázka velikosti vzorku.

Práce je členěna do osmi kapitol. První kapitola je úvodem do dané problematiky. Ve druhé kapitole sjednotíme základní terminologii a značení. Ve třetí kapitole se seznámíme s metodami odhadu vlastností celého souboru. Základním výběrovým plánům pro konečné populace se věnuje čtvrtá kapitola. O spojitých populacích se čtenář dočte v kapitole páté. Lesnické metody používané v praxi jsou představeny v šesté kapitole. Porovnání teoretické části z vybraných výběrů ze čtvrté, páté a šesté kapitoly s výsledky získanými pomocí simulací v softwaru R je publikováno v sedmé kapitole. Návody a postupy k softwaru byly získány z [8]. Kapitola poslední, osmá, je celkovým shrnutím vybraných metod včetně jejich hlavních výhod a nevýhod.

Kapitola 2

Základní terminologie, značení a charakteristiky

V literatuře se často můžeme setkávat s různými názvy a značením jedné a té samé věci. V této kapitole připomeneme a sjednotíme základní značení a terminologii. Značení, terminologie a myšlenky této kapitoly můžeme naleznout v publikacích [2], [7], [11] a [12].

Při šetření, zkoumání či výběru je důležité si uvědomit, co je předmětem a cílem našeho snažení. Předmětem zájmu je statistický soubor zvaný populace, někdy též základní soubor. Populace je množina prvků, jedinců nebo objektů, které tvoří základní soubor pro výběr, a značíme jí $\mathbb{U}$. Při výběrech rozlišujeme dva základní typy populací. První z nich považuje prvky základního souboru za konečnou množinu prvků, druhý považuje populaci za nekonečnou množinu bodů. Platí $\mathbb{U} = \{U_x : x \in \mathcal{D}\}$, kde $\mathcal{D}$ je buď množina daná výčtem, $\mathcal{D} = \{1, 2, \ldots, N\}$, nebo interval, oblast, obecně nějaká podmnožina prostoru $\mathbb{R}^d$. U takového souboru nás zajímá nějaká jeho vlastnost. Tuto vlastnost nazýváme populační charakteristikou. Zkoumaná hodnota vlastnosti spojená s prvkem U_x se obvykle značí y_x. V rámci výběru a zkoumání nás na jednom prvku populace nemusí zajímat nutně jen jedna vlastnost, ale můžeme se zajímat o více atributů najednou, např. výšku, stáří, množství listí na stromě apod. Při výběru vzorku vybíráme nějakou podmnožinu s ze základního souboru $\mathbb{U}$.

Pro diskrétní populace bychom mohli říct, že vybíráme nějakou podmnožinu o velikosti n z celé populace o velikosti N. Rozlišujeme dva základní případy. První je ten, kdy počet n prvků ve vzorku je nenáhodné, před výběrem pevně dané číslo, říkáme též, že výběr má pevný rozsah. Druhý, jak bychom intuitivně očekávali, je takový, kde počet prvků ve vzorku je náhodné číslo n. Každý výběr s má určitou šanci, že bude zvolen. Tato pravděpodobnost se nazývá výběrová pravděpodobnost a značíme ji $p(s)$. Obecně různé výběrové plány můžou mít různé výběrové pravděpodobnosti.

Důležitou informací je počet možných vzorků. Množinu všech možných vzorků značíme Ω a její velikost, $|\Omega|$, je rovna počtu možných vzorků. U většiny výběrových plánů je počet možných vzorků mnohem vyšší, než je počet prvků v populaci. Je to způsobeno tím, že většina plánů připouští, že dva různé vzorky mohou obsahovat společné prvky. Pro pravděpodobnosti výběru a možné vzorky s_i platí základní vztah $\sum_{i=1}^{|\Omega|} p(s_i) = 1$. Pro nekonečně velké populace a pro spojité populace je Ω nekonečná.

Dalším termínem, se kterým se budeme velmi často setkávat a zároveň pro nás bude velmi důležitou informací, je pravděpodobnost zahrnutí. Pro pravděpodobnostní výběr platí, že každý prvek má nenulovou pravděpodobnost zahrnutí ve vzorku.

Definice 1. *Symbolem π_i značíme pravděpodobnost zahrnutí i-tého prvku konečné populace do výběru, tedy*

$$\pi_i = \sum_{s \ni U_i} p(s),$$

kde $s \ni U_i$ značí ty vzorky s, které obsahují i-tý prvek. Tato pravděpodobnost se nazývá pravděpodobnost zahrnutí prvního řádu.

Zavedení tohoto pojmu pro spojitou populaci je uvedeno v podkapitole 3.2.

Rozlišujeme dva základní přístupy k výběru vzorku. První je designově založený a druhý je modelově založený přístup. U designově založených výběrů je jediným náhodným faktorem samotný výběr prvků populace, tedy vzorek je chápán jako realizace náhodného procesu. Máme-li prvky populace, tak množství vlastnosti každého z nich je neznámé, ale fixní. Po výběru je již množství známé. Na druhé straně u modelově založených přístupů je populace považována za realizaci náhodného procesu. Výhoda designového modelu je ta, že nezávisí na předpokladech. Modelový přístup má tu výhodu, že je ve shodě s mnohými výsledky ze statistické analýzy. V praxi je důležité, s jakým přístupem pracujeme, protože nestranný odhad pod jedním přístupem nemusí být nestranný pod druhým.

Uvažujme nyní diskrétní populaci. Diskrétní populací rozumíme takový soubor, který se skládá z konečného počtu rozlišitelných prvků. Prvky označujeme U_1, U_2, U_3, …, U_N. První z populačních charakteristik, která nás bude zajímat, je populační úhrn. Nechť y_i značí hodnotu charakteristiky našeho zájmu spojenou s i-tým prvkem populace. Populační úhrn Y získáme sumací charakteristiky přes všech N prvků populace. Další takovou charakteristikou, ke které se ubírají naše zájmy, je populační průměr $\bar{Y}$, pro který platí $\bar{Y} = Y/N$.

Definice 2. *Nechť je dána konečná populace s prvky $U_1, U_2, \ldots, U_N$. Dále nechť y_i značí hodnotu vlastnosti prvku U_i, $i = 1, \ldots, N$. Potom hodnotu*

$$Y = \sum_{i=1}^{N} y_i$$

nazýváme populační úhrn.

Definice 3. *Populačním průměrem rozumíme hodnotu*

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Obecně můžeme říct, že populační průměr je průměrná hodnota charakteristiky vztažená k jedné jednotce populace. Pro úplnost si uveďme některé další zajímavé populační charakteristiky, se kterými se můžeme setkat. Jedná se zejména o poměr a rozsah. Poměr je definován jako poměr dvou charakteristik, např.

$$R_{y|x} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}.$$

Rozsah je definován jako rozdíl mezi maximální a minimální hodnotou v populaci a můžeme ho považovat za míru rozpětí hodnot v populaci.

Druhým možným případem je spojitá populace. Uvažujme tedy spojitou populaci, ta se přirozeně nedělí do diskrétních subjektů, ani nedovoluje jednodušší popis. Celkový soubor $\mathbb{U}$ se skládá z prvků U_x, kde $x \in \mathcal{D}$, což je nekonečná podmnožina prostoru $\mathbb{R}^d$. V analogii s hodnotou vlastnosti y_i spojenou s i-tým prvkem v diskrétním případě budeme intuitivně uvažovat hodnotu vlastnosti spojenou s bodem x, tu nazýváme hustotou vlastnosti a značíme $y(x)$. V jednodimenzionálním případě nám tato hustota dává množství vlastnosti v délce jednotky nebo na jednotku času. Ve dvoudimenzionálním nebo třídimenzionálním případě je hustota typicky množství vlastnosti na jednotku plochy nebo objemu. Jednou z populačních charakteristik je celkové množství, neboli úhrn, vlastnosti existující přes celý soubor. Populační úhrn definujeme jako integrál hustoty vlastnosti na množině $\mathcal{D}$ a značíme jej opět Y.

Definice 4. *Nechť je dána spojitá populace $\mathbb{U} = \{U_x : x \in \mathcal{D}\}$. Nechť $y(x)$ značí hustotu vlastnosti pro $x \in \mathcal{D}$, potom*

$$Y = \int_{\mathcal{D}} y(x)\,\mathrm{d}x$$

nazveme populačním úhrnem pro spojitou populaci.

Definice 5. *Populační průměr hustoty vlastnosti přes oblast $\mathcal{D}$ je*

$$\bar{Y} = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} y(x)\,\mathrm{d}x = \frac{Y}{|\mathcal{D}|},$$

kde $|\mathcal{D}|$ je d-rozměrná Lebesgueova míra množiny $\mathcal{D}$.

Nyní již víme, co budeme zkoumat. To, co zatím jen tušíme, je, jak budeme naše zkoumání hodnotit. Během pozorování získáme vzorek $s \subseteq \mathbb{U}$. Naším cílem je na základě pozorování, které zahrnuje vzorek s, stanovit odhad populační charakteristiky, jako je úhrn, průměr apod. Odhadem rozumíme algebraické vyjádření z dat vzorku dávající kvantitativní odhad cíleného parametru. Nechť Y značí charakteristiku našeho zájmu a $\hat{Y}$ značí její odhad. V některých situacích budeme zapisovat $\hat{Y}(s)$ namísto $\hat{Y}$, a to pro zdůraznění závislosti na výběru s. Předpokládáme, že hodnoty y_x, $x \in \mathcal{D}$ jsou neznámé a je nemožné vyhodnotit všechny prvky v populaci. To je jedním z důvodů, proč vybíráme vzorek s za účelem odhadu skutečné hodnoty Y. Naším cílem je stanovit odhad co nejpřesnější. Je tedy intuitivní, že čím větší náš vzorek bude, tím více se hodnoty odhadu $\hat{Y}$ budou blížit skutečné hodnotě Y.

Kvalita a přesnost odhadu závisí však i na výběrovém plánu. Obecně platí, že odhad získaný z jedněch napozorovaných dat se bude lišit od odhadu získaného z jiných napozorovaných dat. Tato rozmanitost způsobená rozdílnými vzorky je nazývána výběrová variace. Tyto rozdíly zapříčiňují vznik rozdělení výběru. To nás zajímá, pokud zkoumáme vlastnost odhadu, a ne rozdělení hodnot ve vybrané populaci. Souvislost s velikostí vybraného vzorku se nazývá konzistence. V literatuře se setkáváme s různými přístupy a vysvětleními. My budeme uvažovat přístup, kdy je odhad nazýván konzistentním, pakliže je identicky roven cílenému parametru, kdykoliv vzorek zahrnuje celou populaci. Když v těchto situacích platí, že $\hat{Y} \neq Y$, pak odhad nazýváme nekonzistentním.

Další vlastností odhadu je střední hodnota. Jedná se o vážený průměr všech možných odhadů, kde váha násobící odhad $\hat{Y}(s)$ z jednotlivého vzorku s je pravděpodobnost, že bude daný vzorek vybrán, tedy $p(s)$. Střední hodnota pak může být vyjádřena jako

$$\mathbb{E}\left[\hat{Y}\right] = \int_{\Omega} p(s)\hat{Y}(s)\,\mathrm{d}\mu(s),$$

což pro diskrétní populaci s čítací mírou μ znamená

$$\mathbb{E}\left[\hat{Y}\right] = \sum_{s\in\Omega} p(s)\hat{Y}(s),$$

kde $s \in \Omega$ chápeme jako sumaci přes všechny možné vzorky v daném výběrovém plánu. Ve speciálním případě výběrového plánu, kdy jsou všechny vzorky stejně pravděpodobné, platí, že $p(s) = \frac{1}{|\Omega|}$ je konstantní, což zjednodušuje výraz $\mathbb{E}\left[\hat{Y}\right]$ na prostý aritmetický průměr

$$\mathbb{E}\left[\hat{Y}\right] = \frac{1}{|\Omega|}\sum_{s\in\Omega} \hat{Y}(s).$$

Velmi často budeme užívat vlastnost vychýlení. Vychýlení je definováno jako rozdíl mezi střední hodnotou a populační charakteristikou, pro kterou je parametr počítán. Značíme ho $B\left[\hat{Y}:Y\right]$ a počítáme jako

$$B\left[\hat{Y}:Y\right]=\mathbb{E}\left[\hat{Y}\right]-Y.$$

Pokud je odhad $\hat{Y}$ nestranný, tedy $\mathbb{E}\left[\hat{Y}\right]=Y$, pak je vychýlení rovno nule, někdy také říkáme, že $\hat{Y}$ je nevychýlený odhad. Na rozdíl od střední hodnoty je vychýlení odhadu funkcí jednotlivých populačních parametrů. Tedy nemůžeme mluvit o vychýlení nebo nevychýlení bez určení nejen výběrového plánu, ale také populačního parametru, který je odhadován. Velikost rozdílu $\hat{Y}(s)-Y$ je známa jako výběrová chyba.

Jednou z dalších vlastností odhadu je rozptyl. Rozptyl můžeme definovat jako průměrnou čtvercovou vzdálenost mezi jednotlivými vybranými hodnotami a jejich střední hodnotou. Značí se $V\left[\hat{Y}\right]$ a počítáme ho jako

$$V\left[\hat{Y}\right]=\int_{\Omega}p(s)\left(\hat{Y}(s)-\mathbb{E}\left[\hat{Y}\right]\right)^{2}\mathrm{d}\mu(s).$$

Pro diskrétní populaci s čítací mírou μ přechází předchozí vztah na

$$V\left[\hat{Y}\right]=\sum_{s\in\Omega}p(s)\left(\hat{Y}(s)-\mathbb{E}\left[\hat{Y}\right]\right)^{2}.$$

Porovnáváme-li dva nevychýlené odhady, pak ten s menším rozptylem je lepší. Některé situace a odhady vyžadují znalost standardní chyby. Standardní chyba je definována jako odmocnina z rozptylu odhadu, tedy $\sqrt{V\left[\hat{Y}\right]}$.

Dalším pojmem a informací o kvalitě odhadu je střední čtvercová odchylka. Střední čtvercová odchylka je podobně jako rozptyl pravděpodobnostně vážený průměr čtvercových vzdáleností mezi $\hat{Y}(s)$ a Y. Pomocí symbolů zapisujeme jako $MSE\left[\hat{Y}:Y\right]$ a počítáme jako

$$MSE\left[\hat{Y}:Y\right]=\int_{\Omega}p(s)\left(\hat{Y}(s)-Y\right)^{2}\mathrm{d}\mu(s).$$

Tento vzorec je v případě konečné populace možné psát jako

$$MSE\left[\hat{Y}:Y\right]=\sum_{s\in\Omega}p(s)\left(\hat{Y}(s)-Y\right)^{2}.$$

Použijeme-li definice pro $B\left[\hat{Y}:Y\right]$ a $V\left[\hat{Y}\right]$, pak po menších úpravách dostaneme, že

$$MSE\left[\hat{Y}:Y\right]=V\left[\hat{Y}\right]+\left(B\left[\hat{Y}:Y\right]\right)^2.$$

Je zřejmé, že když $\mathbb{E}\left[\hat{Y}\right]=Y$, pak $MSE\left[\hat{Y}:Y\right]$ a $V\left[\hat{Y}\right]$ jsou totožné. Užitečnost střední čtvercové odchylky je ta, že umožňuje lépe posoudit interpretaci odhadu. Při porovnávání dvou odhadů je ten s menší střední čtvercovou odchylkou více přesný. Kromě čtvercové odchylky budeme uvažovat i absolutní odchylku

$$MAE\left[\hat{Y}:Y\right]=\int_\Omega p(s)\left|\hat{Y}(s)-Y\right|\,\mathrm{d}\mu(s).$$

Je důležité si uvědomit základní rozdíl mezi $\hat{Y}$ jako odhadem nějaké populační charakteristiky Y a kvalitou jako $V\left[\hat{Y}\right]$. Jmenovitě $\hat{Y}$ je náhodná veličina, zatímco $V\left[\hat{Y}\right]$ nikoliv.

Kapitola 3

Metody odhadu

V předchozí kapitole jsme uvedli charakteristiky, které budeme odhadovat, a vlastnosti odhadu. V této kapitole uvedeme, jak budeme tyto charakteristiky odhadovat. Použité vzorce včetně jejich podrobnějšího odvození jsou dostupné ve článku [3] a ve skriptech [12].

3.1 Lineární odhad úhrnu, Horvitzův-Thompsonův odhad

Uvažujme diskrétní populaci. Základním ukazatelem je populační úhrn. Lineární odhad úhrnu bude metoda spojená s výběry, které si představíme v pozdějších kapitolách. Pro hrubou představu to jsou zejména prostý náhodný výběr, Poissonův výběr a systematický výběr.

Hodnoty $y_1, y_2, \ldots, y_N$ považujeme za konstanty. Nepovažujeme je za náhodné veličiny nebo jejich realizace. Jedná se o designově založený plán. Jediný náhodný vliv zde bude spočívat v tom, že s je náhodně zvolený výběr. Rozdělení odhadů bude dáno použitým výběrovým plánem.

Definice 6. *Nechť $s \subseteq \mathbb{U}$ je náhodně zvolený výběr. Lineárním odhadem úhrnu Y nazveme náhodnou veličinu $\hat{Y} = \hat{Y}(s)$, $s \subseteq \mathbb{U}$, definovanou vztahem*

$$\hat{Y} = \sum_{i:U_i \in s} y_i w_i, \tag{3.1}$$

kde w_i jsou libovolné váhy.

Váhy jsou obecně náhodné veličiny z důvodu závislosti na výběru s. Můžeme proto psát $w_i(s)$ namísto w_i. Pro váhy musí být splněno $w_i(s) = 0$ pro každé i takové, že $U_i \notin s$. V některých případech je výhodné vyjádřit tyto váhy jako násobek nějaké konstanty a indikátoru zahrnutí.

Definice 7. *Indikátor zahrnutí, někdy jen indikátor, je náhodná veličina definovaná jako*

$$I_i = \begin{cases} 1, & \text{když } U_i \in s, \\ 0, & \text{když } U_i \notin s. \end{cases}$$

Platí tedy $w_i(s) = w_i I_i$ a po dosazení do (3.1) dostáváme pro odhad úhrnu vztah

$$\hat{Y} = \sum_{i=1}^{N} y_i w_i I_i,$$

kde $I_1, I_2, \ldots, I_N$ jsou indikátory zahrnutí.

Výběrovou odhadovou strategií rozumíme dvojici (p, w), kde $\{p(s), s \subseteq \mathbb{U}\}$ je výběrový plán a $\{w_i(s), s \subseteq \mathbb{U},\ i = 1, \ldots, N\}$ jsou váhy.

Definice 8. *Nechť pro váhy w_i platí $w_i(s) = \frac{1}{\pi_i}$ pro všechna i taková, že $U_i \in s$ a $w_i(s) = 0$ pro všechna i taková, že $U_i \notin s$. Potom lineární odhad definovaný vztahem (3.1) nazýváme prostý lineární odhad. V literatuře se častěji setkáme s názvem Horvitzův-Thompsonův odhad.*

Tento odhad je tedy definovaný pro ty výběrové plány, jejichž pravděpodobnosti zahrnutí splňují $\pi_i > 0$, $1 \leq i \leq N$, jako

$$\hat{Y} = \sum_{i:U_i \in s} \frac{y_i}{\pi_i}. \tag{3.2}$$

Tvrzení 1. *Nutnou a postačující podmínkou pro to, aby odhad úhrnu Y definovaný vztahem (3.1) byl nevychýlený, je $\mathbb{E}w_i = 1$, $1 \leq i \leq N$.*

Důsledkem předchozího tvrzení je, že prostý lineární odhad (3.2) je nevychýlený. Pro výběrové plány, pro které známe pravděpodobnosti zahrnutí π_i, můžeme využít prostý lineární odhad.

V předchozí kapitole jsme zmínili pojem střední čtvercové odchylky, kterou označujeme jako $MSE\left[\hat{Y} : Y\right]$ a počítáme ji $\mathbb{E}\left(\hat{Y} - Y\right)^2$.

Tvrzení 2. *Pro obecný lineární odhad definovaný vztahem (3.1) má vzorec pro výpočet střední čtvercové odchylky tvar*

$$\mathbb{E}\left(\hat{Y} - Y\right)^2 = \mathbb{E}\left[\sum_{i=1}^{N} y_i(w_i - 1)\right]^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \mathbb{E}\left[(w_i - 1)(w_j - 1)\right]. \tag{3.3}$$

Tvrzení 3. *Pokud je $\hat{Y}$ nevychýlený odhad, pak je $\mathbb{E}w_i = 1,\ 1 \leq i \leq N$, a vzorec (3.3) přechází ve tvar*

$$\mathbb{E}\left(\hat{Y} - Y\right)^2 = V\left[\hat{Y}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \left(\mathbb{E}w_i w_j - 1\right). \tag{3.4}$$

V definici 1 jsme zavedli pravděpodobnosti zahrnutí prvního řádu. Tím rozumíme zahrnutí jednoho prvku do výběru. Pro další účely budeme potřebovat pravděpodobnost zahrnutí dvou prvků.

