UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Vybrané metody výběru

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2011

Vypracoval:
Bc. Romana Schneiderová
AME, V. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně za vedení RNDr. Karla Hrona, Ph.D. a že jsem v seznamu literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 31. března 2011

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zaslouží poděkování moje rodina, která mě po celou dobu studia podporovala.

Obsah

Úvod

Cílem této práce je přehledně shrnout jednotlivé metody pravděpodobnostního výběru v teorii výběrových šetření, a to zejména stratifikovaný, skupinový, systematický a vícestupňový výběr.

V praxi dochází k různým výběrovým situacím, proto jsou náklady na pořízení a zpracování výběru různé, ať už se to týká peněz nebo času. Důležité je také porovnání výše uvedených metod, zjišťování jejich předností nebo naopak jejich nevýhod. Následně pak volíme vždy takovou výběrovou metodu, pro kterou budou náklady pro danou populaci (základní soubor) minimální. Další etapou po pořízení výběru je odhadování výběrových charakteristik, a to výběrového úhrnu a výběrového průměru. Pomocí těchto odhadů zjišťujeme požadovanou vlastnost o výběru, kterou pak můžeme zobecnit pro celou populaci. Každá metoda je ilustrována na příkladě a na závěr jsou teoretické poznatky demonstrovány na reálných datech z Českého statistického úřadu.

Metody výběrových šetření nachází především uplatnění v oblasti tzv. úřední statistiky (official statistics), u nás provozované již zmíněným Českým statistickým úřadem, ale i některými ministerstvy a orgány státní správy. Dalšími oblastmi uplatnění jsou průzkumy veřejného mínění, marketingová šetření nebo výběrová zpracování evidence pojistných událostí ve velkých pojišťovnách.

První kapitola seznamuje čtenáře s pojmem statistického šetření a jeho různými podobami a děleními.

Druhá kapitola je již věnována konkrétním metodám výběrového šetření; jedná se o základní metody, které je třeba objasnit, neboť na nich jsou založené ostatní metody výběru.

Třetí a zároveň nejdůležitější kapitola popisuje čtyři uvedené složitější metody výběru.

Nakonec je uvedena zmíněná studie s reálnými daty, která jsou zpracována vybranými metodami (datový soubor je přiložen na CD).

Stěžejní literaturou se pro tuto diplomovou práci stala kniha Sampling od S. K. Thompsona.

Snažila jsem se, aby tato práce byla pro čtenáře srozumitelná a pomohla jim lépe se orientovat v problematice výběrových šetření. Upustila jsem proto (až na výjimky) od odvozování vzorců, které si čtenář může dohledat v uvedené literatuře, a zabývala se především jejich aplikací v praxi.

1 Obecný úvod k výběrovým šetřením

Při psaní této kapitoly jsem použila literaturu [1], [5] a [9].

U statistického šetření je potřeba si vymezit předmět zkoumání, obsah zkoumání a metody, pomocí nichž provedeme celé šetření. Předmětem statistického šetření je vyšetřování *statistického souboru (populace, základní soubor)*. Každá populace se skládá z určitého počtu jednotek, na nichž se provádí šetření (zjišťujeme o nich požadované údaje) z hlediska určitých statistických znaků. Znakem je myšlena vlastnost jednotky, která ji charakterizuje a kterou zjišťujeme. Je-li populace složena např. z osob, můžeme u nich zjišťovat pohlaví, věk, v jakém regionu žijí, socioekonomické faktory atd. Podle velikosti statistického souboru se můžeme rozhodnout, zda provedeme statistické šetření úplné nebo neúplné.

1.1 Úplná a neúplná statistická šetření

Úplné statistické šetření provádíme, pokud chceme zjistit požadované údaje u všech jednotek z populace. Příkladem je stav hospodářských zvířat v zemědělských podnicích k určitému dni v České republice, žádný ze závodů nemůže být ze soupisu vyloučen. Úplné šetření podává přesné charakteristiky (úhrny a průměry) nejen o souboru, ale i o každé jednotce zvlášť, je tudíž nezastupitelné tam, kde se požaduje informace o každé statistické jednotce. Avšak z praktického hlediska je u rozsáhlých souborů toto šetření časově náročné, nákladné a mohlo by přinést i zcela chybné údaje u části souboru. Proto jeho meze použitelnosti jsou značně omezené a spíše se v praxi využívá neúplné statistické šetření.

Při *neúplném statistickém šetření* buď úmyslně nebo náhodně vybíráme určitý počet jednotek, na kterých pak provádíme šetření. Z výsledků pak usuzujeme pro celý soubor. Neúplné šetření nám tedy poskytuje přesné charakteristiky za prošetřenou část a za celý soubor jen přibližné hodnoty (odhady) těchto charakteristik. Ne ovšem každé šetření se dá zobecnit pro celý soubor, a proto rozlišujeme různé druhy neúplného šetření.

Mezi hlavní druhy neúplného šetření patří: anketa, metoda základního masívu, průzkum a výběrová šetření.

Anketa je takový druh šetření, kdy se dotazujeme daného okruhu osob na předem pečlivě zvolené otázky, které se týkají určitého problému, avšak jen malá část z nich je ochotna poskytnout odpovědi. Většinou se ale nedělají závěry na celek, neboť mezi odpovědí nebo jejím odmítnutím a dotazovanou skutečností bývá dosti úzký vztah. Například osoby s vysokým příjmem často neodpoví pravdivě na otázku o jejich výši příjmů.

Metodu základního masívu lze užít tehdy, skládá-li se statistický soubor z několika velkých a velkého počtu malých jednotek (např. podniky ve stavebním průmyslu). Potom se šetření provede na velkých jednotkách a malé můžeme vynechat. Sice se tímto sníží pracnost a prošetří se převážná část souboru, nicméně nám tato metoda nedovoluje získané charakteristiky zobecnit pro celý soubor, protože v neprošetřené části mohou jednotky vykazovat jiné zákonitosti.