Definice 9. *Symbolem π_{ij} značíme pravděpodobnost zahrnutí i-tého a j-tého prvku do výběru, tedy*

$$\pi_{ij} = \sum_{s \ni U_i, U_j} p(s).$$

Tato pravděpodobnost se nazývá pravděpodobnost zahrnutí druhého řádu.

Tvrzení 4. *Pokud je $\hat{Y}$ prostý lineární odhad s vahami $w_i = \frac{I_i}{\pi_i}$, potom je nevychýlený a pro jeho rozptyl dostáváme z (3.4) vzorec*

$$V\left[\hat{Y}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \left(\frac{\mathbb{E}I_i I_j}{\pi_i \pi_j} - 1\right) = \sum_{i=1}^{N} y_i^2 \left(\frac{1}{\pi_i} - 1\right) + \sum_{\substack{1 \leq i,j \leq N \\ i \neq j}}\sum y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right). \tag{3.5}$$

Tvrzení 5. *V případě, že výběr má pevný rozsah, pak můžeme vzorec (3.5) přepsat do tvaru*

$$V_{YG}\left[\hat{Y}\right] = \frac{1}{2}\sum_{\substack{1 \leq i,j \leq N \\ i \neq j}}\sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \left(\pi_i \pi_j - \pi_{ij}\right). \tag{3.6}$$

Definice 10. *Vzorec na pravé straně v rovnosti (3.6) nazýváme Yatesovo-Grundyho formulí.*

Jak vidíme z rovnosti (3.3), střední kvadratická odchylka závisí i na hodnotách y_i náležících prvkům, které nejsou ve výběru. Pro konkrétní výběr ji neumíme spočítat. Vyjdeme z předpokladu, že známe pravděpodobnosti zahrnutí druhého řádu $\pi_{ij}, 1 \leq i, j \leq N$. Pakliže chceme odhadnout střední kvadratickou odchylku, budeme potřebovat následující tvrzení.

Tvrzení 6. *Statistika*

$$e\left(\hat{Y}(s) - Y\right)^2 = \sum_{i:U_i \in s,\, j:U_j \in s}\sum \frac{y_i y_j}{\pi_{ij}} \mathbb{E}\left[\left(w_i - 1\right)\left(w_j - 1\right)\right]$$

je nevychýleným odhadem střední čtvercové odchylky $\mathbb{E}\left(\hat{Y}(s) - Y\right)^2$ dané vzorcem (3.3).

Pro prostý lineární odhad definovaný vzorcem (3.2) máme následující tvrzení.

Tvrzení 7. *Je-li $\hat{Y}$ prostý lineární odhad s vahami $w_i = \frac{I_i}{\pi_i}$, pak*

$$e\left(\hat{Y}(s) - Y\right)^2 = \hat{V}\left[\hat{Y}\right] = \sum_{i:U_i \in s,\, j:U_j \in s} \sum \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right). \tag{3.7}$$

je nevychýleným odhadem rozptylu $V\left[\hat{Y}\right]$ daného vzorcem (3.4).

Speciálně dostáváme, že

$$\begin{aligned} e\left(\hat{Y}(s) - Y\right)^2 = \hat{V}_{YG}\left[\hat{Y}\right] &= \sum_{i:U_i \in s} \frac{{y_i}^2}{\pi_i} \left(\frac{1}{\pi_i} - 1\right) + \sum_{\substack{i:U_i \in s,\, j:U_j \in s \\ i \neq j}} \sum y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) \\ &= \frac{1}{2} \sum_{\substack{i:U_i \in s,\, j:U_j \in s \\ i \neq j}} \sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1\right) \end{aligned} \tag{3.8}$$

je nevychýleným odhadem rozptylu prostého lineárního odhadu s pevnou velikostí n. Tento odhad bývá označován jako Yatesův-Grundyho odhad rozptylu odhadu.

3.2 Rozšíření Horvitzova-Thompsonova odhadu na spojité populace

Následující odstavce byly vytvořeny na základě článku [3], kde rovněž nalezneme většinu vzorců, nebo jejich ideu. Horvitzův-Thompsonův odhad je takový, který za váhy bere hodnoty $w_i = \frac{I_i}{\pi_i}$. Pro naše účely předpokládáme, že populace $\mathbb{U}$ se skládá z nekonečně mnoha prvků, bodů, které můžeme popsat pomocí množiny $\mathcal{D} \subseteq \mathbb{R}^d$. Vlastnost, která nás bude zajímat v bodě x, budeme značit $y(x)$.

Naším cílem je odhadnout populační úhrn $Y = \int_{\mathcal{D}} y(x)\,\mathrm{d}x$. Pro zjednodušení uvažujme, že je vybrán vzorek o pevné velikosti n. Výběrový plán definujeme pomocí pravděpodobnostní míry P na $\mathcal{D}^n$, pro kterou existuje hustota f, neboli platí

$$P(B) = \int_B f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

pro každou měřitelnou $B \subseteq \mathcal{D}^n$. Vzorek s je vybrán náhodně z rozdělení P. Označme f_i marginální hustotu sdružené hustoty f, tj.

$$f_i(x) = \int_{\mathcal{D}^{n-1}} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n.$$

V analogii s diskrétním případem definujeme pravděpodobnosti zahrnutí $\pi(x)$ a budeme předpokládat, že jsou kladné pro každé $x \in \mathcal{D}$.

Definice 11. *Pro $x \in \mathcal{D}$ definujeme pravděpodobnost zahrnutí jako*

$$\pi(x) = \sum_{i=1}^{n} f_i(x).$$

Je zřejmé, že $\int_{\mathcal{D}} \pi(x)\,\mathrm{d}x = n$. Pro případy, kdy je π spojitá, existuje přirozená interpretace pro funkci π. V tomto případě můžeme $\pi(x)$ považovat za lokální míru počtu vybraných bodů na jednotku plochy.

Dále definujme funkce párové hustoty zahrnutí na $\mathcal{D} \times \mathcal{D}$.

Definice 12. *Pro $i \neq j$ nechť f_{ij} je marginální hustota i-té a j-té složky, neboli*

$$f_{ij}(x, x') = \int_{\mathcal{D}^{n-2}} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_{j-1}, x', x_{j+1}, \ldots, x_n) \\ \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_{j-1}\,\mathrm{d}x_{j+1} \cdots \mathrm{d}x_n,$$

pokud $i < j$. Pro $x, x' \in \mathcal{D}$ definujeme funkci párové hustoty zahrnutí jako

$$\pi(x, x') = \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \sum f_{ij}(x, x').$$

Pravděpodobnosti zahrnutí $\pi(x)$ a $\pi(x, x')$ hrají obdobnou roli ve spojitých případech jako pravděpodobnosti zahrnutí π_i a π_{ij} v konečných populacích.

Definice 13. *Horvitzův-Thompsonův odhad úhrnu Y, který uvažujeme pro nějaký vzorek s, je roven*

$$\hat{Y}(s) = \sum_{x \in s} \frac{y(x)}{\pi(x)}. \tag{3.9}$$

Nyní formulujeme základní tvrzení týkající se některých vlastností odhadů pro případ spojitých populací.

Tvrzení 8. *Pokud je funkce y buď omezená, nebo nezáporná, a $\pi(x) > 0$ pro každé $x \in \mathcal{D}$, potom odhad $\hat{Y}(s)$ definovaný vzorcem (3.9) je nestranný odhad úhrnu Y.*

Tvrzení 9. *Pokud je funkce y omezená, $\pi(x) > 0$ pro každé $x \in \mathcal{D}$ a $\int_{\mathcal{D}} \frac{1}{\pi(x)}\,\mathrm{d}x < \infty$, potom existuje $V\left[\hat{Y}\right]$ a je dán podobně jako v (3.5) vztahem*

$$V\left[\hat{Y}\right] = \int_{\mathcal{D}} \frac{y(x)^2}{\pi(x)}\,\mathrm{d}x + \int_{\mathcal{D}} \int_{\mathcal{D}} y(x)y(x') \left(\frac{\pi(x, x') - \pi(x)\pi(x')}{\pi(x)\pi(x')} \right) \mathrm{d}x\,\mathrm{d}x'. \tag{3.10}$$

Když navíc $\pi(x, x') > 0$ pro každé $x, x' \in \mathcal{D}$, pak pro odhad rozptylu analogicky ke vzorci (3.7) je

$$\hat{V}\left[\hat{Y}\right] = \sum_{x \in s} \left(\frac{y(x)}{\pi(x)} \right)^2 + \sum_{x, x' \in s} \sum y(x)y(x') \frac{\pi(x, x') - \pi(x)\pi(x')}{\pi(x)\pi(x')}.$$

Tyto vzorce jsou dostatečně obecné pro většinu aplikací.

Stejně tak jako v diskrétním případě existuje zjednodušení vzorce rozptylu $\hat{Y}(s)$ počítaného v rovnosti (3.10) na tvar známý jako Yatesův-Grundyho vzorec. Jeho odhad je dán podobně jako ve vztahu (3.8).

Definice 14. *Pravou stranu v rovnosti*

$$\hat{V}_{YG}\left[\hat{Y}\right] = \frac{1}{2}\sum_{\substack{x,x'\in s\\ x\neq x'}}\sum \frac{\pi(x)\pi(x') - \pi(x,x')}{\pi(x,x')}\left(\frac{y(x)}{\pi(x)} - \frac{y(x')}{\pi(x')}\right)$$

nazýváme Yatesovo-Grundyho odhadem.

Důležitý speciální případ dostaneme, když $f(x_1,\ldots,x_n) = g(x_1)\ldots g(x_n)$ pro nějakou pravděpodobnostní hustotu g na $\mathcal{D}$. To znamená, že vzorek s je tvořen n-ticí $x_1,\ldots,x_n$ nezávislých stejně rozdělených bodů v $\mathcal{D}$. Odhad úhrnu na základě i-tého vybraného bodu x_i je podíl hustoty vlastnosti a pravděpodobnostní hustoty g, tedy

$$\hat{Y}_i = \frac{y(x_i)}{g(x_i)}. \tag{3.11}$$

Definice 15. *Kombinovaný odhad n nezávislých výběrů je dán jako*

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i, \tag{3.12}$$

kde $\hat{Y}_i = \frac{y(x_i)}{g(x_i)}$, $i = 1,\ldots,n$.

Pravděpodobnosti zahrnutí jsou $\pi(x) = ng(x)$, tudíž (3.12) je speciálním případem (3.9). Protože $\pi(x,x') = n(n-1)g(x)g(x')$, lze zjednodušit vzorec 3.10.

Tvrzení 10. *Pro rozptyl kombinovaného odhadu definovaného ve vzorci (3.12) platí*

$$V\left[\hat{Y}\right] = \frac{1}{n}\left(\int_{\mathcal{D}}\frac{y(x)^2}{g(x)}\,\mathrm{d}x - Y^2\right). \tag{3.13}$$

Tento rozptyl se dá pro $n > 1$ nevychýleně odhadnout jako

$$\hat{V}\left[\hat{Y}\right] = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{Y}\right)^2}{n(n-1)}. \tag{3.14}$$

Kapitola 4

Výběrové plány pro konečné populace

V této kapitole představíme výběrové plány pro konečné populace. Hlavní myšlenky této kapitoly můžeme najít zejména ve [4] a [12]. Konečnou populací rozumíme množinu prvků $\mathbb{U} = \{U_x : x \in \mathcal{D}\}$, kde $\mathcal{D}$ je množina daná výčtem, tedy $\mathcal{D} = \{1, 2, \ldots, N\}$.

4.1 Plány se stejnými pravděpodobnostmi

Jak již název sám napovídá, bude se jednat o takové výběrové plány, kde pravděpodobnost zahrnutí π_i každého prvku U_i populace $\mathbb{U}$ bude stejná. Mezi takové plány patří například prostý náhodný výběr, systematický výběr, Bernoulliho výběr a další. Ve zkratce uveďme některá základní fakta k některým z nich.

4.1.1 Prostý náhodný výběr

Prostý náhodný výběr dělíme na dva základní typy. Prvním z nich je prostý náhodný výběr bez opakování, při němž se vybrané prvky nemohou v jednom vzorku opakovat. Druhým je prostý náhodný výběr s opakováním a zde počet opakování ve vzorku není nijak omezen, snad jen velikostí vzorku. Důležité je, že takový výběr má pevnou velikost n. Prostý náhodný výběr se vyznačuje tím, že všechny vzorky jsou stejně pravděpodobné. A obráceně, každý plán, ve kterém jsou všechny vzorky stejně pravděpodobné, je prostý náhodný výběr. Symbolem $p(s)$ značíme pravděpodobnost výběru vzorku s a symbolem π_i značíme pravděpodobnost zahrnutí prvku U_i populace do výběru s. V prostém náhodném výběru platí, že jsou obě tyto pravděpodobnosti konstantní. Obecně se budeme setkávat s případy, kdy n je mnohem menší než N.

Řekněme něco o základních vlastnostech a charakteristikách prostého náhodného výběru bez opakování. Počet možných vzorků velikosti n, které mohou být vybrány z populace $\mathbb{U}$ o velikosti N diskrétních prvků, je roven $|\Omega| = \frac{N!}{n!(N-n)!} = \binom{N}{n}$. Každý z těchto $|\Omega|$ možných vzorků má stejnou pravděpodobnost výběru $p(s)$, a ta je rovna $p(s) = \frac{1}{|\Omega|}$. Pro pravděpodobnosti zahrnutí prvního řádu platí

$$\pi_i = \frac{n}{N}, \quad 1 \leq i \leq N$$

a pro pravděpodobnosti zahrnutí druhého řádu platí

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \quad 1 \leq i \neq j \leq N.$$

Definice 16. *Výběrovým průměrem rozumíme náhodnou veličinu*

$$\bar{y} = \frac{1}{n} \sum_{i:U_i \in s} y_i.$$

Při dalších výběrech a vzorcích budeme používat výběrový průměr a populační průměr. Podívejme se blíže na základní vzorce pro lineární odhad úhrnu a pro jeho rozptyl. Rovnost (3.2) přejde v případě, že za výběrový plán volíme prostý náhodný výběr o rozsahu n, ve tvar

$$\hat{Y} = \sum_{i:U_i \in s} \frac{y_i}{\pi_i} = \frac{N}{n} \sum_{i:U_i \in s} y_i = \bar{y}N, \tag{4.1}$$

kde $\bar{y}$ je výběrový průměr. Odhad definovaný v předchozím vzorci je nevychýlený. Rozptyl počítaný pomocí vzorce (3.6) přechází ve tvar

$$\begin{aligned} V\left[\hat{Y}\right] = V\left[N\bar{y}\right] &= \frac{1}{2} \sum_{\substack{1 \leq i,j \leq N \\ i \neq j}} \sum \left(\frac{y_i}{n/N} - \frac{y_j}{n/N} \right)^2 \frac{n(N-n)}{N^2(N-1)} \\ &= \frac{1}{2} \sum_{\substack{1 \leq i,j \leq N \\ i \neq j}} \sum \left(y_i - y_j\right)^2 \frac{N-n}{(N-1)n} = \frac{N(N-n)}{(N-1)n} \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2, \end{aligned} \tag{4.2}$$

kde $\bar{Y}$ je populační průměr.

Pakliže chceme odhadnout populační průměr $\bar{Y}$, použijeme vzorec

$$\hat{\bar{Y}} = \frac{1}{N}\hat{Y} = \bar{y}.$$

Tento odhad je nevychýlený a jedná se o výběrový průměr s rozptylem

$$V\left[\bar{y}\right] = \frac{1}{N^2} V\left[\hat{Y}\right] = \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 .$$

Výraz

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 \tag{4.3}$$

odpovídá rozptylu v popisné statistice. Tímto koeficientem charakterizujeme variabilitu sledované charakteristiky v populaci. Rozptyl výběrového průměru tedy můžeme psát jako

$$V\left[\bar{y}\right] = \frac{\sigma_y^2}{n}\left(1 - \frac{n}{N}\right), \tag{4.4}$$

kde $\left(1 - \frac{n}{N}\right)$ je tzv. konečnostní násobitel. Ten pro $n \ll N$ můžeme pokládat za 1.

Poslední charakteristikou, která nás bude zajímat, je střední čtvercová odchylka. Dosadíme-li do vzorce (3.8) za π_i hodnotu $\frac{n}{N}$ a za π_{ij} hodnotu $\frac{n(n-1)}{N(N-1)}$, dostaneme odhad rozptylu odhadu daného vzorcem (4.2), který bude tvaru

$$\hat{V}\left[\hat{Y}\right] = \frac{1}{2} \sum_{i:U_i \in s,\, j:U_j \in s} \sum \left(y_i - y_j\right)^2 \frac{N-n}{(n-1)n^2} = \frac{N(N-n)}{(n-1)n} \sum_{i:U_i \in s} \left(y_i - \bar{y}\right)^2 . \tag{4.5}$$

Alternativní možností, jak získat odhad rozptylu, je hledat jej přímo. Stačí si uvědomit, že ve vzorci (4.2) potřebujeme odhadnout pouze výraz daný vzorcem (4.3), který můžeme pro prostý náhodný výběr nevychýleně odhadnout jako

$$\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_{i:U_i \in s} \left(y_i - \bar{y}\right)^2 . \tag{4.6}$$

Celkově tedy dostáváme

$$\hat{V}\left[\hat{Y}\right] = \frac{N(N-n)}{n} \hat{\sigma}_y^2 = \frac{N(N-n)}{n} \frac{1}{n-1} \sum_{i:U_i \in s} \left(y_i - \bar{y}\right)^2 ,$$

což je nevychýlený odhad rozptylu odhadu vlastnosti Y totožný se vzorcem (4.5).

V případě, že bychom chtěli nevychýleně odhadnout rozptyl výběrového průměru $\bar{y}$ daného vzorcem (4.4), použijeme vzorce

$$\hat{V}\left[\bar{y}\right] = \frac{1}{n}\left(1 - \frac{n}{N}\right) \hat{\sigma}_y^2 .$$

Druhým typem je prostý náhodný výběr s opakováním. Pro velké populace je pravděpodobnost, že bude nějaký prvek vybrán vícekrát, malá a model se blíží tomu

bez opakování. Obecně platí, že odhad pro plán s opakováním je méně přesný než obdobný plán bez opakování. Pro pravděpodobnost výběru platí $p(s) = \frac{1}{|\Omega|} = \frac{1}{N^n}$. Pravděpodobnost výběru prvku U_i je $p_i = \frac{1}{N}$. Pro pravděpodobnosti zahrnutí platí

$$\pi_i = 1 - (1 - p_i)^n = 1 - \left(1 - \frac{1}{N}\right)^n.$$

Jako nevychýlený odhad úhrnu Y můžeme vzít Horvitzův-Thompsonův odhad $\hat{Y}$ daný vzorcem (3.2), ale obvyklejší odhad je

$$\hat{Y}_{opak} = \frac{1}{n} \sum_{i:U_i \in s} \frac{y_i}{p_i} = \frac{N}{n} \sum_{i:U_i \in s} y_i = N\bar{y}, \tag{4.7}$$

což je totožný odhad co do tvaru s Horvitzovo-Thompsonovým odhadem pod prostým náhodným výběrem bez opakování. Je důležité si uvědomit, že obecně $\hat{Y} \neq \hat{Y}_{opak}$. Rozptyl odhadu $\hat{Y}_{opak}$ je

$$V\left[\hat{Y}_{opak}\right] = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2, \tag{4.8}$$

který pod prostým náhodným výběrem s opakováním přechází v

$$V\left[\hat{Y}_{opak}\right] = V\left[N\bar{y}\right] = N^2 \frac{N-1}{N} \frac{\sigma_y^2}{n}.$$

Tyto rozptyly se dají nevychýleně odhadnout jako

$$\hat{V}\left[\hat{Y}_{opak}\right] = \frac{1}{n(n-1)} \sum_{i:U_i \in s} \left(\frac{y_i}{p_i} - \hat{Y}_{opak}\right)^2,$$

respektive

$$\hat{V}\left[N\bar{y}\right] = N^2 \frac{\hat{\sigma}_y^2}{n} = \frac{N^2}{n(n-1)} \sum_{i:U_i \in s} \left(y_i - \bar{y}\right)^2.$$

Pro úplnost uveďme rozptyl a odhad rozptylu pro výběrový průměr $\bar{y}$. Rozptyl je dán vzorcem

$$V\left[\bar{y}\right] = \frac{V\left[\hat{Y}_{opak}\right]}{N^2} = \frac{1}{nN^2} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 = \frac{N-1}{N} \frac{\sigma_y^2}{n}$$

a odhad tohoto rozptylu je dán jako

$$\hat{V}\left[\bar{y}\right] = \frac{\hat{\sigma}_y^2}{n} = \frac{1}{n(n-1)} \sum_{i:U_i \in s} \left(y_i - \bar{y}\right)^2,$$

kde σ_y^2 a $\hat{\sigma}_y^2$ jsou dány vzorci (4.3) a (4.6).