Průzkum je obdobou výběrového šetření, kterému jsou věnovány následují odstavce, avšak liší se od něj především svým obsahem a počtem jednotek zahrnutých do výběru. U výběrového šetření máme k dispozici data, která jsou objektivně měřitelná, kdežto u průzkumu nás zajímají postoje a názory dotazovaných na určitý problém, proto je i počet vybraných jednotek daleko menší. Abychom objektivně zjistili názor dotazovaného, používá se u průzkumu řada psychologických a sociologických metod. Užívá se především při průzkumech veřejného mínění, v různých sociologických průzkumech apod. Jeho výhodou (stejně jako u výběrového šetření) je, že výsledky průzkumu lze aplikovat na celý soubor.

1.2 Výběrová šetření

Výběrové šetření je jedním z nejdůležitějších druhů neúplného statistického šetření. Hlavní myšlenkou tohoto šetření je, že ze statistického souboru (populace) vybereme určitý počet jednotek do výběru, na kterém pak provádíme šetření, a z výsledků pak usuzujeme pro celý soubor. Rozlišujeme dva druhy výběrového šetření: záměrný výběr a náhodný výběr.

Záměrný výběr je charakterizován tím, že závisí na rozhodnutí zadavatele, které jednotky budou zahrnuty do výběru a které nikoliv. Právě na těchto vybraných jednotkách by se mělo co nejlépe provádět zamýšlené statistické šetření. Získané charakteristiky můžeme rozšířit na celý soubor, avšak výsledky mohou být příliš subjektivně zatíženy.

Při *náhodném (pravděpodobnostním) výběru* jsou jednotky vybrány zcela náhodně, nezávisle na sobě a na mínění zadavatele šetření. V této práci se nadále budu zabývat jen náhodným výběrem.

1.3 Náhodný (pravděpodobnostní) výběr

U náhodného výběru uvažujeme, že celkový počet jednotek v celé populaci je N a počet jednotek zahrnutých do výběru a následně zkoumaných je n. Každá jednotka je do výběru zahrnuta s určitou pravděpodobností. Nejčastěji má každá jednotka stejnou pravděpodobnost, že bude vybrána. Ale samozřejmě existují i výběry s nestejnými pravděpodobnostmi, o nichž se zmíníme později.

Jak už bylo uvedeno, zkoumá se pouze ta část populace, která je vybrána, čili n jednotek. Proto budou výsledkem tzv. výběrové charakteristiky, a to zejména výběrový průměr a výběrový úhrn, což jsou odhady příslušných populačních charakteristik, tj. populačního průměru a populačního úhrnu. Populační ukazatelé jsou určeny z celého souboru N jednotek a jsou tedy skutečnými hodnotami. Chceme, aby odhady byly co nejpřesnější, což nezávisí jen na rozsahu výběru ale i na způsobu výběru jednotek. Proto využíváme náhodný výběr jednotek s již předem známými pravděpodobnostmi, neboť odhady, které získáme, jsou statistickými odhady. A proto lze jejich přesnost při daném rozsahu výběru objektivně změřit, můžeme stanovit i interval, v němž se bude téměř jistě nacházet skutečná hodnota. Dále můžeme říci, že tyto odhady jsou konzistentní, pokud s rostoucím rozsahem výběru konvergují ke skutečné hodnotě, a nestranné, pokud skutečnou hodnotu v průměru ani nepodhodnocují ani nenadhodnocují při každém rozsahu výběru. Proto bychom se měli vyvarovat toho, že do výběru budeme zahrnovat jen ty jednotky, u kterých snáze zjistíme požadované údaje. Tím bychom naopak

docílili velkého odlišení výběrových ukazatelů od populačních. Např. při zkoumání tělesných rozměrů kojenců je pohodlnější vybrat kojence z jeslí, není ovšem vždy zaručeno, že se tam nachází reprezentativní vzorek celé populace.

2 Základní metody výběru

V této kapitole jsem nejvíce vycházela z literatury [1], [2], [7], a dále pak i z literatury [4], [5] a [6].

Pokud se provádí statistické šetření na základě výběru, je třeba si stanovit metodu, pomocí které bude výběr pořízen. Především je dobré vědět, jaká je velikost a struktura statistického souboru, dále pak, jak moc si jsou nebo nejsou podobné jednotky v souboru a jaké charakteristiky chceme odhadovat, a to proto, abychom stanovili tu správnou výběrovou metodu.

V praxi častěji používaný je *výběr se stejnými pravděpodobnostmi*, protože je teoreticky jednodušší a v praxi snazší. Znamená to tedy, že každá jednotka z celé populace má stejnou pravděpodobnost, že bude do výběru zařazena.

Další metodou je *výběr s nestejnými pravděpodobnostmi*, kdy každá jednotka má různou pravděpodobnost, že se dostane do výběru.

Druhou možností, podle níž můžeme náhodné výběry dělit, je s vracením nebo bez vracení vybraných jednotek zpět do základního souboru po jejich prošetření. V prvním případě se jedná o *výběr s vracením*, kdy každá vybraná a prošetřená jednotka je do souboru vrácena ještě předtím, než je vybrána další, a může se tedy ve výběru opakovat. Využívá se především při výběrech s nestejnými pravděpodobnostmi. V druhém případě jde o *výběr bez vracení* a to znamená, že vybranou jednotku již do souboru nevracíme, tudíž nemá možnost opakovaného výběru. Z teoretického hlediska je výhodnější ten první, protože výběry jednotlivých jednotek jsou realizovány nezávisle na sobě, a proto vzorce i úvahy budou jednodušší. V praxi je naopak tendence více užívat výběry bez vracení.

2.1 Prostý náhodný výběr

Prostý náhodný výběr je nejjednodušší metodou náhodného výběru a spočívá v přímém výběru jednotek se stejnými pravděpodobnostmi. Tento výběr je základem pro teorii ostatních, složitějších, výběrů, a proto je nezbytné ho uvést hned na začátku. Dále budeme mluvit o prostém náhodném výběru bez vracení, což je výběrová metoda, ve které je n různých jednotek vybráno z N jednotek celé

populace takovým způsobem, že každá možná kombinace vybraných n jednotek má stejnou pravděpodobnost, že bude do výběru zahrnuta. Po vybrání jednotky už ji nevracíme zpět, takže každá jednotka se ve výběru může vyskytnout jen jednou. Pravděpodobnost, že bude i-tá jednotka zařazena do výběru je

$$\pi_i = \frac{n}{N}.$$

Metody jiné než prostý náhodný výběr mohou také dávat každé jednotce stejnou pravděpodobnost zahrnutí do výběru, ale jen u prostého náhodného výběru má každý možný výběr složený z n jednotek stejnou pravděpodobnost, že bude realizován. Po výběru jednotek přecházíme k další stránce šetření, což je odhadování základních charakteristik.

2.1.1 Odhadování populačního průměru a úhrnu

Zkoumaný číselný znak (proměnnou) označíme y, přičemž hodnoty, kterých nabývá na jednotlivých jednotkách populace, značíme $y_1, y_2, ..., y_N$.

Populační průměr μ se určí jako průměr y-nových hodnot z celé populace

$$\mu = \frac{1}{N}\left(y_1 + y_2 + \cdots + y_N\right) = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{1}$$

a výběrový průměr $\bar{y}$ jako průměr y-nových hodnot ve výběru,

$$\bar{y} = \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right) = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2}$$

kde y_i je hodnota zjišťovaného znaku na i-té jednotce ve výběru. U prostého náhodného výběru je výběrový průměr $\bar{y}$ nestranným odhadem populačního průměru μ, tedy $E(\bar{y}) = \mu$.

Speciálním případem prostého náhodného výběru je odhadování proporcí. Pokud například chceme zjistit podíl mužů ve zkoumané populaci, potom y-ové hodnoty nabývají pouze dvou hodnot, 0 a 1. V případě, že i-tá jednotka bude splňovat danou vlastnost (muž), pak $y_i = 1$, pokud tuto vlastnost splňovat nebude (žena),

je $y_i = 0$. Pro výpočet proporce v populaci s danou vlastností užijeme (1) nebo lze také odvodit speciální vztahy [7].

Jako míru variability budeme používat rozptyl. Konečný populační rozptyl je dán vztahem

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2$$

a výběrový rozptyl je definován jako

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 . \tag{3}$$

U prostého náhodného výběru také platí, že výběrový rozptyl s^2 je nestranným odhadem konečného populačního rozptylu σ^2, tj. $E(s^2) = \sigma^2$. Dále pak rozptyl odhadu $\bar{y}$ u prostého náhodného výběru je

$$var\,(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n} \tag{4}$$

a nestranný odhad tohoto rozptylu (rovněž tedy platí, že $E[\widehat{var}(\bar{y})] = var(\bar{y})$) je

$$\widehat{var}\,(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{s^2}{n}. \tag{5}$$

Odmocninu z rozptylu odhadu nazýváme jeho směrodatnou odchylkou. Pro směrodatnou odchylku z (5) ale neplatí, že by byla nestranným odhadem směrodatné odchylky z (4).

Výraz $(N-n)/N$, který můžeme psát také ve tvaru $1-(n/N)$, nazýváme korekčním faktorem (konečnostním násobitelem). Pokud je rozsah populace N ve srovnání s rozsahem výběru n velký, potom výraz n/N je malý, korekční faktor se bude blížit hodnotě 1 a rozptyl výběrového průměru $\bar{y}$ bude mít přibližně hodnotu σ^2/n. Vynechání korekčního faktoru při odhadování rozptylu $\bar{y}$ bude mít za následek, že dostaneme mírně nadhodnocené odhady skutečného rozptylu.

Při výběru z malé populace může mít korekční faktor podstatný vliv na snížení rozptylu tohoto odhadu. Jestliže se totiž rozsah výběru blíží k rozsahu populace, potom korekční faktor se blíží k nule (při prostém náhodném výběru) a tudíž i rozptyl odhadu $\bar{y}$ se blíží k nule.

S průměrem úzce souvisí úhrn hodnot zkoumaného znaku. Jelikož je populační úhrn N-násobek populačního průměru,

$$\tau = \sum_{i=1}^{N} y_i = N\mu,$$

je odhadem populačního úhrnu N-násobek výběrového průměru,

$$\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i, \tag{6}$$

a opět platí, že tento odhad je nestranný, tedy $E(\hat{\tau}) = \tau$.

Jestliže je odhad $\hat{\tau}$ N-násobkem odhadu $\bar{y}$, potom rozptyl odhadu $\hat{\tau}$ je N^2-násobkem rozptylu odhadu $\bar{y}$. Tedy

$$var\left(\hat{\tau}\right) = N^2 var\left(\bar{y}\right) = N\left(N-n\right)\frac{\sigma^2}{n}$$

a nestranný odhad tohoto rozpylu je

$$\widehat{var}\left(\hat{\tau}\right) = N^2\widehat{var}\left(\bar{y}\right) = N\left(N-n\right)\frac{s^2}{n}. \tag{7}$$

Odhad výběrového průměru $\bar{y}$ je náhodná veličina, jejichž hodnota závisí na složení výběru, tzn. z jakých y-ových hodnot byl výběr složen. Pro jakýkoliv jiný výběr může být hodnota $\bar{y}$ vyšší nebo nižší než populační průměr μ, ale střední (očekávaná) hodnota $\bar{y}$ přes všechny možné výběry z populace je vždy rovna μ. Odhad $\bar{y}$ je tedy skutečně právem nazýván nestranným odhadem pro populační průměr μ, protože pravděpodobnost, s níž se očekávání hodnotí, plyne

z pravděpodobností jednotlivých výběrů (vzhledem k metodě). Z toho důvodu nestrannost výběrového průměru vůči populačnímu průměru nezávisí na žádném předpokladu o populaci samotné.