4.1.2 Systematický výběr

Předpokládejme, že jednotky, ze kterých máme vybrat vzorek, jsou uspořádány nezávisle na vlastnosti zkoumání do nějaké posloupnosti. Podstata systematického výběru je v tom, že vybíráme každou a-tou jednotku této posloupnosti. Výběrový interval a reprezentuje počet jednotek mezi po sobě jdoucími prvky ve výběru. Výběr začíná náhodně vybraným prvkem mezi prvními a jednotkami. V anglicky psané literatuře bývá tento výběr označován jako 1-in-a systematický výběr. Při plánování výběru může nastat situace, kdy požadujeme určitou velikost vzorku. V takovém případě se výběrový interval a dopočítá. Běžnějším případem je však ten, kdy a známe předem. Pokud budeme postupovat skrz vzorek, pak velikost populace nemusí být nutně známa předem, bude důsledkem výběru. Velikost jednotlivých vzorků vzniklých systematickým výběrem se nebude lišit o více než jeden prvek. Jde vlastně o takový posun výběru a základní otázka je ta, kde začneme.

Symbolem Ω značíme množinu všech možných vzorků a velikost této množiny je $|\Omega| = a$. Pravděpodobnost výběru každého vzorku z a možných je $p(s) = \frac{1}{a}$. To je v souladu s myšlenkou, že výběr závisí jen na výběru první jednotky, zbytek je pak už dán. Ve většině případů se setkáme se situací, kdy se bude $n \cdot a$ lišit od N o nějakou celočíselnou hodnotu, tedy $N = n \cdot a + c$, kde $c < a$. Kdykoliv bude $c \neq 0$, pak budeme mít nepatrný rozdíl ve velikosti vzorku. Počet prvků je buď $n = \left\lfloor \frac{N}{a} \right\rfloor$, nebo $n = \left\lfloor \frac{N}{a} \right\rfloor + 1$, kde symbolem $\lfloor \cdot \rfloor$ značíme dolní celou část. V porovnání s prostým náhodným výběrem má systematický výběr mnohem hladší rozdělení vzorku. Pravděpodobnosti zahrnutí každé jednotky jsou také $\frac{1}{a}$, tedy v systematickém výběrovém plánu je $\pi_i = p(s)$ pro každý prvek U_i. Populační úhrn lze nevychýleně odhadnout pomocí Horvitzova-Thompsonova odhadu jako

$$\hat{Y}_{sys} = a \sum_{i:U_i \in s} y_i. \tag{4.9}$$

Pakliže tento vztah vydělíme N, dostaneme vzorec pro nevychýlený odhad populačního průměru $\bar{Y}$ daný jako

$$\hat{\bar{Y}}_{sys} = \frac{a}{N} \sum_{i:U_i \in s} y_i,$$

který se liší od výběrového průměru $\bar{y}$, kdykoliv $N \neq n \cdot a$. Je zřejmé, že $\bar{y}$ je nevychýlený odhad průměru $\bar{Y}$, jen když $N = n \cdot a$, zatímco $\hat{\bar{Y}}_{sys}$ je nevychýlený odhad průměru $\bar{Y}$ vždy.

Nechť Y_s značí úhrn vzorku s, tedy

$$Y_s = \sum_{i:U_i \in s} y_i.$$

Potom odhad celkového úhrnu můžeme psát jako

$$\hat{Y}_{sys} = a \cdot Y_s.$$

Rozptyl tohoto odhadu počítáme vzorcem

$$\begin{aligned} V\left[\hat{Y}_{sys}\right] &= V\left[a \cdot Y_s\right] = \frac{1}{a}\sum_{s\in\Omega}\left(a \cdot Y_s - Y\right)^2 \\ &= a \cdot \sum_{s\in\Omega} Y_s^2 - Y^2 = a \cdot \sum_{s\in\Omega}\left(Y_s - \bar{Y}_a\right)^2, \end{aligned} \tag{4.10}$$

kde $\bar{Y}_a = Y/a$. Odpovídající rozptyl $\hat{\bar{Y}}_{sys}$ je

$$V\left[\hat{\bar{Y}}_{sys}\right] = \frac{V\left[\hat{Y}_{sys}\right]}{N^2} = \frac{a}{N^2}\sum_{s\in\Omega}\left(Y_s - \bar{Y}_a\right)^2.$$

4.1.3 Cyklický systematický výběr

V předchozím odstavci jsme uvedli, že výběr vzorku skončí, pokud je vyčerpána velikost vzorku. Variantou k tomuto výběru je cyklický systematický výběr umožňující ve výběru pokračovat znovu od začátku tak, jako kdyby po posledním prvku populace následoval opět první prvek. Ke splnění podmínky pozitivnosti pravděpodobnosti zahrnutí výběr začíná náhodným výběrem prvního prvku ze všech N jednotek populace, nejen z prvních a prvků. Jako důsledek máme, že počet možných vzorků je $|\Omega| = N$. Výhodou je, že je odstraněna rozdílnost velikostí mezi možnými systematickými výběry. Platí tedy, že výběrový průměr $\bar{y}$ je nevychýlený odhad populačního průměru $\bar{Y}$. Pro cyklický systematický výběr velikosti n s pravděpodobnostmi zahrnutí $\pi_i = \frac{n}{N}$ platí $\hat{\bar{Y}} = \frac{NY_s}{n}$ a rozptyl je dán

$$V\left[\hat{\bar{Y}}\right] = V\left[\frac{NY_s}{n}\right] = \frac{1}{N}\sum_{s\in\Omega}\left(\frac{NY_s}{n} - Y\right)^2.$$

Cyklický systematický výběr je vhodný pro ty situace, kdy výběrový poměr $\frac{n}{N}$ je malý. Odhad úhrnu při cyklickém výběru je obvykle méně precizní než ten při 1-in-a systematickém výběru.

4.1.4 Bernoulliho výběr

Tato metoda odpovídá výběru se stejnými pravděpodobnostmi, bez opakování, s prvky populace vybíranými nezávisle a s konstantními pravděpodobnostmi zahrnutí π. Platí tedy $\pi_i = \pi$ pro každý prvek populace. Velikost vzorku se bude mezi

různými výběry lišit. Pro její střední hodnotu platí $\mathbb{E}[n] = N\pi$. Pravděpodobnost výběru vzorku s o velikosti n je

$$p(s) = \pi^n (1-\pi)^{N-n}.$$

Horvitzův-Thompsonův odhad úhrnu daný vzorcem (3.2) se zjednodušuje na

$$\hat{Y} = \frac{1}{\pi} \sum_{i:U_i \in s} y_i = \frac{N}{\mathbb{E}[n]} \sum_{i:U_i \in s} y_i,$$

kde $\mathbb{E}[n]$ je střední hodnota velikosti vzorku. Tento odhad je nevychýlený, ale není konzistentní. Rozptyl odhadu $\hat{Y}$ je dán jako

$$V\left[\hat{Y}\right] = \frac{1-\pi}{\pi} \sum_{i=1}^{N} y_i^2.$$

Zvláštností tohoto odhadu je to, že narozdíl od výše zmíněných odhadů jeho rozptyl s rostoucím počtem jednotek ve výběru neklesá, je po celou dobu konstantní. Nevychýlený odhad rozptylu odhadu úhrnu je dán vzorcem

$$\hat{V}\left[\hat{Y}\right] = \frac{1-\pi}{\pi^2} \sum_{i:U_i \in s} y_i^2.$$

Odhad populačního průměru počítáme jako

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1}{\pi N} \sum_{i:U_i \in s} y_i = \frac{1}{\mathbb{E}[n]} \sum_{i:U_i \in s} y_i.$$

Tento odhad je nevychýlený, ale není konzistentní odhad populačního průměru $\bar{Y}$. Rozptyl odhadu počítáme dle vzorce

$$V\left[\hat{\bar{Y}}\right] = \frac{V\left[\hat{Y}\right]}{N^2},$$

jenž můžeme nevychýleně odhadnout pomocí

$$\hat{V}\left[\hat{\bar{Y}}\right] = \frac{\hat{V}\left[\hat{Y}\right]}{N^2} = \frac{1-\pi}{\pi^2 N^2} \sum_{i:U_i \in s} y_i^2.$$

4.2 Plány s nestejnými pravděpodobnostmi

V následujících odstavcích popíšeme hlavní body o výběrových plánech, ve kterých pravděpodobnosti zahrnutí prvků U_i populace $\mathbb{U}$ nebudou stejné. Největším důvodem pro výběr s nestejnými pravděpodobnostmi zahrnutí je zvýšení preciznosti odhadu.

4.2.1 Výběr dle seznamu

Tento výběr je analogický k prostému náhodnému výběru s opakováním. Jde tedy o výběr s pevnou velikostí vzorku n, který připouští opakování prvků ve vzorku. Pro pravděpodobnosti výběru i-tého prvku U_i platí, že $0 < p_i < 1$ a $\sum_{i=1}^{N} p_i = 1$. Obecně se jednotkám s větší hodnotou sledované vlastnosti přiřazuje větší pravděpodobnost výběru do vzorku. Tak jako v prostém náhodném výběru s opakováním, je odhad úhrnu Y daný vzorcem (4.7), tedy

$$\hat{Y}_{opak} = \frac{1}{n} \sum_{i:U_i \in s} \frac{y_i}{p_i},$$

a rozptyl $V\left[\hat{Y}_{opak}\right]$ lze počítat pomocí vzorce (4.8). Pro odhad populačního průměru $\bar{Y}$ můžeme použít nevychýlený odhad $\hat{\bar{Y}}_{opak} = \frac{\hat{Y}_{opak}}{N}$ s rozptylem $V\left[\hat{\bar{Y}}_{opak}\right] = \frac{V\left[\hat{Y}_{opak}\right]}{N^2}$.

4.2.2 Poissonův výběr

Pro libovolnou posloupnost čísel $p_1, p_2, \ldots, p_N$, $0 \leq p_i \leq 1$, $1 \leq i \leq N$, definujeme výběrový plán pravděpodobnostmi

$$p(s) = \prod_{i:U_i \in s} p_i \prod_{i:U_i \in \mathbb{U} \setminus s} (1 - p_i), \quad s \subseteq \mathbb{U},$$

tento výběrový plán nazveme poissonovským výběrem. Hlavní význam poissonovského výběru je teoretický. Pomocí tohoto výběru je možné definovat jiné výběrové plány. Výhodou je nezávislost indikátorů zahrnutí a také jeho jednoduchá závislost pravděpodobností zahrnutí na konstantách $p_1, p_2, \ldots, p_N$.

Pro pravděpodobnosti zahrnutí při poissonovském výběru platí

$$\begin{aligned} \pi_i &= p_i, \quad 1 \leq i \leq N, \\ \pi_{ij} &= p_i \cdot p_j, \quad 1 \leq i \neq j \leq N. \end{aligned}$$

Horvitzův-Thompsonův odhad úhrnu je dle (3.2) roven

$$\hat{Y} = \sum_{i:U_i \in s} \frac{y_i}{p_i},$$

což je nestranný, ale neefícientní odhad úhrnu Y. Rozptyl pro Poissonův výběr je roven (3.5), kde z nezávislosti indikátorů je druhý sčítanec roven nule, dostáváme tedy

$$V\left[\hat{Y}\right] = \sum_{i=1}^{N} y_i^2 \left(\frac{1}{p_i} - 1\right).$$

4.3 Dvouúrovňový výběr

Je-li základní soubor rozsáhlý, potom z praktických důvodů přistupujeme k výběru ve více úrovních. Skupiny původních prvků populace pokládáme za nové jednotky a z nich vytvoříme nový základní soubor. Takto bychom mohli pokračovat i dále, my se zde však omezíme pouze na dvouúrovňový výběr. Výběr posléze probíhá nejprve od jednotek stojících nejvýše v popsané struktuře. Jednotky, ze kterých vybíráme jako první, se nazývají primární, česky spíše prvoúrovňové, výběrové jednotky. Pro jednoduchost je v textu budeme značit zkratkou PVJ. Každou takto vybranou jednotku považujeme za základní soubor pro další výběr. Tento další výběr, v pořadí druhý, nám dá vznik druhoúrovňových výběrových jednotek. Ty budeme opět pro jednoduchost označovat DVJ. Výše jsme popsali jak vznik první úrovně výběru, tak i vznik druhé úrovně výběru. V případě více úrovní pokračujeme ve výběrech, až se dostaneme k původním prvkům populace. Ty nesou námi sledovanou hodnotu charakteristiky. Budeme se zabývat pouze dvouúrovňovým výběrem, veškeré rozšíření na více úrovní je více než intuitivní. Speciálním případem dvouúrovňového výběru, v němž chybí druhá úroveň, tedy vybrané PVJ jsou vyšetřeny celé, je výběr, kterému se říká skupinkový výběr. O tomto výběru řekneme více v další podkapitole.

Předpokládejme, že máme populaci N prvků rozdělenou do PVJ, jejich počet je M. Značíme je přirozeně $\mathbb{U}_1, \mathbb{U}_2, \ldots, \mathbb{U}_M$. Počet prvků v subpopulaci $\mathbb{U}_i$ je N_i. Je tedy zřejmé, že

$$N = \sum_{i=1}^{M} N_i.$$

Princip výběru je takový, že na první úrovni vybereme m PVJ. Vznikne tak výběrový soubor označovaný jako s^I. Při druhém výběru vybíráme z každé vybrané PVJ soubor DVJ. Symbolem s_i^{II} označujeme výběrový soubor DVJ vybraných při druhém výběru z i-té PVJ vybrané v první úrovni. Celkový soubor s je dán jako sjednocení těchto druhých výběrů přes všechny vybrané PVJ, tedy

$$s = \bigcup_{i:\mathbb{U}_i \in s^I} s_i^{II}.$$

Dalším předpokladem je to, že v každé vybrané PVJ se druhý výběr provádí nezávisle na výběru v ostatních PVJ. Pro přehlednost a lepší pochopení poznamenejme, že v následujícím budeme indexy i a j značit jednotky odpovídající množině PVJ a indexy k a l budeme značit jednotky odpovídající množině DVJ. Pro pravděpodobnosti zahrnutí prvku U_k do vzorku platí

$$\pi_k = \pi_i^I \cdot \pi_{k|i}^{II}.$$

Pro pravděpodobnosti zahrnutí dvou jednotek U_k a U_l do vzorku platí

$$\begin{aligned}
\pi_{kl} &= \pi_i^I \cdot \pi_{k|i}^{II}, \quad k = l,\, U_k \in \mathbb{U}_i, \\
\pi_{kl} &= \pi_i^I \cdot \pi_{kl|i}^{II}, \quad k \neq l,\, U_k,\, U_l \in \mathbb{U}_i, \\
\pi_{kl} &= \pi_{ij}^I \cdot \pi_{k|i}^{II} \cdot \pi_{l|j}^{II}, \quad k \neq l,\, U_k \in \mathbb{U}_i,\, U_l \in \mathbb{U}_j,\, i \neq j.
\end{aligned}$$

Pro úhrn v i-té PVJ počítaný jako

$$Y_i = \sum_{k:U_k \in \mathbb{U}_i} y_k$$

můžeme použít Horvitzova-Thompsonova odhadu, který je nevychýlený, a tvaru

$$\hat{Y}_i = \sum_{k:U_k \in s_i^{II}} \frac{y_k}{\pi_{k|i}^{II}}.$$

Teoretický rozptyl v i-té PVJ je dán jako

$$V_i = \sum_{k:U_k \in \mathbb{U}_i} \sum_{l:U_l \in \mathbb{U}_i} \frac{\pi_{kl|i}^{II} - \pi_{k|i}^{II}\pi_{l|i}^{II}}{\pi_{k|i}^{II}\pi_{l|i}^{II}} y_k y_l$$

a odhad tohoto rozptylu v i-té PVJ je

$$\hat{V}_i = \sum_{k:U_k \in s_i^{II}} \sum_{l:U_l \in s_i^{II}} \frac{\pi_{kl|i}^{II} - \pi_{k|i}^{II}\pi_{l|i}^{II}}{\pi_{kl|i}^{II}\pi_{k|i}^{II}\pi_{l|i}^{II}} y_k y_l.$$

Ve dvouúrovňovém výběru Horvitzův-Thompsonův odhad populačního úhrnu

$$Y = \sum_{k=1}^{N} y_k = \sum_{i=1}^{M} Y_i$$

je roven

$$\hat{Y}_{2st} = \sum_{i:\mathbb{U}_i \in s^I} \frac{\hat{Y}_i}{\pi_i^I}. \tag{4.11}$$

Rozptyl tohoto odhadu je dán jako

$$V\left[\hat{Y}_{2st}\right] = V_{PVJ} + V_{DVJ}. \tag{4.12}$$

Pro jednotlivé složky platí

$$V_{PVJ} = \sum_{i=1}^{M} \sum_{j=1}^{M} \left(\pi_{ij}^I - \pi_i^I \pi_j^I\right) \frac{Y_i}{\pi_i^I} \frac{Y_j}{\pi_j^I},$$

respektive

$$V_{DVJ} = \sum_{i=1}^{M} \frac{V_i}{\pi_i^I}.$$

Tyto rozptyly můžeme odhadnout jako

$$\hat{V}_{PVJ} = \sum_{i:\mathbb{U}_i \in s^I} \sum_{j:\mathbb{U}_j \in s^I} \frac{\pi_{ij}^I - \pi_i^I \pi_j^I}{\pi_{ij}^I} \frac{\hat{Y}_i}{\pi_i^I} \frac{\hat{Y}_j}{\pi_j^I} - \sum_{i:\mathbb{U}_i \in s^I} \frac{1 - \pi_i^I}{\left(\pi_i^I\right)^2} \hat{V}_i$$

a

$$\hat{V}_{DVJ} = \sum_{i:\mathbb{U}_i \in s^I} \frac{\hat{V}_i}{\left(\pi_i^I\right)^2},$$

což dohromady dává

$$\hat{V}\left[\hat{Y}_{2st}\right] = \sum_{i:\mathbb{U}_i \in s^I} \sum_{j:\mathbb{U}_j \in s^I} \frac{\pi_{ij}^I - \pi_i^I \pi_j^I}{\pi_{ij}^I} \frac{\hat{Y}_i}{\pi_i^I} \frac{\hat{Y}_j}{\pi_j^I} + \sum_{i:\mathbb{U}_i \in s^I} \frac{\hat{V}_i}{\pi_i^I}.$$

Uvažujme nyní, že provádíme prostý náhodný výběr v obou úrovních výběru. Tedy nejdříve vybíráme PVJ. Těch vybereme m z celkového počtu M. Poté v každé z m vybraných PVJ vybereme DVJ. Počet vybraných DVJ z i-té PVJ bude n_i a celkový počet prvků v i-té PVJ je N_i. Výše uvedené vzorce se tedy zjednoduší. Odhad úhrnu daný vzorcem (4.11) na

$$\hat{Y}_{2st} = \frac{M}{m} \sum_{i:\mathbb{U}_i \in s^I} \hat{Y}_i = \frac{M}{m} \sum_{i:\mathbb{U}_i \in s^I} \frac{N_i}{n_i} \sum_{k:U_k \in s_i^{II}} y_k = \frac{M}{m} \sum_{i:\mathbb{U}_i \in s^I} N_i \bar{y}_i, \tag{4.13}$$

kde $\bar{y}_i$ je výběrový průměr příslušný s_i^{II}. Rozptyl v tomto případě přejde z rovnosti (4.12) na tvar

$$\begin{aligned} V\left[\hat{Y}_{2st}\right] = {} & \frac{M(M-m)}{m} \frac{1}{M-1} \sum_{i=1}^{M} \left(\sum_{k:U_k \in \mathbb{U}_i} y_k - \sum_{i=1}^{M} \frac{\sum_{l:U_l \in \mathbb{U}_i} y_l}{M} \right)^2 \\ & + \frac{M}{m} \sum_{i=1}^{M} \frac{N_i \left(N_i - n_i\right)}{n_i} \frac{1}{N_i - 1} \sum_{k:U_k \in \mathbb{U}_i} \left(y_k - \frac{\sum_{l:U_l \in \mathbb{U}_i} y_l}{N_i} \right)^2. \end{aligned} \tag{4.14}$$

Tento rozptyl můžeme odhadnout jako

$$\begin{aligned} \hat{V}\left[\hat{Y}_{2st}\right] = {} & \frac{M(M-m)}{m} \frac{1}{m-1} \sum_{i:\mathbb{U}_i \in s^I} \left(\frac{N_i}{n_i} \sum_{k:U_k \in s_i^{II}} y_k - \frac{\sum_{i:\mathbb{U}_i \in s^I} \frac{N_i}{n_i} \sum_{l:U_l \in s_i^{II}} y_l}{m} \right)^2 \\ & + \frac{M}{m} \sum_{i:\mathbb{U}_i \in s^I} \frac{N_i \left(N_i - n_i\right)}{n_i} \frac{1}{n_i - 1} \sum_{k:U_k \in s_i^{II}} \left(y_k - \frac{\sum_{l:U_l \in s_i^{II}} y_l}{n_i} \right)^2. \end{aligned} \tag{4.15}$$