Rozptyl výběrového průměru závisí na variabilitě jednotek v populaci, čím větší je variabilita v populaci, tím větší je i variabilita výběrového průměru a naopak. Rozptyl výběrového průměru také závisí, nepřímo úměrně, na rozsahu výběru. Největší hodnoty nabývá rozptyl, pokud je roven populačnímu rozptylu σ^2, a to je v případě(až na korekční faktor), kdy vybereme pouze jednu jednotku. Čím bude rozsah výběru větší, tím menší rozptyl bude. Rozptyl bude roven nule pro maximální možnou hodnotu n, tj. n rovno N. Rozptyly odhadů jsou, jak bylo uvedeno výše, také nestrannými odhady jejich populačních protějšků.

2.1.2 Prostý náhodný výběr s vracením

Principem je, že vybíráme n jednotek z celkového počtu N jednotek v celé populaci, přičemž každou vybranou jednotku vracíme zpět ještě před výběrem další, tedy jedna jednotka může být vybrána vícekrát. Těchto n výběrů je nezávislých a každá jednotka v populaci má stejnou pravděpodobnost zahrnutí do výběru. Výhodou této metody je, že tedy nemusíme dávat pozor na to, která jednotka byla zahrnuta do výběru více než jednou. Avšak pro daný rozsah výběru n je prostý náhodný výběr s vracením méně eficientní než prostý náhodný výběr bez vracení.

Výběrový průměr n pozorování $\bar{y}_n$ je tvaru

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

a platí, jestliže je jednotka vybrána vícekrát, potom její y-ová hodnota je v odhadu užita více než jedenkrát.

Rozptyl $\bar{y}_n$ je

$$var(\bar{y}_n) = \frac{1}{nN}\sum_{i=1}^{N} (y_i - \mu)^2 = \frac{N-1}{nN}\sigma^2.$$

Z toho můžeme usoudit, že rozptyl výběrového průměru prostého náhodného výběru bez vracení je $(N-n)/(N-1)$ krát menší než rozptyl výběrového průměru s vracením. I zde lze říci, že rozptyl výběrového průměru závisí nepřímo úměrně na rozsahu výběru, ovšem u výběru s vracením můžeme rozsah výběru zvětšovat neomezeně, takže nulový rozptyl dostaneme pro n rovno nekonečnu.

Nestranný odhad rozptylu $\bar{y}_n$ je

$$\widehat{var}(\bar{y}_n) = \frac{s^2}{n}.$$

Odhad $\bar{y}_n$ závisí na tom, kolikrát byla každá jednotka vybrána. Pokud totiž budu mít dva výběry se stejným souborem navzájem si různých jednotek, ale s jejich rozdílným opakováním ve výběrech, může to obecně přinášet rozdílné odhady.

Počet různých jednotek obsažených ve výběru nazýváme efektivní rozsah výběru a značíme ν. Potom $\bar{y}_\nu$ bude výběrový průměr různých jednotek, jehož výpočet je

$$\bar{y}_\nu = \frac{1}{\nu} \sum_{i=1}^{\nu} y_i.$$

Odhad $\bar{y}_\nu$ je nestranným odhadem populačního průměru. Můžeme říci, že rozptyl $\bar{y}_\nu$ je menší než rozptyl $\bar{y}_n$, ale stále ještě nebude menší než rozptyl výběrového průměru $\bar{y}$ prostého náhodného výběru bez vracení.

2.1.3 Odvození pro náhodný výběr

U prostého náhodného výběru můžeme střední hodnotu výběrového průměru odvodit pomocí alternativního rozdělení. Pro každou i-tou jednotku z populace definujeme ukazatel proměnné z_i, pro něhož platí, jestliže je i-tá jednotka zahrnuta do výběru, je $z_i = 1$, v opačném případě $z_i = 0$. Potom výběrový průměr můžeme psát ve tvaru

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{N} y_i z_i,$$

kde y_i je číslo a z_i je náhodná veličina se střední hodnotou $E(z_i) = P(z_i = 1) = n/N$. Pak střední hodnota výběrového průměru je

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} y_i E(z_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu.$$

Tímto jsme dokázali, že výběrový průměr $\bar{y}$ je nestranným odhadem populačního průměru μ.

Rozptyl náhodného výběru může být odvozen obdobně,

$$var(\bar{y}) = var\left(\frac{1}{n}\sum_{i=1}^{N} y_i z_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 var(z_i) + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j cov(z_i, z_j)\right].$$

Protože z_i je náhodná veličina s alternativním rozdělením, je $var(z_i) = (n/N)(1-n/N)$.

Počet výběrů obsahující obě jednotky i a j, kde $i \neq j$, je $\binom{N-2}{n-2}$, a tak pravděpodobnost zahrnutí obou jednotek do výběru je $\binom{N-2}{n-2}/\binom{N}{n} = n(n-1)/[N(N-1)]$. Výraz $z_i z_j$ je roven nule kromě případu, kdy jsou obě jednotky zahrnuty do výběru, tedy

$$E(z_i z_j) = P(z_i = 1, z_j = 1) = \frac{n(n-1)}{N(N-1)}.$$

Kovariance je

$$cov(z_i, z_j) = E(z_i z_j) - E(z_i)E(z_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{-n(1-n/N)}{N(N-1)}.$$

Potom rozptyl výběrového průměru je

$$var(\bar{y}) = \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{i\neq j} y_i y_j\right].$$

Protože platí vztah

$$\sum_{i=1}^{N}(y_i-\mu)^2 = \sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} y_i)^2}{N} = \frac{1}{N}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{i\neq j} y_i y_j\right],$$

rozptyl se zjednoduší do tvaru

$$var(\bar{y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N-1} = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}.$$

Pro prostý náhodný výběr s vracením se střední hodnota a rozptyl výběrového průměru získají pomocí obvyklých statistických vlastností výběrového průměru z nezávislých a stejně rozdělených náhodných veličin. Pravděpodobnost výběru i-té jednotky je $p_i = 1/N$. Pravděpodobnost, že i-tá jednotka bude zahrnuta do výběru (jednou nebo vícekrát) je $\pi_i = 1 - (1 - 1/N)^n$. Očekávaný počet zahrnutí i-té jednotky do výběru je n/N.