4.4 Skupinkový výběr

Speciálním případem dvouúrovňového výběru je ten, ve kterém se ve druhé úrovni provádí kompletní šetření. Proč právě skupinkový výběr nám bude jasné, pokud si představíme, jak takový výběr vypadá. Výběr, nebo spíše měření DVJ, zde probíhá v rámci prvků, které jsou k sobě nějak přidruženy, tedy po skupinkách. Nechť populace $\mathbb{U} = \{U_1, U_2, \ldots, U_N\}$ je rozdělena do subpopulací PVJ nazývaných skupinky. Ty jsou značeny $\mathbb{U}_1, \mathbb{U}_2, \ldots, \mathbb{U}_M$. Z této populace je vybrán vzorek. Počet prvků subpopulace $\mathbb{U}_i$ je N_i. Tedy platí

$$\mathbb{U} = \bigcup_{i=1}^{M} \mathbb{U}_i$$

a

$$N = \sum_{i=1}^{M} N_i.$$

Skupinkový výběr je prováděn tak, že v první fázi je vybrán vzorek s^I skupinek o velikosti m z celkového počtu M. Ve druhé fázi je každý prvek ve vybrané skupince pozorován a změřen. Pravděpodobnosti zahrnutí na úrovni skupinek jsou jako obvykle značeny π_i^I a π_{ij}^I, kde $\pi_{ii}^I = \pi_i^I$. Na úrovni prvků původní populace máme $\pi_k = \pi_i^I$ pro každý prvek U_k ve skupince $\mathbb{U}_i$. Dále pak $\pi_{kl} = \pi_i^I$, když oba prvky U_k a U_l náleží stejné skupince $\mathbb{U}_i$ a poslední možností je $\pi_{kl} = \pi_{ij}^I$, když U_k a U_l náleží odlišným skupinkám $\mathbb{U}_i$ a $\mathbb{U}_j$. Úhrn v i-té skupince lze počítat jako

$$Y_i = \sum_{k:U_k \in \mathbb{U}_i} y_k, \quad i = 1, \ldots, M.$$

Horvitzův-Thompsonův odhad úhrnu $Y = \sum_{i=1}^{M} Y_i$ je

$$\hat{Y} = \sum_{i:\mathbb{U}_i \in s^I} \frac{Y_i}{\pi_i^I}$$

s rozptylem

$$V\left[\hat{Y}\right] = \sum_{i=1}^{M} \sum_{j=1}^{M} \left(\pi_{ij}^I - \pi_i^I \pi_j^I\right) \frac{Y_i}{\pi_i^I} \frac{Y_j}{\pi_j^I}.$$

Tento rozptyl odhadujeme jako

$$\hat{V}\left[\hat{Y}\right] = \sum_{i:\mathbb{U}_i \in s^I} \sum_{j:\mathbb{U}_j \in s^I} \frac{\pi_{ij}^I - \pi_i^I \pi_j^I}{\pi_i^I \pi_j^I} \frac{Y_i}{\pi_i^I} \frac{Y_j}{\pi_j^I}.$$

V případě pevné velikosti vzorku můžeme použít Yatesovo-Grundyho vzorec definovaný vzorcem (3.6).

4.5 Stratifikovaný výběr

Stratifikovaný výběr je specifický tím, že v něm úmyslně rozdělujeme populaci $\mathbb{U}$ do dvou či více menších subpopulací. Tyto subpopulace se nepřekrývají, jsou disjunktní a jsou nazývány strata, odtud název stratifikovaný výběr. Základní soubor tedy můžeme považovat za disjunktní sjednocení konečného počtu podsouborů. Ke každému stratu přistupujeme jednotlivě. Jako obvykle jedním z cílených parametrů bude populační úhrn značený Y, případně další populační vlastnosti a parametry.

Stratifikovaný výběr a jeho vznik byl motivován požadavkem odhadu úhrnu nebo průměru nějaké vlastnosti v celkové populaci, kde stejnou vlastnost můžeme vyšetřit pro každé stratum zvlášť. Stratifikace populace před výběrem vzorku je často výhodou, protože umožňuje tvůrci výběru přizpůsobit plán dle potřeb a vlastností každého strata. Vzhledem k tomu, že k výběru dochází nezávisle uvnitř každého strata, výběrový plán se může mezi straty lišit. Jinou motivaci bychom mohli nalézt v administrativní výhodě. Např. prvky, které jsou zájmem ankety, mohou být v různých stratech kontrolovány různými osobami a to také přináší jistou výhodu. Ze statistického hlediska je stratifikace velmi užitečný nástroj ke zvýšení preciznosti odhadovaných populačních parametrů. Růst preciznosti je způsoben případy, kdy homogenita vlastnosti ve stratu je větší než v nestratifikované populaci. Jinými slovy je možné odhadnout úhrn Y více precizně se stratifikovaným výběrovým plánem než s nestratifikovaným plánem.

Budeme předpokládat, že v každém stratu aplikujeme zvolený výběrový plán nezávisle na ostatních stratech. Kritérium stratifikace závisí na okolnostech. Nechť H značí počet strat. Každý prvek $U_i,\ i = 1, \ldots, N$ populace $\mathbb{U}$ musí být přiřazen právě do jedné subpopulace. Symbolem $\mathbb{U}_h$ budeme značit subpopulaci náležící stratu h, kde $h = 1, \ldots, H$. Výběrový soubor získaný v h-té oblasti značíme s_h. Je zřejmé, že $s_h \subseteq \mathbb{U}_h,\ h = 1, \ldots, H$. Celkový výběrový soubor je sjednocením oblastních výběrů $s = \bigcup_{h=1}^{H} s_h$. Oba tyto aspekty, tedy specifikace stratifikace i počet strat H, závisí na rozhodnutí tvůrce plánu. Nechť N značí počet jednotek v populaci $\mathbb{U}$. Počet jednotek v subpopulaci $\mathbb{U}_h$ budeme značit N_h. Tedy z principu stratifikace máme

$$N = N_1 + N_2 + \cdots + N_H = \sum_{h=1}^{H} N_h.$$

Položme

$$Y_h = \sum_{i:U_i \in \mathbb{U}_h} y_i, \quad h = 1, 2, \ldots, H.$$

Hodnotu Y_h nazýváme stratifikovaný úhrn a udává nám celkové množství vlastnosti v h-tém stratu. Poznamenejme, že sčítáme přes všech N_h prvků ve stratu. Celkový

úhrn vlastnosti ve všech stratech je

$$Y = \sum_{i=1}^{N} y_i = \sum_{h=1}^{H} Y_h.$$

Definice 17. *Lineárním odhadem při stratifikovaném výběru nazýváme náhodnou veličinu*

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h,$$

kde $\hat{Y}_h$ jsou lineární odhady stratifikovaných úhrnů.

Pro každý jednotlivý odhad úhrnu ve stratu h ze vztahu (3.1) platí

$$\hat{Y}_h = \sum_{i:U_i \in s_h} y_i w_i.$$

Uvažujme výběrovou strategii, ve které v každém stratu požadujeme odhad každého úhrnu Y_h pomocí Horvitzova-Thompsonona odhadu $\hat{Y}_h$. Výběrový plán nebudeme nikterak specifikovat, protože může být v každém stratu odlišný. Předpokládejme, že n_h je velikost vzorku stanoveného výběrovým plánem v subpopulaci $\mathbb{U}_h$. Je zřejmé, že $n_h \leq N_h$ a že celková velikost stratifikovaného vzorku je

$$n = \sum_{h=1}^{H} n_h.$$

Nechť Ω_h značí množinu všech možných vzorků a $|\Omega_h|$ značí počet možných rozdílných vzorků pod výběrovým plánem pro subpopulaci $\mathbb{U}_h$. Množinu všech možných vzorků intuitivně značíme Ω a počet rozdílných vzorků pod stratifikovaným plánem je pak

$$|\Omega| = \prod_{h=1}^{H} |\Omega_h|.$$

Tvrzení 11. *Při volbě vah tak, že $\mathbb{E}\left[\hat{Y}_h\right] = Y_h$, ihned vidíme, že i $\mathbb{E}\left[\hat{Y}\right] = Y$. Pro nevychýlený odhad $\hat{Y}$ je $\mathbb{E}\left(\hat{Y} - Y\right)^2 = V\left[\hat{Y}\right]$ a vzhledem k nezávislosti výběrů v jednotlivých stratech dostáváme*

$$V\left[\hat{Y}\right] = \sum_{h=1}^{H} V\left[\hat{Y}_h\right], \tag{4.16}$$

kde $V\left[\hat{Y}_h\right]$ *je dán vzorcem (3.4) použitým na h-té stratum, tedy*

$$V\left[\hat{Y}_h\right] = \sum_{i:U_i\in\mathbb{U}_h}\sum_{j:U_j\in\mathbb{U}_h} y_i y_j \left(\mathbb{E}w_i w_j - 1\right).$$

Ze vzorce (4.16) je patrné, že čím jsou jednotlivé rozptyly odhadů ve stratech menší, tím je menší celkový rozptyl odhadu $\hat{Y}$. Je tedy výhodné mít zkoumaná strata homogenní vzhledem ke zkoumané vlastnosti. Z odvození vzorce (4.16) je zřejmé, že pro odhad rozptylu odhadu úhrnu dostáváme

$$\hat{V}\left[\hat{Y}\right] = \sum_{h=1}^{H} \hat{V}\left[\hat{Y}_h\right]. \tag{4.17}$$

Nyní uvažujme v každé oblasti prostý lineární odhad zmíněný v definici 8. Dostáváme

$$\hat{Y} = \sum_{h=1}^{H}\sum_{i:U_i\in s_h} \frac{y_i}{\pi_i}$$

a z tvrzení 1 a 11 víme, že platí $\mathbb{E}\left[\hat{Y}\right] = Y$. Aplikací Yatesovy-Grundyho formule uvedené ve vzorci (3.6) na jednotlivá strata dostaneme vzorec pro rozptyl odhadu úhrnu

$$V_{YG}\left[\hat{Y}\right] = \frac{1}{2}\sum_{h=1}^{H}\left(\sum_{\substack{i:U_i\in\mathbb{U}_h,\, j:U_j\in\mathbb{U}_h\\ i\neq j}}\sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \left(\pi_i\pi_j - \pi_{ij}\right)\right).$$

Obdobnou aplikací vzorce (3.8) obdržíme odhad rozptylu odhadu úhrnu jako

$$\hat{V}_{YG}\left[\hat{Y}\right] = \frac{1}{2}\sum_{h=1}^{H}\left(\sum_{\substack{i:U_i\in s_h,\, j:U_j\in s_h\\ i\neq j}}\sum \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \left(\frac{\pi_i\pi_j}{\pi_{ij}} - 1\right)\right).$$

Je zřejmé, že když $\hat{V}\left[\hat{Y}_h\right]$ nevychýleně odhaduje $V\left[\hat{Y}_h\right]$, tak $\hat{V}\left[\hat{Y}\right]$ nevychýleně odhaduje $V\left[\hat{Y}\right]$.

Populační průměr na jednotku v $\mathbb{U}_h$ je

$$\bar{Y}_h = \frac{Y_h}{N_h},$$

který může být nevychýleně odhadnut jako

$$\hat{\bar{Y}}_h = \frac{\hat{Y}_h}{N_h}.$$

Je-li potřeba vyjádřit populační průměr vlastnosti na jednotku populace, pak

$$\bar{Y} = \frac{Y}{N}$$

a přirozený odhad je

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N}.$$

Po menších algebraických úpravách můžeme $\bar{Y}$ vyjádřit jako

$$\begin{aligned}\bar{Y} &= \frac{N_1 \cdot \bar{Y}_1 + N_2 \cdot \bar{Y}_2 + N_3 \cdot \bar{Y}_3 + \cdots + N_H \cdot \bar{Y}_H}{N} \\ &= \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2 + \frac{N_3}{N}\bar{Y}_3 + \cdots + \frac{N_H}{N}\bar{Y}_H \\ &= W_1 \cdot \bar{Y}_1 + W_2 \cdot \bar{Y}_2 + W_3 \cdot \bar{Y}_3 + \cdots + W_H \cdot \bar{Y}_H,\end{aligned}$$

kde $W_h = \frac{N_h}{N}$ obvykle nazýváme váha strata pro stratum h. Každá váha strata W_h je poměr jednotek $\mathbb{U}_h$, které jsou ve stratu h, vůči všem jednotkám souboru. Tedy $0 \leq W_h \leq 1$ a $\sum_{h=1}^{H} W_h = 1$. Jako alternativní vyjádření $\hat{\bar{Y}}$ používáme

$$\hat{\bar{Y}} = W_1\hat{\bar{Y}}_1 + W_2\hat{\bar{Y}}_2 + W_3\hat{\bar{Y}}_3 + \cdots + W_H\hat{\bar{Y}}_H = \sum_{h=1}^{H} W_h\hat{\bar{Y}}_h.$$

Podobně můžeme vyjádřit alternativně $V\left[\hat{\bar{Y}}\right]$ rovností

$$V\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} W_h^2\, V\left[\hat{\bar{Y}}_h\right],$$

což můžeme odhadnout jako

$$\hat{V}\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} W_h^2\, \hat{V}\left[\hat{\bar{Y}}_h\right].$$

4.5.1 Stratifikovaný náhodný výběr

Výběrový plán je v tomto případě takový, že v každém jednotlivém stratu provádíme prostý náhodný výběr. Jedná se o speciální případ obecného stratifikovaného výběru, kde mohou být v každém stratu prováděny různé plány. Pravděpodobnost zahrnutí prvku $U_i \in \mathbb{U}_h$ je

$$\pi_i = \frac{n_h}{N_h}.$$

Obecně platí, že stratifikovaný náhodný výběr nepatří mezi výběrové plány se stejnými pravděpodobnostmi.

Dříve jsme definovali pojmy výběrový průměr a populační průměr. Nechť s_h značí vzorek n_h prvků vybraných z $\mathbb{U}_h$. Pro účely této podkapitoly budeme symbolem $\bar{y}_h$ označovat výběrový průměr v h-tém stratu počítaný jako

$$\bar{y}_h = \frac{1}{n_h} \sum_{i:U_i \in s_h} y_i.$$

a symbolem $\bar{Y}_h$ populační průměr v h-tém stratu daný vzorcem

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i:U_i \in \mathbb{U}_h} y_i.$$

Analogicky s úvodem čtvrté kapitoly, nechť máme

$$\sigma^2_{y,h} = \frac{1}{N_h - 1} \sum_{i:U_i \in \mathbb{U}_h} \left(y_i - \bar{Y}_h\right)^2.$$

Nevychýlený odhad $\sigma^2_{y,h}$ dle prostého náhodného výběru bez opakování je

$$\hat{\sigma}^2_{y,h} = \frac{1}{n_h - 1} \sum_{i:U_i \in s_h} \left(y_i - \bar{y}_h\right)^2.$$

Vzorce pro odhad úhrnu, rozptyl odhadu úhrnu a jeho odhad jsou

$$\hat{Y} = \sum_{h=1}^{H} N_h \bar{Y}_h,$$

$$V\left[\hat{Y}\right] = \sum_{h=1}^{H} \frac{N_h - n_h}{n_h/N_h} \sigma^2_{y,h},$$

$$\hat{V}\left[\hat{Y}\right] = \sum_{h=1}^{H} \frac{N_h - n_h}{n_h/N_h} \hat{\sigma}^2_{y,h}.$$

Kapitola 5

Výběrové plány pro spojité populace

V následujících odstavcích se budeme věnovat situacím, kdy není možné danou populaci rozdělit na diskrétní jednotky. Při psaní této kapitoly byly inspirací některé odstavce z [2] a dále pak hlavním zdrojem informací publikace [4], a to zejména čtvrtá kapitola. Uvažujme spojitou populaci prvků, které značíme U_x, kde $x \in \mathcal{D} \subseteq \mathbb{R}^d$. Většina metod bude analogických k diskrétním situacím s tím rozdílem, že místo součtů budeme používat integrály. Zároveň zde nepracujeme s pravděpodobnostmi výběru, ale s hustotami pravděpodobnosti.

5.1 Metoda Monte Carlo

Základní motivací je výpočet úhrnu vlastnosti na oblasti $\mathcal{D}$ definované pomocí funkce $y(x)$. Cílený parametr však nebývá úhrn vlastnosti přes celou populaci z definice 4, ale mnohem častěji je předmětem našeho zájmu populační průměr daný definicí 5. Známe-li analytické vyjádření funkce $y(x)$, úhrn Y můžeme spočítat. V případě, že toto vyjádření nemáme k dispozici nebo není možné, užijeme metodu Monte Carlo k získání odhadu $\hat{Y}$, případně $\hat{\bar{Y}}$ z měření hodnot $y(x)$ v náhodně vybraných bodech populace $\mathbb{U}$. Metoda Monte Carlo je nejjednodušší strategie na odhadování cíleného parametru, který je vyjádřen integrálem. Protože spojité populace zahrnují nekonečně mnoho bodů, výběrové pravděpodobnosti definované pro diskrétní populace nemohou být vyjádřeny spolu s výběrovým plánem. Na jejich místě se používají pravděpodobnostní funkce hustoty.

Pro výběr bodů x, ve kterých měříme $y(x)$, definujeme pravděpodobnostní hustotu $g(x)$, která je kladná na oblasti $\mathcal{D}$. V metodě Monte Carlo použijeme rovnoměrnou pravděpodobnostní hustotu, tedy

$$g(x) = \frac{1}{|\mathcal{D}|}, \quad x \in \mathcal{D}.$$

Podívejme se na základní odhady úhrnů a jejich rozptyl. V podkapitole 3.2 jsme uvedli rozšíření Horvitzova-Thompsonova odhadu na spojitou populaci. Vygenerujme bod x_i dle hustoty g. Nevychýlený odhad úhrnu Y je podíl hustoty vlastnosti a pravděpodobnostní funkce hustoty v x_i a je dán vzorcem (3.11), tedy

$$\hat{Y}_i = \frac{y(x_i)}{g(x_i)}.$$

Pro rovnoměrnou hustotu použitou v metodě Monte Carlo se $\hat{Y}_i$ zjednoduší na $\hat{Y}_i = |\mathcal{D}|\, y(x_i)$. Kombinovaný odhad $\hat{Y}$ počítaný z výběrů $n > 1$ bodů je dán vzorcem (3.12). Dosazením a úpravou dostaneme kombinovaný odhad $\hat{Y}$ pro metodu Monte Carlo ve tvaru

$$\hat{Y} = \frac{|\mathcal{D}|}{n} \sum_{i=1}^{n} y(x_i).$$

Vzorek s je v tomto případě tvořen n-ticí $(x_1, \ldots, x_n)$ nezávislých náhodných bodů. Rozptyl odhadu $\hat{Y}$ počítáme dle rovnosti (3.13). Tento rozptyl bývá nevychýleně odhadován vztahem (3.14).

Populační průměr lze nevychýleně odhadnout jako

$$\hat{\bar{Y}}_i = \frac{\hat{Y}_i}{|\mathcal{D}|}.$$

Pro metodu Monte Carlo se nám odhad opět zjednoduší na pouhou míru hustoty vlastnosti v bodě x_i, tedy $\hat{\bar{Y}}_i = y(x_i)$. Pro kombinovaný odhad populačního průměru hustoty za použití $n > 1$ výběrů platí

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\bar{Y}}_i$$

a rozptyl odhadu $\hat{\bar{Y}}$, počítaný jako

$$V\left[\hat{\bar{Y}}\right] = \frac{V\left[\hat{Y}\right]}{|\mathcal{D}|^2},$$

lze nevychýleně odhadnout jako

$$\hat{V}\left[\hat{\bar{Y}}\right] = \frac{\hat{V}\left[\hat{Y}\right]}{|\mathcal{D}|^2}.$$

Odhady $\hat{\bar{Y}}$ a $\hat{Y}$ mají nulový rozptyl, pokud hustota vlastnosti stejně tak jako hustota pravděpodobnosti je rovnoměrně rozdělena v oblasti $\mathcal{D}$. V těchto případech

je $\hat{Y}_i = Y$, bez ohledu na to, který bod vybereme, protože

$$\hat{Y}_i = \frac{y(x_i)}{g(x_i)} = \frac{\frac{Y}{|\mathcal{D}|}}{\frac{1}{|\mathcal{D}|}} = Y.$$

V praxi se však s konstantní hustotou setkáme opravdu ojediněle. Bude nám tedy stačit, když se hustota bude blížit konstantě.