Příklad 1.
Na pozemku o výměře 100 m^2 jsou pěstovány brambory. Pozemek je rozdělen na 100 stejně velkých ploch o velikosti 1 m^2. Prostým náhodným výběrem bez vracení vybereme 10, z nichž budeme odhadovat průměrnou a celkovou ztrátu brambor při sklizni a dále vypočítáme směrodatné odchylky obou těchto odhadů. Ztráty (v kg) zjištěné na vybraných plochách jsou: 0.23, 0.08, 0.15, 0.55, 0.32, 0.02, 0.10, 0.18, 0.29 a 0.30.

Odhad průměrné ztráty při sklizni vypočteme pomocí vztahu (2):

$$\bar{y} = \frac{0.23 + 0.08 + \cdots + 0.30}{10} = 0.222 \text{ kg}$$

a odhad celkové ztráty podle vztahu (6):

$$\hat{\tau} = 100 \cdot 0.222 = 22.2 \text{ kg}.$$

Dále vypočteme odhady rozptylů obou těchto odhadů pomocí vztahů (5) a (7) a z nich směrodatné odchylky. Pro jejich výpočet ale nejdříve potřebujeme znát výběrový rozptyl (3),

$$s^2 = \frac{(0.23 - 0.222)^2 + (0.08 - 0.222)^2 + \cdots + (0.30 - 0.222)^2}{10 - 1} = 0.0234 \text{ kg},$$

$$\widehat{var}(\bar{y}) = \left(\frac{100-10}{100}\right)\frac{0.0234}{10} = 0.002106 \text{ kg}$$

a směrodatná odchylka je $\sqrt{0.002106} = 0.04589$ kg,

$$\widehat{var}(\hat{\tau}) = 100^2 \cdot 0.002106 = 21.06 \text{ kg},$$

a směrodatná odchylka tohoto odhadu je $\sqrt{21.06} = 4.589$ kg.

2.2 Konfidenční intervaly pro populační průměr a úhrn

Díky výběrovým metodám máme sice k dispozici odhady populačního průměru a úhrnu, což jsou odhady bodové, ale ty nám obvykle nestačí na to, abychom získali nějakou představu o přesnosti získaného odhadu. Proto se často konstruuje intervalový odhad, tzv. konfidenční interval (interval spolehlivosti), takovým způsobem, že z výběru předem daným způsobem vypočteme dvě výběrové funkce, které budou tvořit dolní a horní mez tohoto intervalu tak, aby skutečná populační hodnota (hodnota hledané populační charakteristiky) jím byla pokryta.

Konfidenční interval pro populační průměr μ si označíme I. Jak už jsme zmínili, interval spolehlivosti musí být stanoven tak, aby v sobě zahrnul skutečnou hodnotu populační charakteristiky, a to s danou pravděpodobností blízkou jedné. Proto zvolíme malé číslo $\alpha \in (0, 1)$, tzv. hladinu významnosti, která představuje pravděpodobnost, že interval nepokryje skutečnou hodnotu. Potom platí, že $P(\mu \in I) = 1-\alpha$. Výraz $1-\alpha$ je označován jako konfidenční koeficient (spolehlivost odhadu), a tedy interval nazýváme $(1-\alpha)100\%$ konfidenční interval. Krajní body intervalu se mohou měnit z výběru na výběr, zatímco parametr μ, i když je neznámý, je pevný. Hladina významnosti α se většinou stanovuje jako 0.01 nebo 0.05. Jestliže zvolíme $\alpha = 0.05$, potom dostaneme 95%-ní interval spolehlivosti, což nám říká, že interval pokryje skutečnou hodnotu μ pro 95% možných výběrů o rozsahu n.

Přibližný $(1-\alpha)100\%$ konfidenční interval pro populační průměr μ při prostém náhodném výběru bez vracení dostaneme ve tvaru

$$\bar{y} \pm t_{n-1,1-\frac{\alpha}{2}} \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{s^2}{n}\right)},$$

kde $t_{n-1,\alpha}$ je příslušný α-kvantil Studentova t-rozdělení o $n-1$ stupních volnosti.

Přibližný $(1-\alpha)100\%$ konfidenční interval pro populační úhrn τ je potom

$$\hat{\tau} \pm t_{n-1,1-\frac{\alpha}{2}} \sqrt{N(N-n)\frac{s^2}{n}}. \tag{8}$$

Rozdělení každé výběrové charakteristiky závisí na rozdělení celé populace (základního souboru). Se zvětšujícím se rozsahem výběru lze velmi dobře rozdělení výběrových charakteristik aproximovat normálním rozdělením. Proto pro rozsahy výběrů větší než 50 se užívá místo Studentova t-rozdělení α-kvantil normálního normovaného rozdělení.

Obecně platí, pokud je $\hat{\theta}$ normálně rozdělený, nestranný odhad parametru θ, potom konfidenční interval pro parametr θ je dán vztahem

$$\hat{\theta} \pm u_{1-\frac{\alpha}{2}} \sqrt{var(\hat{\theta})},$$

kde u_α označuje α-kvantil normálního normovaného rozdělení $N(0,1)$. Konvergence rozdělení výběrových charakteristik k normálnímu rozdělení vyplývá z limitních vět teorie pravděpodobnosti a to především z nejznámější, centrální limitní věty. Odhad tak může mít rozdělení, které se blíží normálnímu rozdělení, dokonce i když původní y-ové hodnoty toto rozdělení nemají. Protože je rozptyl odhadu parametru θ nejčastěji určen (odhadnut) z výběru, dostaneme pak přibližný interval

$$\hat{\theta} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\hat{\theta})}.$$

A jak už jsme se zmínili, pokud je rozsah výběru menší než 50, je vhodnější užít $(1-\alpha/2)$-kvantil Studentova t-rozdělení o $n-1$ stupních volnosti, který nám dává poněkud širší interval, než je tomu v případě příslušného kvantilu normálního normovaného rozdělení.