5.2 Systematický výběr

Systematický výběr využívá pravidelnou mříž bodů, nejčastěji se volí čtvercová, obdélníková, trojúhelníková nebo šestiúhelníková. Označme ji písmenem Δ, např. d-rozměrná čtvercová mříž bodů je $\Delta = h\mathbb{Z}^d$ pro nějaké $h > 0$. Podobně jako u metody Monte Carlo vybereme rovnoměrně náhodně v oblasti $\mathcal{D}$ bod x_i. Vezmeme všechny body tvaru $x_i + \delta$ pro nějaké $\delta \in \Delta$, které padnou do $\mathcal{D}$. Znamená to, že jsme mříž bodů posunuli tak, aby obsahovala x_i. Body posunuté mříže, které padnou do zkoumané oblasti $\mathcal{D}$, pak tvoří $s = (x_i + \Delta) \cap \mathcal{D}$, kde $x_i + \Delta = \{x : \exists \delta \in \Delta,\, x = x_i + \delta\}$. Počet n takto vybraných bodů je obecně náhodný. Odhad založený na měřeních $y(x_i + \delta)$ je roven

$$\hat{Y}_i = \frac{|\mathcal{D}|}{n} \sum_{x \in s} y(x).$$

Ačkoliv je $\hat{Y}_i$ počítán ze dvou či více měření, tato měření nejsou odvozena od nezávislých výběrů. V důsledku toho nemůže být rozptyl $V\left[\hat{Y}_i\right]$ odhadnut nevychýleně. Pro odhad populačního průměru $\bar{Y}$ dostáváme

$$\hat{\bar{Y}}_i = \frac{1}{n} \sum_{x \in s} y(x).$$

5.3 Stratifikovaný výběr

Metoda popsaná v této podkapitole je analogická se stratifikovaným výběrem pro diskrétní populace a její rozšíření na spojité populace je velmi intuitivní. Podrobnější rozbor nalezneme v [3]. Uvažujme spojitou populaci $\mathbb{U}$, kterou rozdělíme do strat $\mathbb{U}_1, \ldots, \mathbb{U}_H$ typicky podle geometrie oblasti $\mathcal{D}$. Poznamenejme, že podoblasti $\mathcal{D}_1, \ldots, \mathcal{D}_H$ příslušné stratům $\mathbb{U}_1, \ldots, \mathbb{U}_H$ jsou disjunktní, ale zároveň pokrývají celou oblast $\mathcal{D}$. Vzorek vybraný v h-tém stratu budeme značit s_h a jeho velikost n_h. Odhad v h-tém stratu je dán jako

$$\hat{Y}_h = \frac{|\mathcal{D}_h|}{n_h} \sum_{i : x_i \in s_h} y(x_i).$$

Výsledný odhad úhrnu není opět ničím jiným, než součtem jednotlivých odhadů ve stratech

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h.$$

Pro rozptyl a odhad rozptylu můžeme použít vzorce 4.16 a 4.17. I tady tento odhad pro rozptyl je mírně nadhodnocený.

Kapitola 6

Lesnické metody

V následující kapitole představíme základní výběrové metody používané v praxi, zejména v lesnictví. Použité vzorce nalezneme v [7]. Většina z nich je kombinací informací získaných přímo v lese a informací dostupných z jiných zdrojů. Tímto zdrojem může být například letecká fotografie krajiny nebo předchozí šetření krajiny. Uvažujme, že doplňující informace je zpracována v první fázi výběru. Tento výběr má obvykle značnou velikost a je značen s_1. Následuje druhý výběr, který se provádí přímo v lese. Jedná se o podmnožinu vzorku vybraného v první fázi. Tyto informace z terénu jsou získávány buď pomocí jednoúrovňové metody, kde jsou stromy vybírány přímo za účelem zjištění hodnoty vlastnosti zájmu, v takovém případě je vzorek označován symbolem s_2, nebo pomocí dvouúrovňové metody, kde v první úrovni jsou stromy vybrány k zjištění přibližné hodnoty vlastnosti zájmu, a následně v druhé úrovni je výběr proveden za účelem získání přesné hodnoty vlastnosti, v takovém případě jsou vzorky značeny s_2 a s_3.

6.1 Terminologie

V této kapitole se setkáváme se specifickým přístupem k problematice výběru v lesnictví. Z tohoto důvodu je nezbytné zavést některé pojmy, značení a základní vlastnosti. Zalesněnou oblast uvažujme jako podmnožinu F euklidovského prostoru $\mathbb{R}^2$. V praxi, kdy se les vlní, se používá rovnoběžná projekce reálného lesa. Velikost této plochy F je označována jako $\lambda(F)$. V lesnictví se obvykle používají jako jednotky hektary. Uvažujme populaci $\mathbb{U}$ čítající N stromů náležících lesu F. Hodnota vlastnosti našeho zájmu bude, jak je zvykem, značena Y pro úhrn a $\bar{Y}$ pro průměr. Poznamenejme, že na stromy pohlížíme jako na bezrozměrné body v ploše $U_1, \ldots, U_N \in F \subseteq \mathbb{R}^2$, ve kterých je hodnota vlastnosti dána. Sluší se taktéž zmínit, že populace $\mathbb{U}$ nemusí nutně znamenat všechny stromy z daného lesa. V aplikacích z této kapitoly uvažujeme pouze ty prvky, jež splňují určité parametry. Jde o hod-

noty zkoumané charakteristiky, které jsou měřeny v tzv. prsní výšce. V odbornější literatuře se setkáme s pojmem výčetní výška. Pro představu, jedná se o hodnoty ve výšce 130 cm od paty kmene. Do populace $\mathbb{U}$ zahrneme pouze ty stromy, které splňují minimální šířku kmene v této referenční výšce. Pro seznámení se základními lesnickými pojmy je vhodná vysokoškolská učebnice [10].

Jestliže máme stromy očíslovány pomocí $1, \ldots, N$, můžeme použít metody určené pro diskrétní populace (viz kapitola 4) a získat odhady úhrnu Y nebo průměru $\bar{Y}$. V praktických situacích ale tento předpoklad není příliš reálný. Obvykle nemáme ani informaci o počtu stromů v lese, natož abychom byli schopni je očíslovat. Přirozenější je způsob výběru, kdy se náhodně zvolí místo v lese a prozkoumají se všechny stromy v blízkém okolí.

Vybíráme náhodný bod x rovnoměrně z F, což znamená, že pro jakoukoliv měřitelnou množinu $B \subseteq \mathbb{R}^2$ je pravděpodobnost, že daný bod x náleží množině B, rovna

$$P(x \in B) = \frac{\lambda(B \cap F)}{\lambda(F)}.$$

Stromy jsou vybrány, pokud náleží kruhu $K_r(x) = \{y \in \mathbb{R}^2 : d(y, x) \leq r\}$ s poloměrem r a středovým bodem x. Nyní můžeme uvažovat indikátor

$$I_i(x) = \begin{cases} 1, & \text{když } U_i \in K_r(x), \\ 0, & \text{když } U_i \notin K_r(x) \end{cases}$$

a N kruhů $K_i(r) = K_r(U_i)$ s konstantním poloměrem r centrovaných kolem stromů, kde U_i značí polohu i-tého stromu. Ze symetrie je zřejmé, že i-tý strom je v kruhu $K_r(x)$ právě tehdy, když náhodný bod x je v i-tém kruhu $K_i(r)$. Z tohoto důvodu dostáváme zřejmý, ale velmi důležitý princip duality

$$I_i(x) = 1 \Leftrightarrow x \in K_i(r).$$

Pravděpodobnost zahrnutí i-tého stromu je dána jako

$$\pi_i = P(I_i(x) = 1) = \mathbb{E}_x I_i(x) = \frac{\lambda(K_i(r) \cap F)}{\lambda(F)}$$

a pravděpodobnosti zahrnutí pro dvojici stromů jako

$$\pi_{ij} = \frac{\lambda(K_i(r) \cap K_j(r) \cap F)}{\lambda(F)}.$$

Až na výjimky u hranic a krajů lesa je tato pravděpodobnost konstantní.

Horvitzův-Thompsonův odhad populačního průměru má tvar $\frac{1}{N} \sum_{i=1}^{N} \frac{I_i(x) y_i}{\pi_i}$. Pokud dělení počtem N nahradíme velikostí plochy lesa, dostaneme

$$y(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) y_i}{\pi_i}.$$

Této funkci se říká lokální hustota. Všimněme si, že hodnotu $\lambda(F)$ nemusíme znát, protože se zkrátí se jmenovatelem zlomku π_i.

Protože x je zvoleno rovnoměrně náhodně v F, lokální hustota svou konstrukcí splňuje

$$\mathbb{E}_x y(x) = \frac{1}{\lambda(F)} \int_F y(x)\,\mathrm{d}x = \frac{1}{\lambda(F)} \sum_{i=1}^{N} y_i = \bar{Y}. \tag{6.1}$$

Tento vzorec ukazuje vztah mezi populačním průměrem ve spojité populaci a diskrétním populačním průměrem. Přístupy uvažující nekonečnou populaci lépe vyhovují lesnické problematice než ty, které uvažují diskrétní populace. Ve zbytku kapitoly budeme symbolem $\bar{Y}$ rozumět populační průměr $\frac{1}{\lambda(F)} \int_F y(x)\,\mathrm{d}x$.

6.2 Jednofázový jednoúrovňový prostý náhodný výběr

Uvažujme množinu s_2 o velikosti n_2 bodů vybraných rovnoměrně a vzájemně nezávisle v lese F. Námi zkoumaná charakteristika je populační průměr $\bar{Y}$. Jednofázový jednoúrovňový odhad populačního průměru při prostém náhodném výběru je definován jako

$$\hat{\bar{Y}} = \frac{1}{n_2} \sum_{x \in s_2} y(x). \tag{6.2}$$

Tento odhad, jak se můžeme sami snadno přesvědčit, je nevychýlený. Rozptyl lokální hustoty je dán vzorcem

$$V\left[y(x)\right] = \frac{1}{\lambda\left(F\right)} \int_F \left(y(x) - \bar{Y}\right)^2 \mathrm{d}x,$$

popřípadě dle vzorce

$$V\left[y(x)\right] = \frac{1}{\lambda(F)^2} \left\{ \sum_{i=1}^{N} \frac{y_i^2\left(1-\pi_i\right)}{\pi_i} + \sum_{\substack{1 \le i,j \le N \\ i \ne j}} \sum \frac{y_i y_j \left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_i \pi_j} \right\}.$$

Rozptyl odhadu počítáme pomocí vzorce

$$V\left[\hat{\bar{Y}}\right] = \frac{1}{n_2} V\left[y(x)\right]. \tag{6.3}$$

Odhad rozptylu odhadu je dán jako

$$\hat{V}\left[\hat{\bar{Y}}\right] = \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{x \in s_2} \left(y(x) - \hat{\bar{Y}}\right)^2. \tag{6.4}$$

6.3 Jednofázový jednoúrovňový skupinkový výběr

Ve čtvrté kapitole říkáme, že námi šetřená oblast se může skládat ze sjednocení M subpopulací. V této podkapitole bude třeba teorii rozšířit. Předpokládejme tedy, že skupinka skládající se z M částí je dána pevnou množinou M vektorů $e_l \in \mathbb{R}^2$, $l = 1, 2, \ldots, M$. Nechť dále pro jakoukoliv borelovskou množinu $A \subseteq \mathbb{R}^2$ značíme symbolem A_l množinu $A + e_l = \{x : \exists a \in A, x = a + e_l\}$. Klíčovým předpokladem je, že množina A je dostatečně veliká, aby splňovala $F \subseteq A_l$ pro všechna $l = 1, 2, \ldots, M$.

Nyní vybíráme náhodně bod x rovnoměrně z množiny A. Je zřejmé, že body $x_l = x + e_l$ jsou rovnoměrně rozděleny v A_l. Poznamenejme, že ty body x_l, které padnou do F, jsou rovnoměrně rozděleny v lese F.

Indikátor množiny F definujeme jako

$$I_F(x) = \begin{cases} 1, & \text{když } x \in F, \\ 0, & \text{když } x \notin F. \end{cases}$$

Počet bodů skupinky spadajících do plochy lesa je náhodná veličina dána vzorcem

$$M(x) = \sum_{l=1}^{M} I_F(x_l).$$

Lokální hustota na úrovni skupinky je tvaru

$$y_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l) y(x_l)}{M(x)}.$$

Může se stát, že $M(x) = 0$, v tom případě dostáváme výraz typu $\frac{0}{0}$ a definujeme ho jako 0.

Obdobně jako v předchozí podkapitole uvažujme množinu s_2 o velikosti n_2 bodů vybraných rovnoměrně a vzájemně nezávisle v A. Jednofázový jednoúrovňový odhad průměru hustoty pro skupinkový výběr je definován jako

$$\hat{\bar{Y}}_c = \frac{\sum_{x \in s_2} M(x) y_c(x)}{\sum_{x \in s_2} M(x)}.$$

Pro úplnost si představme vzorec pro výpočet odhadu rozptylu odhadu

$$\hat{V}\left[\hat{\bar{Y}}_c\right] = \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{x \in s_2} \left(\frac{M(x)}{\bar{M}_2} \right)^2 \left(y_c(x) - \hat{\bar{Y}}_c \right)^2, \tag{6.5}$$

kde $\bar{M}_2 = \frac{1}{n_2} \sum_{x \in s_2} M(x)$ je průměrný počet bodů skupinky spadající do lesa F.

6.4 Jednofázový dvouúrovňový prostý náhodný výběr

V mnoha situacích se setkáváme s problémem vysokých nákladů na zisk informace měřené vlastnosti. Formalizujme si tedy postup řešící tento problém. Pro každý bod $x \in s_2$ jsou stromy vybírány s pravděpodobnostmi π_i. Množina vybraných stromů je značena $s_2(x)$. Z každého vybraného stromu $U_i \in s_2(x)$ získáme aproximaci y_i^* přesné hodnoty y_i. Z konečné množiny $s_2(x)$ vybereme podvzorek stromů $s_3(x) \subseteq s_2(x)$. V definici 7 jsme definovali indikátor zahrnutí. Pro lepší orientaci definujeme prvoúrovňový indikátor jako

$$I_i(x) = \begin{cases} 1, & \text{když } U_i \in s_2(x), \\ 0, & \text{když } U_i \notin s_2(x). \end{cases}$$

Pro každý strom $U_i \in s_3(x)$ měříme přesnou hodnotu y_i. Druhoúrovňový indikátor definujeme jako

$$J_i(x) = \begin{cases} 1, & \text{když } U_i \in s_3(x), \\ 0, & \text{když } U_i \notin s_3(x). \end{cases}$$

Z definice a konstrukce je zřejmé, že $I_i(x)J_i(x) = J_i(x)$. Pro konstrukci dobrého bodového odhadu potřebujeme znát rezidua $R_i = y_i - y_i^*$, která jsou známa jen pro stromy $U_i \in s_3(x)$. Dále definujme pravděpodobnosti p_i jakožto pravděpodobnost zahrnutí stromu U_i do výběru $s_3(x)$ za podmínky, že byl daný strom U_i zahrnutý do výběru $s_2(x)$, tedy $p_i = P(J_i(x) = 1 | I_i(x) = 1)$. Zobecněná lokální hustota $y^*(x)$ je definována vzorcem

$$\begin{aligned} y^*(x) &= \frac{1}{\lambda(F)} \left(\sum_{i=1}^{N} \frac{I_i(x) y_i^*}{\pi_i} + \sum_{i=1}^{N} \frac{I_i(x) J_i(x) R_i}{\pi_i p_i} \right) \\ &= \frac{1}{\lambda(F)} \left(\sum_{i:U_i \in s_2(x)} \frac{y_i^*}{\pi_i} + \sum_{i:U_i \in s_3(x)} \frac{R_i}{\pi_i p_i} \right). \end{aligned}$$

Zde předpokládáme, že predikce y_i^* hodnot y_i je dána externím modelem. To znamená, že odpovídající model není součástí dat posbíraných v šetření.

Pro další pochopení vzorců je třeba zavést následující značení. Symbolem $\mathsf{V}(x)$ budeme značit podmíněný rozptyl $y^*(x)$ při výběru vzorku $s_3(x)$ za podmínky výběru vzorku $s_2(x)$. Vzhledem k tomu, že stromy ve druhé úrovni jsou vybírány nezávisle na ostatních, dostáváme

$$\mathsf{V}(x) = V_{3|2}\left(y^*(x)\right) = V_{3|2}\left(\frac{1}{\lambda(F)} \sum_{i:U_i \in s_3(x)} \frac{R_i}{\pi_i p_i} \right) = \frac{1}{\lambda(F)^2} \sum_{i:U_i \in s_3(x)} \frac{R_i^2 (1 - p_i)}{\pi_i^2 p_i},$$

což využijeme i v následujících odstavcích. Jednofázový dvouúrovňový odhad populačního průměru pro prostý náhodný výběr je definován jako

$$\hat{\bar{Y}}^* = \frac{1}{n_2} \sum_{x \in s_2} y^*(x).$$

Jeho rozptyl je počítán dle vzorce

$$V\left[\hat{\bar{Y}}^*\right] = \frac{1}{n_2 \lambda(F)^2} \sum_{i=1}^{N} \frac{y_i^2 (1-\pi_i)}{\pi} + \sum_{\substack{1 \le i,j \le N \\ i \neq j}} \sum \frac{y_i y_j \left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_i \pi_j}$$
$$+ \frac{1}{n_2 \lambda(F)^2} \left(\sum_{i=1}^{N} \frac{R_i^2}{\pi_i p_i} - \sum_{i=1}^{N} \frac{R_i^2}{\pi_i} \right).$$

Pro úplnost ještě uveďme vzorec pro výpočet odhadu rozptylu odhadu

$$\hat{V}\left[\hat{\bar{Y}}^*\right] = \frac{1}{n_2 \left(n_2 - 1\right)} \sum_{x \in s_2} \left(y^*(x) - \hat{\bar{Y}}^* \right)^2.$$

6.5 Jednofázový dvouúrovňový skupinkový výběr

V této podkapitole se setkáme s množinou A definovanou v podkapitole 6.3. Předpokládejme, že stromy na druhé úrovni jsou vybírány nezávisle na ostatních v každém bodě v každé skupince. Můžeme $y_c(x)$ velmi intuitivně zobecnit na

$$y_c^*(x) = \frac{\sum_{l=1}^{M} I_F(x_l)\, y^*(x_l)}{M(x)},$$

kde $y^*(x_l)$ je zobecněná lokální hustota v bodě x_l. Jednofázový dvouúrovňový bodový odhad pro skupinkový výběr je definován analogicky k $\hat{\bar{Y}}_c$, a to jako

$$\hat{\bar{Y}}_c^* = \frac{\sum_{x \in s_2} M(x)\, y_c^*(x)}{\sum_{x \in s_2} M(x)}.$$

Rozptyl tohoto odhadu je dán vzorcem

$$V\left[\hat{\bar{Y}}_c^*\right] = \frac{1}{n_2} \frac{\mathbb{E}_{x \in A} M(x)^2 \left(y_c(x) - \bar{Y}\right)^2}{\left(\mathbb{E}_{x \in A} M(x)\right)^2} + \frac{\mathbb{E}_{x \in F} \mathsf{V}(x)}{n_2 \mathbb{E}_{x \in A} M(x)},$$

kde $\mathbb{E}_{x \in A} M(x) = \sum_{l=1}^{M} P\left(x_l \in F\right)$. Poznamenejme, že symboly $\mathbb{E}_{x \in F}$ a $\mathbb{E}_{x \in A}$ rozumíme střední hodnotu vzhledem k rovnoměrnému rozdělení na F, respektive na A. Jako u každé metody odhadu si ještě na závěr uveďme vzorec pro výpočet odhadu rozptylu odhadu. Ten je dán následovně:

$$\hat{V}\left[\hat{\bar{Y}}_c^*\right] = \frac{1}{n_2 \left(n_2 - 1\right)} \sum_{x \in s_2} \left(\frac{M(x)}{\bar{M}_2} \right)^2 \left(y_c^*(x) - \hat{\bar{Y}}_c^* \right)^2.$$

6.6 Dvoufázový jednoúrovňový prostý náhodný výběr

Při volbě tohoto postupu nejprve vybíráme v první fázi velký vzorek s_1 o velikosti n_1 bodů, které jsou nezávislé a rovnoměrně rozdělené v lese F. V každém z těchto bodů získáme doplňující informaci, často se jedná o informaci kvalitativního charakteru. Takovým příkladem může být například informace získaná z leteckého nebo satelitního snímku. Ve druhé fázi vybíráme malý vzorek $s_2 \subseteq s_1$ o velikosti n_2 bodů z prvního většího vzorku s_1. V každém bodě $x \in s_2$ získáme ze šetření v terénu lokální hustotu $y(x)$. Jak je zřejmé, pro body $x \in s_1 \setminus s_2$, tedy pro ty náležící velkému vzorku, ale nenáležící malému, máme k dispozici pouze doplňující informaci. I přesto můžeme vytvořit predikci $\hat{y}(x)$ skutečné lokální hustoty $y(x)$. Předpokládejme, že tato predikce je daná externím modelem a není obsažena v datech získaných v aktuálním šetření. Definujme residua $R(x)$ jako rozdíl skutečné lokální hustoty a její predikce, tedy $R(x) = y(x) - \hat{y}(x)$. Nyní můžeme definovat dvoufázový jednoúrovňový odhad pro prostý náhodný výběr jako

$$\hat{\bar{Y}}_{reg} = \frac{1}{n_1} \sum_{x \in s_1} \hat{y}(x) + \frac{1}{n_2} \sum_{x \in s_2} R(x).$$

Dolní index reg značí, že odhad využívá regresního modelu. Poznamenejme, že nepředpokládáme, že průměr residuí je roven nule. Zbývá uvést vzorec pro výpočet rozptylu odhadu

$$V\left[\hat{\bar{Y}}_{reg}\right] = \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \frac{1}{\lambda(F)} \int_F \left(R(x) - \bar{R}\right)^2 \mathrm{d}x + \frac{1}{n_1} \frac{1}{\lambda(F)} \int_F \left(y(x) - \bar{Y}\right)^2 \mathrm{d}x,$$

kde

$$\bar{R} = \frac{1}{\lambda(F)} \int_F R(x)\, \mathrm{d}x.$$

Vzorec pro odhad rozptylu odhadu je

$$\hat{V}\left[\hat{\bar{Y}}_{reg}\right] = \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{x \in s_2} \left(R(x) - \bar{R}_2\right)^2 + \frac{1}{n_1} \frac{1}{n_2 - 1} \sum_{x \in s_2} \left(y(x) - \bar{Y}_2\right)^2, \tag{6.6}$$

kde $\bar{R}_2 = \frac{1}{n_2} \sum_{x \in s_2} R(x)$ a $\bar{Y}_2 = \frac{1}{n_2} \sum_{x \in s_2} y(x)$.

V některých situacích, kdy je doplňující informace získávána z map, leteckých nebo satelitních snímků, můžeme považovat n_1 za konvergující k nekonečnu. V takovém případě se nám předchozí odhad daný vzorcem (6.6) zjednoduší na

$$\hat{V}\left[\hat{\bar{Y}}_{reg}\right] = \frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\left(R(x)-\bar{R}_2\right)^2.$$

6.7 Dvoufázový dvouúrovňový prostý náhodný výběr

Při tomto postupu dojde na všechny tři výběry zmíněné v šesté kapitole. Tedy vybíráme vzorky s_1, s_2 a s_3. Koncept je naprosto totožný jako v předchozím dvoufázovém jednoúrovňovém prostém náhodném výběru s tou výjimkou, že neznámou skutečnou lokální hustotu $y(x)$ v bodě x nahradíme jejím odhadem, tedy zobecněnou lokální hustotou $y^*(x)$. Dvouúrovňový dvoufázový odhad pro prostý náhodný výběr je definován jako

$$\hat{\bar{Y}}^*_{reg} = \frac{1}{n_1}\sum_{x\in s_1}\hat{y}(x) + \frac{1}{n_2}\sum_{x\in s_2}R^*(x),$$

kde $R^*(x) = y^*(x) - \hat{y}(x)$. Při počítání rozptylu odhadu užíváme vzorce

$$V\left[\hat{\bar{Y}}^*_{reg}\right] = \frac{1}{n_1}V[y(x)] + \left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2}V[R(x)] + \frac{1}{n_2}\mathbb{E}_x\mathsf{V}(x)$$

a pro počítání jeho odhadu používáme vzorce

$$\hat{V}\left[\hat{\bar{Y}}^*_{reg}\right] = \left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\left(R^*(x)-\bar{R}^*_2\right)^2 + \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}\left(y^*(x)-\bar{Y}^*_2\right)^2,$$

kde $\bar{R}^*_2 = \frac{1}{n_2}\sum_{x\in s_2}R^*(x)$ a $\bar{Y}^*_2 = \frac{1}{n_2}\sum_{x\in s_2}y^*(x)$.

6.8 Dvoufázový jednoúrovňový skupinkový výběr

Tato metoda je přímým zobecněním předchozích odstavců. V první fázi vybíráme velký vzorek o velikosti n_1 skupinek, jejichž body umístění $x \in s_1$ jsou rovnoměrně a nezávisle rozděleny v $A \supseteq F$. Ve druhé fázi vybíráme podvzorek o velikosti n_2 skupinek z n_1 původních s umístěním $x \in s_2 \subseteq s_1$. Pro jakoukoliv danou skupinku s bodem umístění $x \in s_2$ máme hodnotu lokální hustoty $y(x_l)$ v každém bodě skupinky. Dvoufázový jednoúrovňový odhad pro skupinkový výběr je počítán jako

$$\hat{\bar{Y}}_{c,reg} = \frac{\sum_{x\in s_1}M(x)\hat{y}_c(x)}{\sum_{x\in s_1}M(x)} + \frac{\sum_{x\in s_2}M(x)R_c(x)}{\sum_{x\in s_2}M(x)},$$

kde rezidua na úrovni skupinek jsou definována dle

$$R_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l)\left(y(x_l) - \hat{y}(x_l)\right)}{\sum_{l=1}^{M} I_F(x_l)}.$$

Vzorec pro rozptyl je definován vzorcem

$$V\left[\hat{\bar{Y}}_{c,reg}\right] = \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{\mathbb{E}_{x\in A}M(x)^2\left(R_c(x) - \bar{R}_c\right)^2}{\left(\mathbb{E}_{x\in A}M(x)\right)^2} + \frac{1}{n_1}\frac{\mathbb{E}_{x\in A}M(x)^2\left(y_c(x) - \bar{Y}_c\right)^2}{\left(\mathbb{E}_{x\in A}M(x)\right)^2}.$$

Tento rozptyl je odhadován jako

$$\hat{V}\left[\hat{\bar{Y}}_{c,reg}\right] = \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{1}{(n_2-1)}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(R_c(x) - \hat{R}_2\right)^2 + \frac{1}{n_1}\frac{1}{(n_2-1)}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(y_c(x) - \hat{Y}_2\right)^2,$$

kde $\hat{Y}_2 = \frac{1}{n_2}\sum_{x\in s_2} y_c(x)$ a $\hat{R}_2 = \frac{1}{n_2}\sum_{x\in s_2} R_c(x)$.

6.9 Dvoufázový dvouúrovňový skupinkový výběr

Tato metoda je přímou kombinací předchozích dvou metod. V první fázi dostaneme predikci $\hat{y}(x_l)$ a $\hat{y}_c(x)$. Ve druhé fázi dostáváme zobecněné hustoty $y^*(x_l)$ a $y_c^*(x)$. Dvoufázový dvouúrovňový odhad pro skupinkový náhodný výběr je vyjádřen jako

$$\hat{\bar{Y}}^*_{c,reg} = \frac{\sum_{x\in s_1} M(x)\hat{y}_c(x)}{\sum_{x\in s_1} M(x)} + \frac{\sum_{x\in s_2} M(x)R_c^*(x)}{\sum_{x\in s_2} M(x)},$$

kde zobecněná rezidua na úrovni skupinek jsou tvaru

$$R_c^*(x) = \frac{\sum_{l=1}^{M} I_F(x_l)\left(y^*(x_l) - \hat{y}(x_l)\right)}{\sum_{l=1}^{M} I_F(x_l)}.$$

Rozptyl odhadu úhrnu počítáme jako

$$V\left[\hat{\bar{Y}}^*_{c,reg}\right] = \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{\mathbb{E}_{x\in A}M(x)^2\left(R_c(x) - \bar{R}\right)^2}{\left(\mathbb{E}_{x\in A}M(x)\right)^2} + \frac{1}{n_1}\frac{\mathbb{E}_{x\in A}M(x)^2\left(y_c(x) - \bar{Y}\right)^2}{\left(\mathbb{E}_{x\in A}M(x)\right)^2} + \frac{1}{n_2}\frac{\mathbb{E}_{x\in F}\mathsf{V}(x)}{\mathbb{E}_{x\in A}M(x)}.$$

Odhad tohoto rozptylu počítáme vzorcem

$$\hat{V}\left[\hat{\bar{Y}}^*_{c,reg}\right] = \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{1}{n_2 - 1}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(R^*_c(x) - \hat{R}^*_2\right)^2 + \frac{1}{n_1}\frac{1}{n_2 - 1}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(y^*_c(x) - \hat{Y}^*_2\right)^2,$$

kde $\hat{Y}^*_2 = \frac{1}{n_2}\sum_{x\in s_2} y^*_c(x)$ a $\hat{R}^*_2 = \frac{1}{n_2}\sum_{x\in s_2} R^*_c(x)$.

6.10 Národní inventarizace lesů

Následující odstavce popíší metodu použitou u venkovního sběru dat při Národní inventarizaci lesů. Národní inventarizace lesů, zkráceně NIL, je prováděna na celém území České republiky a má za úkol podat přesné a souhrnné informace o stavu lesů. Šetření probíhá v náhodně vybraných kruhových inventarizačních plochách. Popis metodiky a výsledky prvního cyklu NIL lze nalézt v [6].

Řekněme si několik málo vět ke způsobu, jakým jsou vybírány lokality. Výběr probíhá dle následujícího postupu. Nejprve je vytvořena pomyslná čtvercová síť bodů Δ ukotvená v náhodně vygenerovaném bodě. Vzdálenost mezi sousedními body je l_1. Kolem každého bodu sítě vybíráme střed x_{i1} první inventarizační plochy. Ten je vybrán ve vzdálenosti r_i, kde r_i je rovnoměrně náhodně vybrána z intervalu $(0, l_2)$, pod rovnoměrně náhodně zvoleným úhlem α_{i1} z intervalu $(0°, 360°)$. Střed x_{i2} druhé inventarizační plochy je ve vzdálenosti l_2 od středu první inventarizační plochy x_{i1} pod náhodně zvoleným úhlem α_{i2} z intervalu $(0°, 360°)$. Obě inventarizační plochy mají stejný poloměr l_3. Šetření plochy probíhá pouze v případě, kdy střed leží uvnitř lesa.

Pro úplnou představu o Národní inventarizaci lesů zbývá uvést přesné hodnoty vzdáleností l_1, l_2 a l_3. Vzdálenost l_1 mezi body mřížky je 2 km. Hodnota maximální vzdálenosti l_2 inventarizačního středu od bodu mřížky je 300 metrů. Posledním údajem je velikost poloměru l_3 kruhové inventarizační plochy, která je 12,62 metrů. Plocha každé inventarizační plochy je tedy 500 metrů čtverečních.

Použitý výběrový plán vychází ze systematického výběru, který byl popsán v podkapitole 5.2. Rozdíl je v tom, že nejsou použity přímo vrcholy sítě, ale ty jsou náhodně posunuty. Vybrané body se navíc vyskytují ve skupinkách po dvou.

Označíme-li s_2 množinu těch inventarizačních středů $\{x_{i1}\}$ a $\{x_{i2}\}$, které padnou do lesa F, a n_2 počet bodů tohoto vzorku, pak nevychýlený odhad populačního průměru je

$$\hat{\bar{Y}} = \frac{1}{n_2}\sum_{x\in s_2} y(x).$$

Podobně jako u systematického výběru neexistuje vhodný postup, který by dával nevychýlený odhad rozptylu odhadu.

6.11 Znáhodněný stratifikovaný výběr

Uvažujme obdélníkovou síť bodů $\Delta = \{(ih_1, jh_2) \in \mathbb{R}^2 : i,\, j \in \mathbb{Z}\}$ pro dané $h_1,\, h_2 > 0$. Nechť $z = (z_1,\, z_2)$ je rovnoměrně náhodně vygenerovaný bod v obdélníku $R = [0,\, h_1] \times [0,\, h_2]$. Očíslujeme ty obdélníky $z + R = [z_1,\, z_1 + h_1] \times [z_2,\, z_2 + h_2]$, které protnou F, jako $R_1, \ldots, R_n$. V každém z těchto obdélníků je rovnoměrně náhodně zvolený jeden bod. Z těchto bodů uvažujeme ty, které padnou do lesa F, a vytvoříme z nich vzorek s_2 o velikosti n_2.

Odhad populačního průměru opět počítáme jako

$$\hat{\bar{Y}} = \frac{1}{n_2} \sum_{x \in s_2} y(x).$$

Obdélníky $R_1, \ldots, R_n$ můžeme chápat jako strata, v každém stratu je vybrán jeden bod. Vzhledem k náhodnému posunu o vektor z mluvíme o znáhodněném stratifikovaném výběru, který zobecňuje klasický stratifikovaný výběr z podkapitoly 5.3. Tato metoda byla plánována pro výběr inventarizačních ploch druhého cyklu NIL. Pro více podrobností o tomto výběrovém plánu odkazujeme čtenáře na [1] nebo [9].

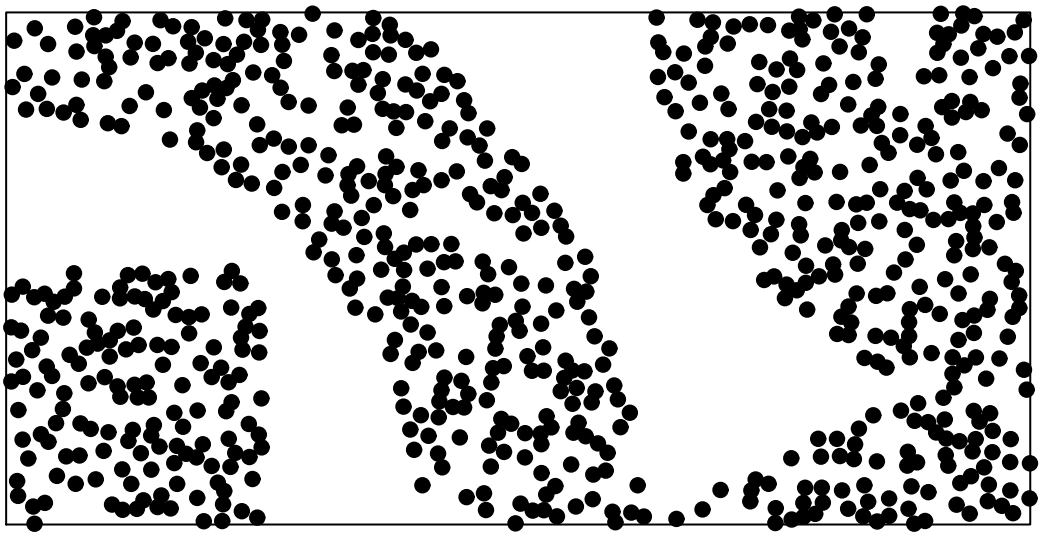
Kapitola 7

Aplikace

V této kapitole se budeme věnovat zkoumání některých metod uvedených v předchozím textu. Pro tyto účely budeme používat jeden konkrétní les. Přesněji řečeno, jedná se o výřez z krajiny obsahující i nezalesněné plochy. Na nich si můžeme představit louku, pole nebo říční tok. Naše oblast je počítačově nasimulována, avšak pro analýzu výběrových metod je stejně hodnotná jako jakákoliv reálná krajina. Pro studii a simulaci metod byl použit statistický software R. Podrobnější instrukce nalezneme v manuálu [8]. Tento software je volně dostupný na stránkách http://www.r-project.org. Velmi dobrý návod pro instalaci a základní manipulaci nalezneme například na stránkách doc. RNDr. Arnošta Komárka, Ph.D.– http://www.karlin.mff.cuni.cz/~komarek/Rko/Rmanual1.pdf. Základní balík neobsahuje rozšíření pro práci s výběrovými metodami a analýzou bodových procesů v prostoru. Je proto nezbytné doinstalovat přídavné balíky. Jde o knihovny spatstat, spsurvey, sampling a RandomFields.

Naše zalesněná oblast F je zasazena do pozorovacího okna o rozměrech 200×100 metrů. V tomto okně máme čtyři zalesněné plochy. V levém dolním rohu je čtvercová plocha o rozměrech 50×50 metrů. Kolem ní máme pás čtvrtiny mezikruží se středem v levém dolním rohu, kde vnější hranice má poloměr 125 metrů a vnitřní hranice má poloměr 80 metrů. V pravé horní části se nachází čtvrtkruh se středem v pravém horním rohu a o poloměru 75 metrů. Poslední, čtvrtou, částí je lichoběžník umístěný do pravé dolní části.

Výše zmíněné oblasti nám říkají, kde v našem pozorovacím okně o celkové rozloze 2 hektary máme rozmístěné stromy. Zalesněná plocha má obsah přibližně 1,42 hektaru. Rozmístění stromů je určeno Matérnovým bodovým procesem typu II s pevným jádrem 2 metry. Definici a základní vlastnosti tohoto modelu teorie bodových procesů lze nalézt v [5, Section 6.5.2]. Každému možnému stromu je přiřazena vlastnost, kterou budeme posléze zkoumat. Onou vlastností je výčetní tloušťka. Jednotlivé tloušťky kmenů jsou určeny z realizace stacionárního logaritmicko–gaussovského náhodného pole, které je nezávislé na polohách stromů. Na obrázku



Obrázek 7.1: Pozorovací okno se zalesněnou částí, polohy stromů jsou vyznačeny puntíky.

7.1 je pro lepší představu znázorněn náš les.

Pro diskrétní přístupy bude naším cílem odhad součtu výčetních tlouštěk jednotlivých stromů v lese. V textu se budeme setkávat s pojmem odhad úhrnu. Oblast, kterou budeme vyšetřovat, čítá 768 stromů. Součet průměrů kmenů stromů je roven 256,8 metrů. Kromě odhadu nás bude zajímat také rozptyl odhadu úhrnu a odhad tohoto rozptylu. Pro každou metodu a volbu parametrů provedeme 100 000 simulací výběru vzorku s. Spočtené odhady zprůměrujeme přes všechny simulace a výsledné hodnoty budeme zapisovat do tabulky, kterou nalezneme u každé popsané metody v této kapitole. Úhrn dané vlastnosti nalezneme ve sloupci s hlavičkou Y. Odhad toho úhrnu bude značen $\hat{Y}$. Hodnoty v těchto sloupcích budou v jednotkách metrů. Sloupec vyjadřující rozptyl odhadu vlastnosti popisujeme jako $V\left[\hat{Y}\right]$. Poslední sloupec bude nadepsán $\hat{V}\left[\hat{Y}\right]$ a budeme v něm mít průměrné hodnoty odhadu rozptylu odhadu sledované charakteristiky. V těchto dvou sloupcích budou hodnoty v metrech čtverečních. Dále bude u každé metody uvedena tabulka popisující kvalitu odhadů. V každé tabulce budou čtyři sloupce. První bude označovat počet vyšetřených stromů. Druhý bude zobrazovat průměrnou hodnotu absolutní odchylky odhadu úhrnu od reálného úhrnu a bude značen $MAE\left[\hat{Y}:Y\right]$. Ve třetím budeme mít analogickou hodnotu pro rozptyl odhadu, tedy absolutní odchylku odhadu rozptylu odhadu úhrnu od rozptylu odhadu úhrnu se značením $MAE\left[\hat{V}:V\right]$. V posledním sloupci budeme mít relativní odchylku odhadu rozptylu v procentech, danou a počítanou jako $RMAE\left[\hat{V}:V\right]=\frac{MAE\left[\hat{V}:V\right]}{V}$. Alternativně bychom pro mí-

ru kvality odhadů mohli místo absolutní chyby použít čtvercovou chybu. V případě nestranných odhadů je ale $MSE\left[\hat{Y}:Y\right] = V\left[\hat{Y}\right]$, takže tuto hodnotu nalezneme rovnou ve sloupci $V\left[\hat{Y}\right]$.

V případě spojitých přístupů bude konvence analogická. Odhadovat budeme populační průměr $\bar{Y} = Y/\lambda(F)$, jehož hodnota je přibližně 18,17 m^{-1} a nalezneme ji ve sloupci s hlavičkou $\bar{Y}$. Odhad tohoto populačního průměru bude značen $\hat{\bar{Y}}$. U některých metod nemáme teoretické vyjádření rozptylu odhadu průměru $V\left[\hat{\bar{Y}}\right]$, v takovém případě použijeme jako jeho aproximaci výběrový rozptyl získaný z provedených simulací, budeme ho značit $\tilde{V}\left[\hat{\bar{Y}}\right]$. Další sloupec bude nadepsán $\hat{V}\left[\hat{\bar{Y}}\right]$ a budeme v něm mít průměrné hodnoty odhadu rozptylu odhadu sledované charakteristiky. Stejně tak, jak je tomu pro diskrétní metody, i zde bude zobrazena průměrná absolutní odchylka $MAE\left[\hat{\bar{Y}}:\bar{Y}\right]$ populačního průměru od jeho odhadu. Kompletní přehled doplní informace o poloměru vyšetřované plochy spolu s průměrným počtem vyšetřených stromů.