Příklad 2.
Zde navážeme na příklad 1 a stanovíme 95% konfidenční interval pro celkovou ztrátu brambor při sklizni.

Pro výpočet užijeme vztah (8). Hodnota Studentova t-rozdělení získaná z tabulek je $t_9(0.975) = 2.26$, potom

$$22.2 \pm 2.26\sqrt{21.06} = 22.2 \pm 10.37 = \langle 11.83 \text{ kg}, 32.57 \text{ kg} \rangle .$$

Získaný interval je poněkud širší vzhledem k variabilitě výběrových hodnot a malému rozsahu výběru.

2.3 Výběr s nestejnými pravděpodobnostmi

Zatímco u prostého náhodného výběru měly všechny jednotky stejnou pravděpodobnost zahrnutí do výběru, zde se budeme zabývat metodou, u které připustíme různé pravděpodobnosti zahrnutí jednotlivých jednotek do výběru. Tento druh náhodného výběru můžeme užít ve složitějších výběrech, a to zejména ve skupinovém nebo vícestupňovém výběru, kterými se budeme zabývat později. Přechod od jednodušší metody výběru ke složitější má smysl pouze tehdy, zvýší-li se tím eficience (vydatnost) odhadů, tedy bude-li rozptyl těchto odhadů značně menší.

Jednotky v populaci se často vyznačují nestejnou velikostí (rozlohou) a tím i nestejným významem, a proto může mít každá z nich různou pravděpodobnost zahrnutí do výběru. Velikost výběrové jednotky je totiž spjata s určitou vlastností (statistickým znakem), a proto výběrová jednotka může být velká z hlediska jednoho znaku a z hlediska ostatních znaků malá. Jelikož chceme získat co nejlepší odhady, můžeme pravděpodobnosti jednotkám přiřadit vědomě, a to podle jejich významu z hlediska zjišťovaného statistického znaku. Významnější jednotky pak mají vyšší pravděpodobnost zahrnutí do výběru.

Jestliže máme například studované území rozděleno na pozemky s nestejnou velikostí, mohlo by se zdát žádoucí, přiřadit větší pravděpodobnost zahrnutí

větším pozemkům. To můžeme udělat tak, že rovnoměrně vybereme body ze studované oblasti a zahrneme pozemek do výběru vždy, když do něj padne vybraný bod.

2.3.1 Výběr s vracením: Hansen–Hurwitzův odhad

Nejdříve se budeme zabývat teoreticky jednodušším modelem při výběrech s nestejnými pravděpodobnostmi, a to výběrem s vracením (jednotka se může ve výběru vyskytnout vícekrát). Protože vracíme po každém tahu jednotku zpět do základního souboru, jsou jednotlivé tahy vzájemně nezávislé náhodné pokusy. Pravděpodobnost výběru i-té jednotky z populace označíme p_i pro $i = 1, 2, \ldots, N$. Pak nestranný odhad populačního úhrnu τ (Hansen–Hurwitzův odhad) je

$$\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}, \tag{9}$$

rozptyl tohoto odhadu je

$$var\left(\hat{\tau}_p\right) = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - \tau\right)^2$$

a nestranný odhad tohoto rozptylu je

$$\widehat{var}\left(\hat{\tau}_p\right) = \frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{\tau}_p\right)^2. \tag{10}$$

Nestranný odhad populačního průměru μ můžeme pak psát ve tvaru $\hat{\mu}_p = (1/N)\hat{\tau}_p$, jeho rozptyl jako $var(\hat{\mu}_p) = (1/N^2)var(\hat{\tau}_p)$ a odhad tohoto rozptylu je $\widehat{var}(\hat{\mu}_p) = (1/N^2)\widehat{var}(\hat{\tau}_p)$.

Jelikož jsou odhady obou těchto charakteristik nestranné, tak můžeme říci, že rozptyl těchto odhadů je totožný se střední kvadratickou chybou.

Co se týče eficience odhadů, tak z pohledu na vzorce odhadů populačního úhrnu a průměru můžeme dojít k závěru, že pravděpodobnosti p_i by měly být co

nejvíce úměrné hodnotám y_i. Pokud by v ideálním případě byly pravděpodobnosti výběru p_i zcela úměrné hodnotám y_i, pak by podíl y_i/p_i byl konstantní a Hansen–Hurwitzův odhad by měl nulový rozptyl (odhad by byl zcela přesný). Rozptyl tak bude malý, jestliže pravděpodobnosti výběru budou přibližně úměrné hodnotám y_i. Samozřejmě ale hodnoty y_i nejsou známy před výběrem, jsou zjišťovány až po provedení výběru, a proto se nemohou použít k určení pravděpodobností. Snažíme se tak najít nějakou pomocnou proměnnou (jako jsou velikosti jednotek), která je přibližně úměrná hodnotám y_i a pomocí níž zvolíme pravděpodobnosti výběru.

Přibližný $(1-\alpha)100\%$ konfidenční interval pro populační úhrn je

$$\hat{\tau}_p \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\hat{\tau}_p)}, \tag{11}$$

kde u_α je příslušný α-kvantil normálního normovaného rozdělení. Při rozsahu výběru menším než 50 jednotek použijeme místo normálního normovaného příslušný α-kvantil Studentova t-rozdělení o $(n-1)$ stupních volnosti.

2.3.2 Horvitz–Thompsonův odhad

Tato metoda, ať už s vracením nebo bez vracení jednotek do základního souboru, pracuje s pravděpodobnostmi π_i zahrnutí i-té jednotky do výběru, pro $i = 1, 2, \ldots, N$, nikoli s pravděpodobnostmi vybrání této jednotky. Chceme tak získat kvalitnější odhady a vyhnout se podmíněným pravděpodobnostem vybrání v dalších tazích. Nestranný odhad populačního úhrnu τ (Horvitz–Thompsonův odhad) je

$$\hat{\tau}_\pi = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i},$$

kde ν je efektivní rozsah výběru, což je počet různých jednotek ve výběru. Tento odhad nezávisí na tom, kolikrát může být jednotka vybrána. Každá jednotka ve výběru je použita pouze jednou.