7.1 Prostý náhodný výběr

Pro první analýzu použijeme nejjednodušší metodu, kterou je prostý náhodný výběr uvedený v podkapitole 4.1.1. Uvažujme tedy náš les uvedený výše. Jak jsme uvedli dříve, odhad úhrnu se počítá dle rovnosti (4.1), kde $N = 768$, n odpovídá velikosti výběru a y_i značí výčetní tloušťky jednotlivých stromů. Provedeme náhodnou generaci určitého počtu stromů. V těchto stromech zjistíme požadovanou vlastnost. Tuto generaci provedeme 100 000-krát, a poté dané odhady zprůměrujeme. Dostáváme výsledky, které jsou shrnuty v tabulce 7.1.

n	Y	$\hat{Y}$	$V\left[\hat{Y}\right]$	$\hat{V}\left[\hat{Y}\right]$
5	256,80	256,86	272,27	271,75
10	256,80	256,82	135,24	135,17
15	256,80	256,85	89,57	89,41
20	256,80	256,82	66,73	66,64
25	256,80	256,79	53,03	52,95
50	256,80	256,78	25,62	25,59
100	256,80	256,78	11,92	11,91

Tabulka 7.1: Průměr odhadů přes 100 000 simulací pro prostý náhodný výběr.

Použili jsme výběry o velikostech 5, 10, 15, 20, 25, 50 a 100 stromů, což vidíme v prvním sloupci. Druhý sloupec udává hodnotu populačního úhrnu. Ve třetím

sloupci je průměr vypočtených Horvitzových-Thompsonových odhadů úhrnu, které máme definované vzorcem (4.1). Další sloupec udává hodnotu rozptylu odhadu úhrnu počítanou dle vztahu (4.2), ta závisí na rozsahu výběru n. Poslední sloupec nám udává průměrnou hodnotu odhadu rozptylu odhadu úhrnu, který je definován ve vzorci (4.5). Výsledky jsou v souladu s teoretickými závěry, $\hat{Y}$ je nevychýlený odhad úhrnu a jeho rozptyl klesá se zvětšujícím se rozsahem výběru. Tento rozptyl můžeme odhadnout nestranným způsobem.

Pro porovnání kvality našeho postupu nás zajímají vzájemné rozdíly hodnot a jejich odhadů. Ve sloupcích tabulky 7.2 máme počet vyšetřených stromů, průměrnou absolutní odchylku úhrnu lesa od odhadu úhrnu, průměrnou absolutní odchylku rozptylu odhadu úhrnu od jeho odhadu a relativní absolutní chybu odhadu rozptylu. Je jasné, že kvalita odhadů roste se zvětšujícím se n

n	$MAE\left[\hat{Y}:Y\right]$	$MAE\left[\hat{V}:V\right]$	$RMAE\left[\hat{V}:V\right]$
5	13,15	148,40	54,51 %
10	9,29	50,13	37,06 %
15	7,56	26,72	29,83 %
20	6,54	17,05	25,55 %
25	5,83	12,01	22,65 %
50	4,03	4,01	15,63 %
100	2,75	1,26	10,61 %

Tabulka 7.2: Porovnání kvality odhadů pro prostý náhodný výběr.

Obrázek A.1 znázorňuje, jak by takový prostý náhodný výběr 50 stromů mohl vypadat.

7.2 Systematický výběr

Druhou zkoumanou metodou aplikovanou na naší zalesněnou oblast je systematický výběr popsaný v podkapitole 4.1.2. Volili jsme sedm různých rozsahů vzorku. Jejich velikost n je uvedena v prvním sloupci tabulky 7.3. Druhý sloupec, stejně jako v příkladu předchozím, nám udává skutečný populační úhrn. Populační úhrn je odhadován pomocí Horvitzova-Thompsonova odhadu definovaného v rovnosti (4.9) a průměr odhadovaných hodnot ze 100 000 simulací vidíme ve třetím sloupci. Rozptyl odhadu úhrnu je počítán pomocí rovnosti (4.10) a jeho výsledky jsou uvedeny v předposledním sloupci. Odhad rozptylu odhadu úhrnu máme zobrazen v posledním sloupci a pro jeho výpočet jsme použili rovnosti (4.5). Zatímco $\hat{Y}$ je nevychýlený odhad úhrnu, tak u odhadu rozptylu se projevuje mírné vychýlení. Je

to způsobeno tím, že jsme použili odhad (4.5), který je určen pro prostý náhodný výběr.

n	Y	$\hat{Y}$	$V\left[\hat{Y}\right]$	$\hat{V}\left[\hat{Y}\right]$
4	256,80	256,80	326,32	340,78
8	256,80	256,79	176,43	168,97
16	256,80	256,81	87,35	82,31
24	256,80	256,80	46,40	55,72
32	256,80	256,81	31,30	40,65
64	256,80	256,81	16,74	19,50
128	256,80	256,80	10,85	8,95

Tabulka 7.3: Průměr odhadů přes 100 000 simulací pro systematický výběr.

Stejně jako v předchozím výběru porovnáme kvalitu jako průměrnou absolutní odchylku úhrnu lesa od odhadu úhrnu a průměrnou absolutní odchylku rozptylu odhadu úhrnu od jeho odhadu. Výsledné hodnoty jsou v tabulce 7.4. Opět vidíme, že máme nevychýlené odhady, jejichž kvalita odhadů stoupá s rostoucím rozsahem výběru.

n	$MAE\left[\hat{Y}:Y\right]$	$MAE\left[\hat{V}:V\right]$	$RMAE\left[\hat{V}:V\right]$
4	14,63	209,55	64,21 %
8	10,46	57,62	32,66 %
16	8,20	20,75	23,75 %
24	5,76	13,09	28,21 %
32	6,30	9,96	31,80 %
64	4,39	2,88	17,17 %
128	1,78	1,91	17,59 %

Tabulka 7.4: Porovnání kvality odhadů pro systematický výběr.

7.3 Dvouúrovňový výběr

Dalším výběrem využívajícím diskrétní populaci je dvouúrovňový výběr. Tento výběr máme podrobně popsán v podkapitole 4.3. Předpokládejme, že pro naše účely jsme les rozdělili do čtyřiceti shodných obdélníků o rozměrech 20×25 metrů. V první úrovni výběru vybíráme prostým náhodným výběrem určitý počet m obdélníků, ve kterých provedeme další šetření. V takto vybrané podoblasti našeho lesa provedeme prostý náhodný výběr stromů. Ve vybraných stromech naměříme šetřenou

vlastnost a odhadneme její charakteristiky. Můžeme tedy říci, že provádíme prostý náhodný výběr v obou úrovních výběru.

Podívejme se na výsledky uvedené v tabulce 7.5. V prvním sloupci máme počet vybraných obdélníků. Druhý sloupec nám říká, kolik stromů v daném obdélníku chceme vybrat. V tomto okamžiku může nastat situace, že ve vybraném obdélníku nebude možné vybrat daný počet stromů. V takovém případě vybereme maximální počet možných stromů. Ve třetím sloupci máme průměrný celkový počet vyšetřených stromů. Ve čtvrtém sloupci je celkový součet všech tlouštěk stromů v lese. Průměr odhadů úhrnu počítaných vzorcem (4.13) je uveden v dalším sloupci. V předposledním sloupci uvádíme rozptyl odhadu úhrnu počítaný dle vzorce (4.14). Poslední sloupec nám udává průměrnou hodnotu odhadu rozptylu odhadu úhrnu, který je dán vzorcem (4.15). Vysoké hodnoty rozptylu $V\left[\hat{Y}\right]$ oproti prostému a systematickému výběru jsou způsobeny velkou variabilitou v rámci výběru první úrovně. Zatímco tři obdélníky neobsahují vůbec žádné stromy, v některých úhrn přesahuje 1,5 násobek průměrného úhrnu na jeden obdélník. V rozkladu (4.12) jsou hodnoty V_{PVJ} daleko větší než V_{DVJ}. Pro volbu 5, 10 a 20 obdélníků jsou V_{PVJ} postupně 2818,73, 1208,03 a 402,68.

m	n_i	n	Y	$\hat{Y}$	$V\left[\hat{Y}\right]$	$\hat{V}\left[\hat{Y}\right]$
5	2	9,25	256,80	256,81	2 970,57	2 950,36
5	4	18,12	256,80	256,85	2 887,61	2 881,21
5	8	35,36	256,80	256,80	2 846,18	2 843,90
10	2	18,50	256,80	256,89	1 283,95	1 266,60
10	4	36,25	256,80	256,89	1 242,47	1 234,90
10	8	70,74	256,80	256,90	1 221,75	1 219,88
20	2	37,01	256,80	256,83	440,64	421,59
20	4	72,51	256,80	256,83	419,90	411,34
20	8	141,51	256,80	256,84	409,54	406,23

Tabulka 7.5: Průměr odhadů přes 100 000 simulací pro dvouúrovňový výběr.

Porovnání kvality jednotlivých odhadů a simulací je v tabulce 7.6. Vidíme, že kvalita závisí hlavně na počtu vybraných obdélníků.

Grafická interpretace jednoho z možných dvouúrovňových výběrů je na obrázku A.2. Na něm je znázorněn výběr deseti obdélníků, kde v každém obdélníku vybíráme čtyři stromy.

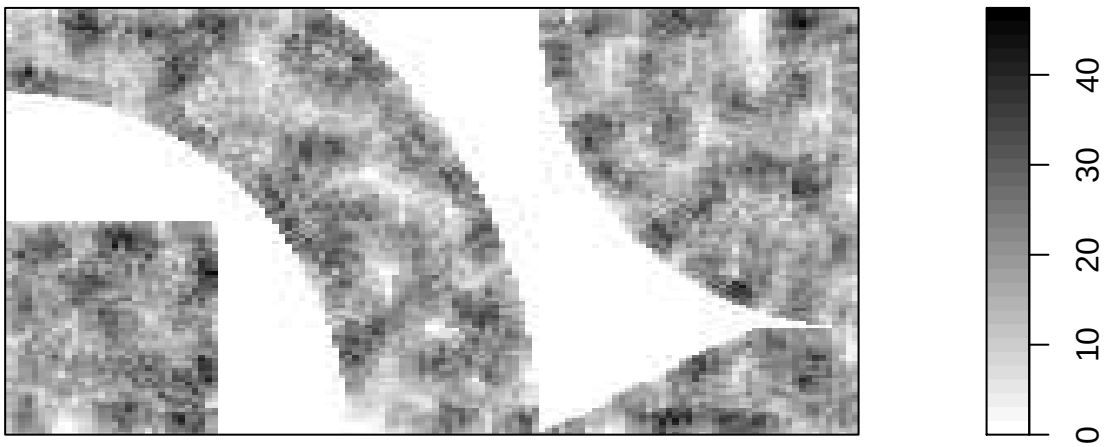
Výše popsané a provedené metody by se daly zařadit do kategorie pracující s diskrétním přístupem šetření populací. V dalších podkapitolách se blíže podíváme na výběry pracující se spojitými principy při šetření.

m	n_i	n	$MAE\left[\hat{Y}:Y\right]$	$MAE\left[\hat{V}:V\right]$	$RMAE\left[\hat{V}:V\right]$
5	2	9,25	44,09	1 438,20	48,41 %
5	4	18,12	43,52	1 417,35	49,08 %
5	8	35,36	43,22	1 406,72	49,42 %
10	2	18,50	28,72	364,40	28,38 %
10	4	36,25	28,27	354,56	28,54 %
10	8	70,74	28,02	351,05	28,73 %
20	2	37,01	16,80	70,38	15,97 %
20	4	72,51	16,39	66,34	15,80 %
20	8	141,51	16,19	64,80	15,82 %

Tabulka 7.6: Porovnání kvality odhadů pro dvouúrovňový výběr.

7.4 Jednofázový jednoúrovňový prostý náhodný výběr

V tomto odstavci předvedeme aplikaci metody zmíněné jako první v šesté kapitole. Jedná se o výběr, kdy náhodně vybereme určitý počet bodů v našem lese a kolem každého bodu utvoříme kruh. Všechny stromy, které do takto vzniklých kruhů patří, zahrneme do výběru. Připomeňme, že v tomto výběrovém plánu je naší zkoumanou charakteristikou populační průměr definovaný rovností (6.1), který můžeme chápat, jako spojitý průměr lokální hustoty $y(x)$. Pro náš les je graf lokální hustoty s volbou poloměru $r = 5\,\mathrm{m}$ znázorněn na obrázku 7.2.



Obrázek 7.2: Graf lokální hustoty s poloměrem 5 metrů.

Pro první typ výběru jsme zvolili poloměr vyšetřované kruhové oblasti $r = 3\,\mathrm{m}$ a počet vybraných ploch byl 10, 25 a 50. Pro poloměr nastavený na $r = 5\,\mathrm{m}$ jsme volili 5, 10 a 20 kruhových ploch. Výsledné zprůměrované hodnoty přes 100 000 simulací máme v tabulce 7.7. První sloupec nám udává zvolený poloměr vyšetřované kruhové plochy. Ve druhém sloupci máme počet vybraných kruhových ploch. Průměrný počet vyšetřených stromů máme ve sloupci třetím. Čtvrtý sloupec nám udává skutečný populační průměr. Ten, jak je intuitivní, je roven součtu tlouštěk jednotlivých stromů vydělenému celkovou plochou zalesněné oblasti. Průměr odhadů populačního průměru počítaných dle vzorce (6.2) vidíme v pátém sloupci. Sloupec předposlední udává rozptyl odhadu populačního průměru a je počítám za pomoci vzorce (6.3). Poslední sloupec hodnot je dán vzorcem (6.4) a jedná se o odhad rozptylu odhadu populačního průměru. Tabulka 7.7 nám potvrzuje, že odhady $\hat{\bar{Y}}$ i $\hat{V}\left[\hat{\bar{Y}}\right]$ jsou nevychýlené.

r	n_2	n	$\bar{Y}$	$\hat{\bar{Y}}$	$V\left[\hat{\bar{Y}}\right]$	$\hat{V}\left[\hat{\bar{Y}}\right]$
3	10	14,63	18,17	18,16	13,59	13,55
3	25	36,59	18,17	18,16	5,44	5,43
3	50	73,21	18,17	18,17	2,72	2,72
5	5	19,63	18,17	18,15	9,63	9,61
5	10	39,26	18,17	18,15	4,81	4,82
5	20	78,54	18,17	18,16	2,41	2,41

Tabulka 7.7: Průměr odhadů přes 100 000 simulací pro jednofázový jednoúrovňový prostý náhodný výběr.

r	n_2	n	$MAE\left[\hat{\bar{Y}} : \bar{Y}\right]$	$MAE\left[\hat{V} : V\right]$	$RMAE\left[\hat{V} : V\right]$
3	10	14,63	2,93	4,93	36,27
3	25	36,59	1,86	1,23	22,65
3	50	73,21	1,32	0,43	15,93
5	5	19,63	2,48	5,18	53,81
5	10	39,26	1,75	1,78	36,91
5	20	78,54	1,24	0,62	25,66

Tabulka 7.8: Porovnání kvality odhadů pro jednofázový jednoúrovňový prostý náhodný výběr.

Ve sloupcích tabulky 7.8 máme kromě poloměru kruhové oblasti, počtu vyšetřených kruhových ploch a průměrného počtu vyšetřených stromů také průměrnou

absolutní odchylku populačního průměru lesa od odhadu průměru, respektive průměrnou absolutní a relativně absolutní odchylku rozptylu odhadu průměru od jeho odhadu.

Na obrázku A.3 je znázorněn prostý náhodný výběr 20 kruhových oblastní o poloměru 5 metrů.

7.5 Jednofázový jednoúrovňový systematický výběr

Dalším přístupem souvisejícím se spojitou populací je systematický výběr. Podstata tohoto výběru je obdobná tomu předchozímu s tím rozdílem, že zde vybíráme náhodně pouze jeden bod. Veškeré další body jsou dány posunem po vertikální a horizontální ose. Vznikne nám tedy mříž bodů, kolem kterých vytvoříme kruhy o daném poloměru a v kruzích, jejichž střed je uvnitř lesa, provedeme šetření. Pro naše účely jsme volili pro poloměr kruhové plochy 3 metry jednotlivé posuny na horizontální ose 40 metrů a na vertikální 20 metrů pro první výběr, pro druhý výběr 30 metrů a 12 metrů a pro poslední výběr 20 metrů a 5 metrů. Pro poloměr plochy 5 metrů byly první posuny 60 metrů na horizontální ose a 30 metrů na vertikální ose, dále pak 50 a 20 metrů a poslední posun 40 a 10 metrů. Naše výsledky máme sumarizovány v tabulce 7.9. První sloupec udává poloměr vyšetřované kruhové plochy. Druhý sloupec udává průměrný počet vyšetřených ploch. Ve třetím sloupci máme průměrný celkový počet vyšetřených stromů. Ve čtvrtém sloupci je zobrazen skutečný populační průměr. Pátý sloupec udává průměrnou hodnotu odhadu populačního průměru. Šestý sloupec obsahuje hodnoty průměrné absolutní odchylky populačního průměru od jeho odhadu. Aproximaci rozptylu odhadu průměru máme ve sloupci předposledním. Poslední sloupec obsahuje průměrné hodnoty odhadu rozptylu počítaného dle vzorce (6.4). Jelikož lokality nejsou vybírány nezávisle, není tento odhad nevychýlený.

r	n_2	n	$\bar{Y}$	$\hat{\bar{Y}}$	$MAE\left[\hat{\bar{Y}} : \bar{Y}\right]$	$\tilde{V}\left[\hat{\bar{Y}}\right]$	$\hat{V}\left[\hat{\bar{Y}}\right]$
3	17,67	25,88	18,17	18,11	2,37	8,58	7,64
3	39,63	58,09	18,17	18,15	1,58	3,66	3,45
3	141,35	206,91	18,17	18,16	0,86	1,11	0,96
5	8,13	31,87	18,17	18,11	2,02	6,22	6,23
5	14,13	55,51	18,17	18,20	1,27	2,59	3,54
5	35,34	138,84	18,17	18,13	1,11	1,80	1,36

Tabulka 7.9: Průměr odhadů přes 100 000 simulací pro jednofázový jednoúrovňový systematický výběr.

Pro názornost systematického výběru se podívejme na obrázek A.4, kde máme systematický výběr s horizontálním posunem 40 metrů a s vertikálním posunem 10 metrů.

7.6 Jednofázový jednoúrovňový skupinkový výběr

V podkapitole 6.3 máme definovaný skupinkový výběr. Tento výběr je kombinací prostého náhodného výběru a systematického výběru pro spojité populace. Pro poloměr 3 i 5 metrů volíme stejný tvar skupinky. Skupinka se skládá ze čtyř bodů umístěných do vrcholů kosočtverce, jehož střed souměrnosti je námi vybraný náhodný bod. Přitom delší úhlopříčka leží v horizontálním směru a vrcholy jsou ve vzdálenosti 30 metrů od sebe a kratší úhlopříčka leží ve vertikálním směru a vrcholy jsou ve vzdálenosti 20 metrů. Tyto čtyři body tvoří středy kruhových oblastí. Náhodně vybíráme 5, 10 a 25 středů skupinky. Výsledky provedené simulace opakované 100 000-krát jsou zobrazeny v tabulce 7.10. V prvním sloupci máme poloměr jedné kruhové plochy. Ve druhém je zobrazený počet vyšetřených kruhových ploch. Sloupec třetí udává průměrný počet vyšetřených stromů. Čtvrtý sloupec nám udává hodnotu populačního průměru, který odhadujeme ve sloupci pátém. Šestý sloupec nám udává průměrné absolutní odchylky odhadu průměru od jeho skutečných hodnot. V předposledním sloupci nalezneme aproximace rozptylu odhadu populačního průměru. Průměrnou hodnotu odhadu rozptylu odhadu populačního průměru máme ve sloupci posledním. Výpočet tohoto odhadu jsme prováděli pomocí vzorce (6.5).

r	n_2	n	$\bar{Y}$	$\hat{\bar{Y}}$	$MAE\left[\hat{\bar{Y}} : \bar{Y}\right]$	$\tilde{V}\left[\hat{\bar{Y}}\right]$	$\hat{V}\left[\hat{\bar{Y}}\right]$
3	10,24	14,98	18,17	18,17	3,02	14,73	13,11
3	20,49	29,97	18,17	18,16	2,09	6,91	6,61
3	51,18	74,84	18,17	18,14	1,30	2,68	2,65
5	10,24	40,21	18,17	18,17	1,80	5,31	4,82
5	20,49	80,42	18,17	18,16	1,25	2,50	2,40
5	51,18	200,87	18,17	18,15	0,78	0,97	0,95

Tabulka 7.10: Průměr odhadů přes 100 000 simulací pro jednofázový jednoúrovňový skupinkový výběr.

Při skupinkovém výběru je velmi častým jevem, že do zalesněné plochy spadnou pouze některé kruhové plochy. Na obrázku A.5 vidíme téměř všechny možné kombinace. Nalezneme zde dvě plné skupinky, a to v levé dolní části a v levé části pásu. Dále zde máme jak tři kruhové plochy z celé skupinky, tak dvě kruhové plochy, a dokonce i pouze jednu kruhovou plochu ze čtyř původních.