Pravděpodobnost, že budou obě jednotky i a j zahrnuty do výběru, značíme

π_{ij}. Rozptyl odhadu $\hat{\tau}_\pi$ je

$$var(\hat{\tau}_\pi) = \sum_{i=1}^{N} \left(\frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{N} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_i} \right) y_i y_j$$

a nestranný odhad tohoto rozptylu je

$$\begin{aligned} \widehat{var}(\hat{\tau}_\pi) &= \sum_{i=1}^{\nu} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^{\nu} \sum_{j \neq i} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_i} \right) \frac{y_i y_j}{\pi_{ij}} \\ &= \sum_{i=1}^{\nu} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i} \right) y_i^2 + 2 \sum_{i=1}^{\nu} \sum_{j>i} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right) y_i y_j, \end{aligned}$$

jestliže všechny společné pravděpodobnosti zahrnutí π_{ij} jsou větší než nula.

Nestranný odhad populačního průměru μ má tvar $\hat{\mu}_\pi = (1/N)\hat{\tau}_\pi$, jeho rozptyl je ve tvaru $var(\hat{\mu}_\pi) = (1/N^2) var(\hat{\tau}_\pi)$ a odhad tohoto rozptylu je $\widehat{var}(\hat{\mu}_\pi) = (1/N^2)\widehat{var}(\hat{\tau}_\pi)$.

Ze vzorců pro odhady vidíme, že k jejich výpočtu potřebujeme znát kromě hodnot y_i i pravděpodobnosti zahrnutí π_i. Avšak jak není lehké navrhnout systémy výběrů, které by získaly požadované nestejné pravděpodobnosti zahrnutí, tak není ani lehké vypočítat tyto pravděpodobnosti zahrnutí pro dané systémy výběrů. Obvykle se vypočítají z již daných pravděpodobností výběru založených na silné korelaci mezi pomocnou proměnnou a hodnotami y_i, jak tomu bylo u Hansen–Hurwitzova odhadu. Opět zde platí, jestliže pravděpodobnosti zahrnutí π_i budou přibližně úměrné hodnotám y_i, pak rozptyl Horvitz–Thomsonova odhadu bude malý.

Pro metody, ve kterých je efektivní rozsah výběru ν spíše pevný než náhodný, může být psán rozptyl ve tvaru

$$var(\hat{\tau}_\pi) = \sum_{i=1}^{N} \sum_{j<i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

a nestranný odhad tohoto rozptylu je pak dán vztahem

$$\widehat{var}(\hat{\tau}_\pi) = \sum_{i=1}^{\nu} \sum_{j<i} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

za předpokladu, že všechny společné pravděpodobnosti zahrnutí π_{ij} jsou větší než nula.

Přibližný $(1-\alpha)100\%$ konfidenční interval pro populační úhrn je

$$\hat{\tau}_\pi \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\hat{\tau}_\pi)},$$

kde u_α je příslušný α-kvantil normálního normovaného rozdělení. Při rozsahu výběru menší než 50 jednotek můžeme nahradit normální normované rozdělení Studentovým t-rozdělením o $(\nu - 1)$ stupních volnosti.

Ačkoliv je odhad rozptylu $\widehat{var}(\hat{\tau}_\pi)$ nestranný, jeho výpočet je poněkud zdlouhavý a s některými metodami může dávat i záporné odhady. Proto ho můžeme nahradit jiným odhadem, který sice není nestranný, ale lépe se s ním počítá a někdy má i menší rozptyl. Pro každou (různou) jednotku y_i, kde $i = 1, 2, \ldots, \nu$, určíme

$$t_i = \frac{\nu y_i}{\pi_i},$$

kde t_i pak představuje odhad populačního úhrnu a jejich aritmetický průměr je Horvitz–Thomsonův odhad. Výběrový rozptyl t_i je následně dán vztahem

$$s_t^2 = \frac{1}{\nu - 1} \sum_{i=1}^{\nu} (t_i - \hat{\tau}_\pi)^2$$

a alternativní odhad rozptylu je konečně

$$\widetilde{var}(\hat{\tau}_\pi) = \left(\frac{N - \nu}{N} \right) \frac{s_t^2}{\nu}.$$

Příklad 3.
Úkolem je odhadnout celkový počet uzavřených manželství v určitém okrese a vypočítat směrodatnou odchylku tohoto odhadu. Okres je složen ze 30 obcí, přičemž vybereme s vracením 6 obcí, a to s pravděpodobnostmi úměrnými počtu

jejich obyvatel vzhledem k celkovému počtu obyvatel v okrese, který činí 215 000 osob. Hodnoty získané ve vybraných obcích jsou:

Obec (i)	Počet obyvatel (x_i)	Počet manželství (y_i)	Vypočítané p-sti (p_i)
1	6 300	600	0.03
2	2 800	450	0.013
3	500	100	0.002
4	1 100	180	0.005
5	8 900	750	0.04
6	2 800	450	0.013

Pro výběr s vracením užijeme Hansen–Hurwitzův odhad (9),

$$\hat{\tau}_p = \frac{1}{6}\left(\frac{600}{0.03} + \frac{450}{0.013} + \cdots + \frac{450}{0.013}\right) = 32330,$$

pro výpočet jeho odhadu rozptylu vztah (10):

$$\widehat{var}(\hat{\tau}_p) = \frac{1}{6(6-1)}\left[\left(\frac{600}{0.03} - 32330\right)^2 + \cdots + \left(\frac{450}{0.013} - 32330\right)^2\right] = 22419518$$

a směrodatná odchylka tohoto odhadu je $\sqrt{22419518} = 4735$.
Dále chceme zjistit 95% konfidenční interval pro celkový počet uzavřených manželství v tomto okrese pomocí vztahu (11). Z tabulek zjistíme hodnotu Studentova rozdělení $t_5(0.975) = 2.57$, potom

$$32330 \pm 2.57\sqrt{22419518} = 32330 \pm 12169 = \langle 20161, 44499 \rangle .$$

2.4 Podílový odhad

Předpokladem pro získání podílového odhadu je znát navíc x-ové hodnoty pro celou populaci (představují hodnoty pomocné proměnné), y-ové jsou ty, co nás zajímají. Existuje mezi nimi lineární vztah v tom smyslu, že pokud x_i se blíží k nule, potom i y_i se blíží k nule. Například pokud se velikost území bude blížit nule, počet zvířat na nich žijících bude zaručeně nula.