7.7 Národní inventarizace lesů

Metodu, kterou si představíme v této podkapitole, je postup použitý při Národní inventarizaci lesů a popsaný v podkapitole 6.10. Musíme ovšem, s ohledem na velikost našeho pozorovacího okna, poupravit vzdálenosti označované jako l_1, l_2 a l_3. Pro oba poloměry kruhové plochy 3 a 5 metrů jsme volili totožné velikosti vzdáleností. V prvním případě je $l_1 = 25\,\mathrm{m}$ a $l_2 = 10\,\mathrm{m}$. Ve druhém případě jsou vzdálenosti 33,3 a 15 metrů a v posledním jsou vzdálenosti voleny jako 50 a 20 metrů. Vzdálenost l_3, která odpovídá poloměru r kruhové plochy, je buď 3, anebo 5 metrů. Výsledky máme sumarizovány v tabulce 7.11, kde máme poloměr kruhové plochy, průměrný počet vyšetřených stromů, populační průměr, průměrný odhad populačního průměru, průměrnou absolutní odchylku odhadu populačního průměru od jeho skutečné hodnoty, aproximaci rozptylu populačního průměru a konečně průměr odhadů tohoto rozptylu. Pro výpočet tohoto odhadu jsme využili vzorec (6.4) určený pro prostý náhodný výběr.

r	n	$\bar{Y}$	$\hat{\bar{Y}}$	$MAE\left[\hat{\bar{Y}} : \bar{Y}\right]$	$\tilde{V}\left[\hat{\bar{Y}}\right]$	$\hat{V}\left[\hat{\bar{Y}}\right]$
3	66,23	18,17	18,16	1,35	2,84	3,02
3	37,24	18,17	18,16	1,82	5,23	5,42
3	16,55	18,17	18,18	2,80	12,43	12,52
5	177,76	18,17	18,17	0,76	0,90	1,07
5	99,93	18,17	18,16	1,07	1,79	1,92
5	44,38	18,17	18,17	1,64	4,27	4,44

Tabulka 7.11: Průměr odhadů přes 100 000 simulací pro NIL.

Jak by jedna taková realizace výběru touto metodou mohla vypadat, máme znázorněno na obrázku A.6.

7.8 Znáhodněný stratifikovaný výběr

Postup uvedený v této podkapitole odpovídá podkapitole 6.11. Pro naše účely jsme pro poloměr 3 metry volili velikosti obdélníkových mřížek 40×20 metrů, 30×12 metrů a 20×5 metrů. Pro poloměr 5 metrů jsou velikosti mřížek 60×30 metrů, 50×20 metrů a 40×10 metrů. Výsledky máme zobrazeny v tabulce 7.12, kde máme poloměr kruhové plochy, průměrný počet vyšetřených stromů, populační průměr, průměrný odhad populačního průměru, průměrnou absolutní odchylku odhadu populačního průměru od jeho skutečné hodnoty, aproximaci rozptylu populačního průměru a konečně průměr odhadů tohoto rozptylu. Pro výpočet tohoto odhadu jsme využili vzorec (6.4) určený pro prostý náhodný výběr.

r	n	$\bar{Y}$	$\hat{\bar{Y}}$	$MAE\left[\hat{\bar{Y}} : \bar{Y}\right]$	$\tilde{V}\left[\hat{\bar{Y}}\right]$	$\hat{V}\left[\hat{\bar{Y}}\right]$
3	25,88	18,17	18,17	2,21	7,68	7,83
3	57,52	18,17	18,16	1,46	3,34	3,48
3	206,99	18,17	18,17	0,73	0,84	0,96
5	30,90	18,17	18,18	2,00	6,36	6,50
5	55,52	18,17	18,16	1,45	3,35	3,49
5	138,88	18,17	18,17	0,89	1,25	1,37

Tabulka 7.12: Průměr odhadů přes 100 000 simulací pro znáhodněný stratifikovaný výběr.

Příklad znáhodněného stratifikovaného výběru s obdélníkovou sítí 50×20 metrů máme na obrázku A.7.

Kapitola 8

Závěr

V této práci jsme se zabývali výběrovými šetřeními uplatnitelnými v lesnictví. Práce je členěna do tří základních celků. V prvním, v kapitole 1 a 2, se věnujeme úvodu do problematiky, důvodům a příčinám vzniku populačních šetření. Dále sjednocujeme základní značení a hlavní pojmy problematiky výběrových metod. Několik odstavců věnujeme hlavním aspektům a cílům šetření, mezi nimi jsou i charakteristiky, které počítáme a odhadujeme. Vysvětlujeme základní dva přístupy, diskrétní a spojitý, které rozlišují případy, kdy je populace tvořena konečnou nebo nekonečnou množinou prvků.

Ve druhé části, kam spadají kapitoly 3, 4, 5 a 6, získáváme teoretický základ. Nejdříve popisujeme obecně metody šetření. Další kapitola pojednává o některých konkrétních diskrétních metodách a jejich teoretických výpočtech. Ta je následována obdobným popisem pro spojité populace. Teoretická část vrcholí přehledem metod používaných v praxi společně s jejich vzorci pro výpočet jednotlivých charakteristik a odhadů. Zde najdeme kvalitní a ucelený obraz o výběrových šetřeních používaných v lesnictví.

Pro většinu zájemců o problematiku šetření v přírodě je nejdůležitější třetí část, která je v kapitole 7. Aplikace teoretických poznatků z předchozích kapitol je jedním z hlavních přínosů práce. Několik vzájemně rozdílných metod použitých v prostředí, které můžeme označit za reálný les, nabízí více než zajímavé porovnání. Pro každou ze zvolených metod bylo provedeno dostatečné množství simulací, abychom výsledky považovali za dostatečně vypovídající o dané metodě. Pro simulaci bylo použito programu R. Zároveň se setkáváme u každé metody s jistou variací, zejména ve velikosti vzorku. To může být, například při omezených prostředcích na šetření, neméně důležitý úhel pohledu. Samozřejmostí je doplnění popsaných metod o vizuální interpretaci.

Literatura

[1] L. Barabesi and S. Franceschi: Sampling properties of spatial total estimators under tessellation stratified designs, *Environmetrics*, **22**, 271–278, 2011.

[2] W. G. Cochran: *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York, 1977.

[3] C. B. Cordy: An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe, *Statistics & Probability Letters*, **18**, 353–362, 1993.

[4] T. G. Gregoire and H. T. Valentine: *Sampling Strategies for Natural Resources and the Environment*, Chapman & Hall/CRC, Boca Raton, 2008.

[5] J. Illian, A. Penttinen, H. Stoyan and D. Stoyan: *Statistical Analysis and Modelling of Spatial Point Patterns*, John Wiley & Sons, Chichester, 2008.

[6] Kolektiv autorů: *Národní inventarizace lesů v České republice 2001-2004*, Ústav pro hospodářskou úpravu lesů Brandýs nad Labem, ČTK REPRO, 2007.

[7] D. Mandallaz: *Sampling Techniques for Forest Inventories*, Chapman & Hall/CRC, Boca Raton, 2008.

[8] R Development Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[9] D. L. Stevens: Variable density grid-based sampling designs for continuous spatial populations, *Environmetrics*, **8**, 167–195, 1997.

[10] Š. Šmelko: *Dendrometria*, Technická univerzita ve Zvolene, Zvolen, 2000.

[11] M. E. Thompson: *Theory of Sample Survey*, Chapman & Hall/CRC, London, 1997.

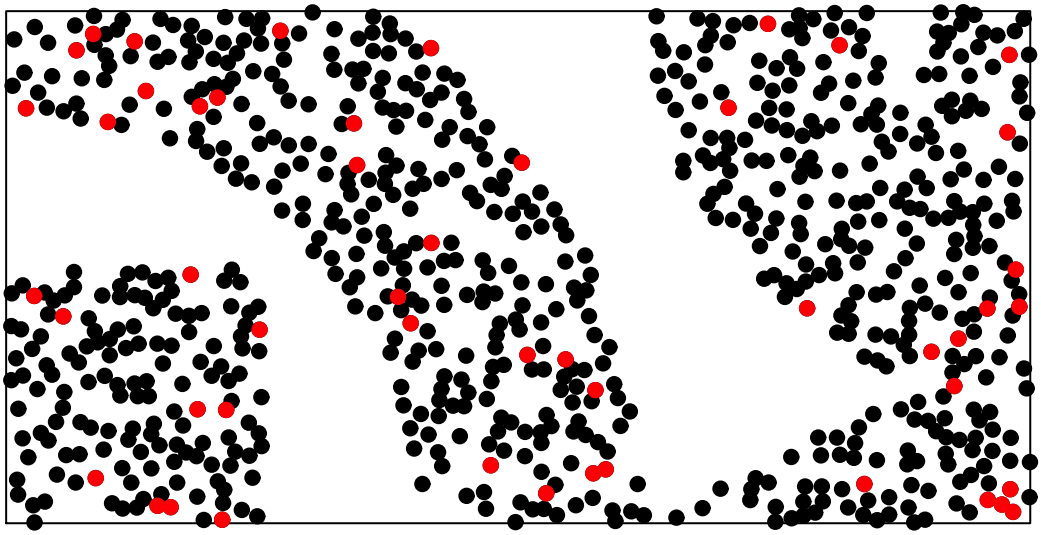
[12] D. Vorlíčková: *Výběry z konečných souborů*, Univerzita Karlova, Praha, 1985.

Seznam obrázků

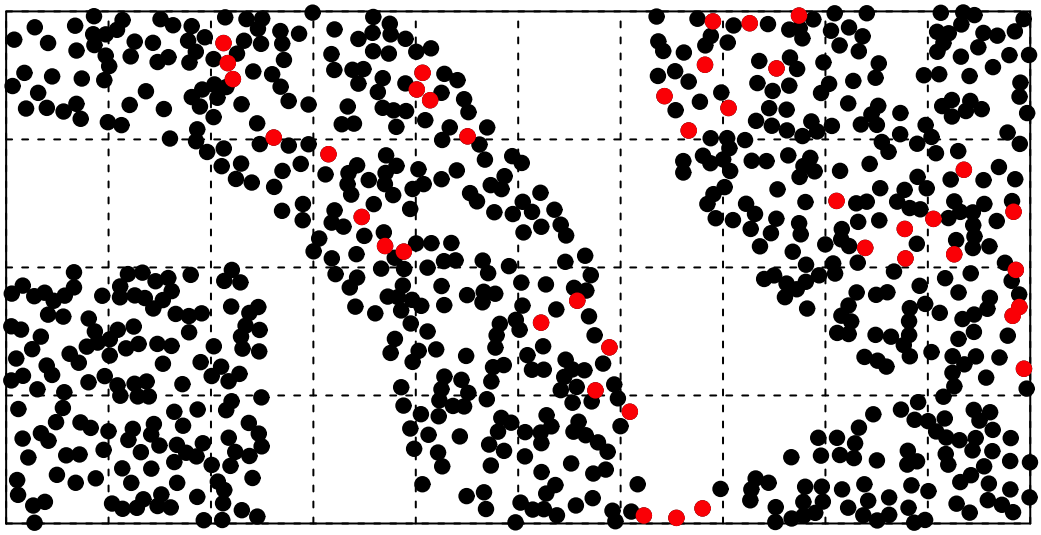
Seznam tabulek

Příloha A

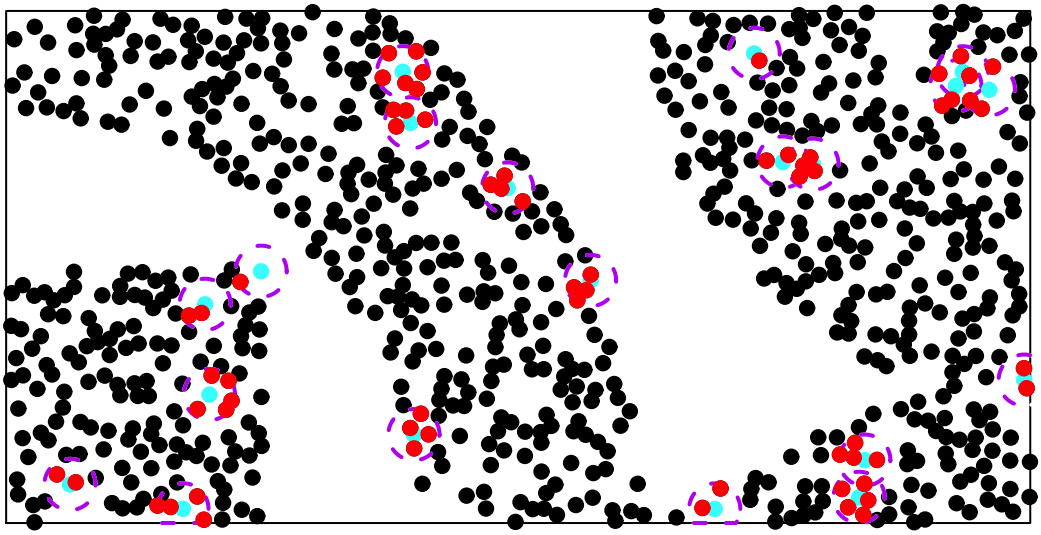
Grafická znázornění výběrů



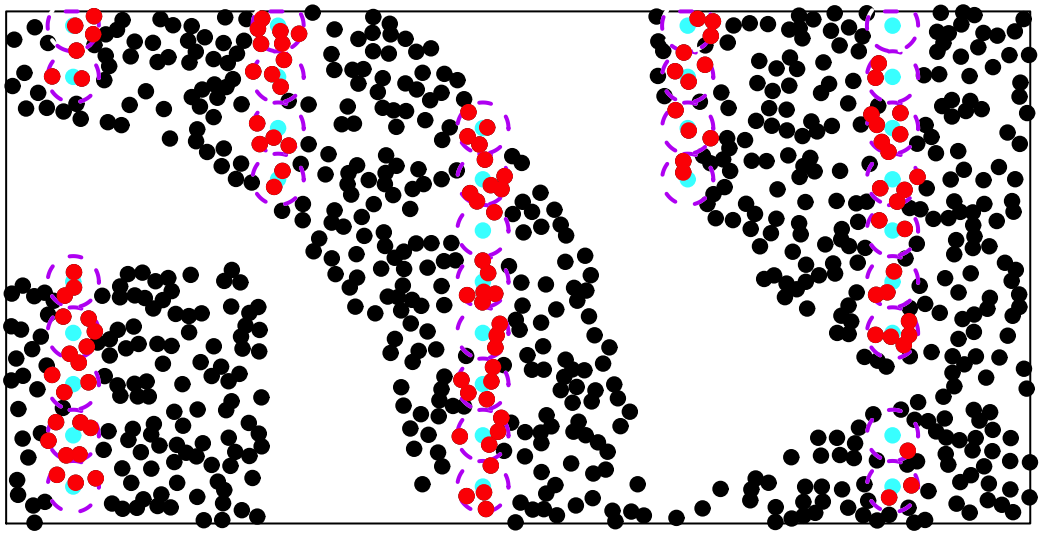
Obrázek A.1: Prostý náhodný výběr 50 stromů v lese.



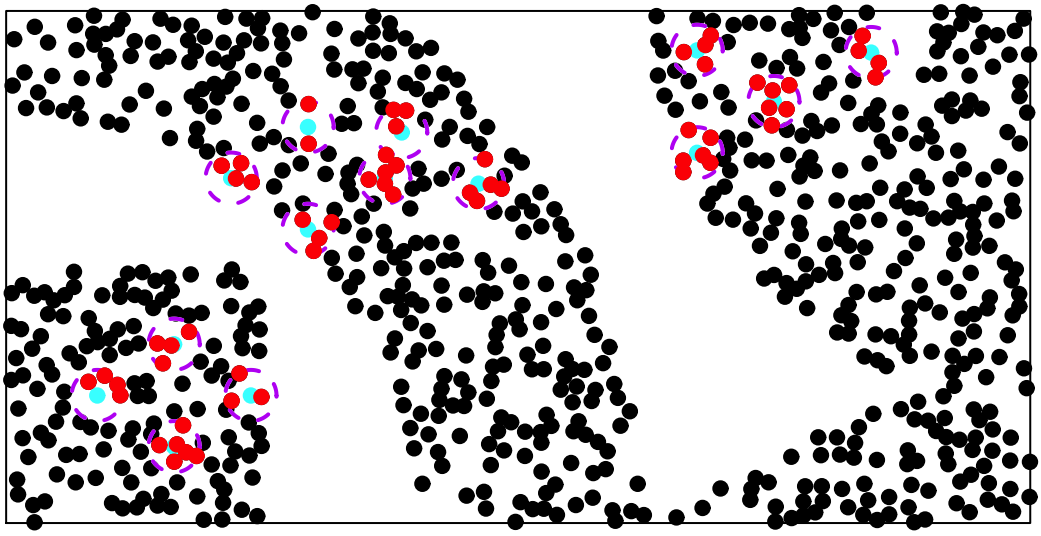
Obrázek A.2: Dvouúrovňový výběr 10 obdélníků v lese, kde v každém obdélníku vybíráme 4 stromy.



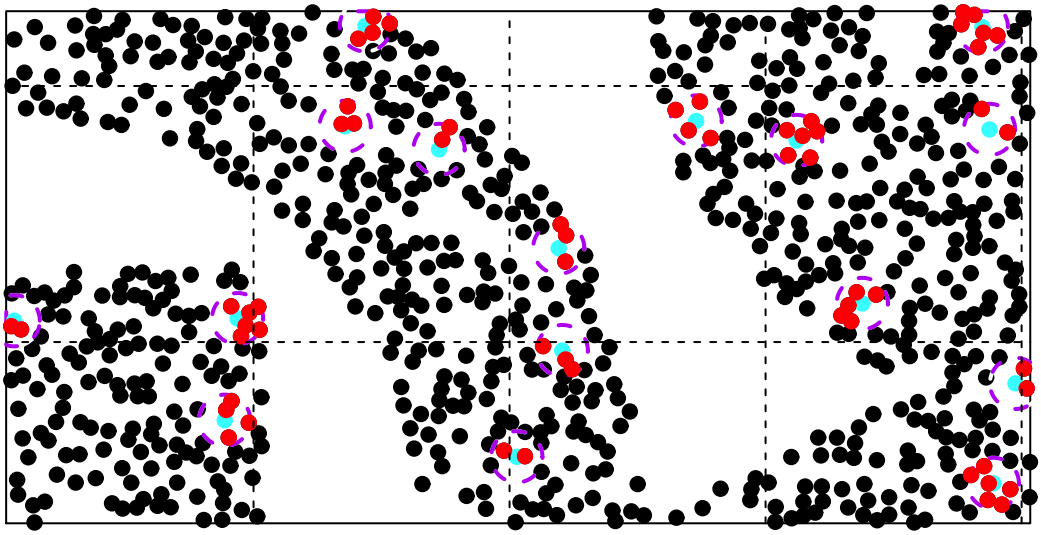
Obrázek A.3: Jednofázový prostý výběr 20 kruhových oblastí o poloměru 5 metrů.



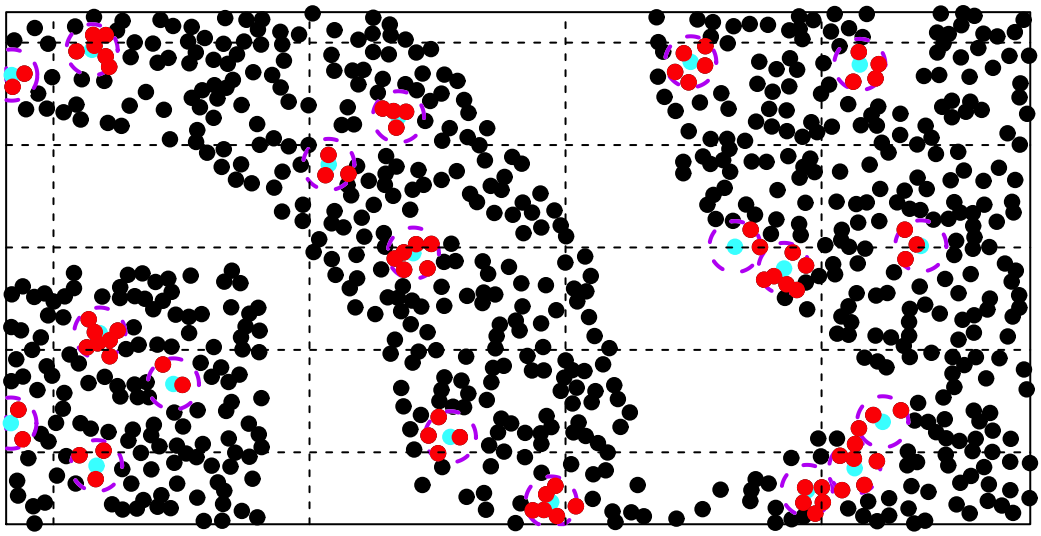
Obrázek A.4: Jednofázový systematický výběr oblastí v lese s horizontálním posunem 40 metrů a vertikálním posunem 10 metrů.



Obrázek A.5: Jednofázový skupinkový výběr 5 skupinek.



Obrázek A.6: Realizace metody NIL se čtvercovou sítí se vzdáleností sousedních bodů 50 metrů.



Obrázek A.7: Realizace znáhodněného stratifikovaného výběru s obdélníkovou sítí 50×20 metrů.