Označíme $\tau_x = \sum_{i=1}^{N} x_i$ populační úhrn x-ových hodnot a $\mu_x = \tau_x/N$ populační průměr x-ových hodnot. Tito populační ukazatelé pro x-ové proměnné jsou známé. Předmětem našeho zájmu jsou odhady populačního průměru μ a úhrnu τ pro y-ové hodnoty.

Pro prostý náhodný výběr n jednotek je výběr y-ových hodnot zaznamenán spolu s odpovídajícími x-ovými hodnotami. Populační podíl R je definován vztahem

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\tau_y}{\tau_x}$$

a výběrový podíl r je

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}. \tag{12}$$

Podílový odhad populačního průměru μ je potom

$$\hat{\mu}_r = r\mu_x. \tag{13}$$

Protože podílový odhad není nestranný (jedná se o podíl dvou náhodných veličin), pro srovnání jeho eficience vůči ostatním odhadům užijeme střední kvadratickou chybu odhadu. Střední kvadratická chyba podílového odhadu je dána vztahem $mse(\hat{\mu}_r) = E(\hat{\mu}_r - \mu)^2$. Pro nestranný odhad se střední kvadratická chyba rovná rozptylu, ale pro odhad, který není nestranný, se střední kvadratická chyba rovná rozptylu plus čtverci biasu: $mse(\hat{\mu}_r) = var(\hat{\mu}_r) + [E(\hat{\mu}_r) - \mu]^2$. Pro podílový odhad je čtvercový bias vzhledem k rozptylu malý, takže vztah pro aproximaci střední kvadratické chyby je stejný jako u rozptylu:

$$var(\hat{\mu}_r) \approx \left(\frac{N-n}{N}\right)\frac{\sigma_r^2}{n},$$

kde

$$\sigma_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2.$$

Podílový odhad má sklon k tomu být přesnější než výběrový průměr y-ových hodnot pro populace, pro které je σ_r^2 menší než σ^2. Což je případ populací, ve kterých jsou x-ové a y-ové hodnoty silně korelované. Podílový odhad je tedy zatížen malou chybou.

Odhad střední kvadratické chyby nebo rozptylu podílového odhadu (hodnoty populační nahradíme hodnotami výběrovými) je

$$\widehat{var}(\hat{\mu}_r) = \left(\frac{N-n}{N}\right)\frac{s_r^2}{n}, \tag{14}$$

kde

$$s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2. \tag{15}$$

Ale tento odhad má tendenci produkovat velké hodnoty pro výběry mající velké hodnoty $\bar{x}$ a malé hodnoty pro výběry mající malé hodnoty $\bar{x}$. Proto byl navrhnut upravený odhad

$$\widetilde{var}(\hat{\mu}_r) = \left(\frac{\mu_x}{\bar{x}}\right)^2 \widehat{var}(\hat{\mu}_r). \tag{16}$$

Přibližný $100(1-\alpha)\%$ konfidenční interval pro populační průměr μ, založen na aproximaci normálním rozdělením, je ve tvaru

$$\hat{\mu}_r \pm t_{n-1,1-\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\mu}_r)},$$

kde $t_{n-1,1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$-kvantil Studentova t-rozdělení o $n-1$ stupních volnosti. Odhad rozptylu pod odmocninou můžeme nahradit upraveným odhadem $\widetilde{var}(\hat{\mu}_r)$.

Podílový odhad populačního úhrnu τ je dán vztahem

$$\hat{\tau}_r = N\hat{\mu}_r = r\tau_x.$$

Rozptyl tohoto odhadu získáme, pokud rozptyl odhadu populačního průměru $var(\hat{\mu}_r)$ vynásobíme hodnotou N^2.

Pro odhadování populačního podílu R užíváme výběrový podíl r. I když není nestranný, lze o jeho nestrannosti hovořit v případě velký rozsahů výběru. Rozptyl aproximace je

$$var(r) \approx \left(\frac{N-n}{N\mu_x^2}\right)\frac{\sigma_r^2}{n},$$

odhad tohoto rozptylu je potom

$$\widehat{var}(r) = \left(\frac{N-n}{N\mu_x^2}\right)\frac{s_r^2}{n}$$

nebo jeho upravená varianta

$$\widetilde{var}(r) = \left(\frac{N-n}{N\bar{x}^2}\right)\frac{s_r^2}{n}.$$

Příklad 4.

Prostým náhodným výběrem byly vybrány 4 domácnosti z celkového počtu 14 500 domácností ve městě se 42 050 obyvateli. Budeme odhadovat průměrné výdaje domácností na jídlo za týden. U vybraných domácností byly zjištěny tyto údaje:

Domácnosti (i)	Počet osob v domácnosti (x_i)	výdaje/týden v Kč (y_i)
1	4	1 500
2	3	1 200
3	2	900
4	4	2 000

Zde srovnáváme užití dvou metod výpočtu. Nejdříve budeme počítat nestranný odhad, který nevyužívá pomocnou informaci o populaci. Průměrné výdaje odhadneme pomocí vztahu (2) a rozptyl tohoto odhadu podle vztahu (5):

$$\bar{y} = \frac{1500 + 1200 + 900 + 2000}{4} = 1\,400 \text{ Kč}.$$

Pro výpočet rozptylu je třeba znát ještě výběrový rozptyl (3),

$$s^2 = \frac{(1500 - 1400)^2 + \cdots + (2000 - 1400)^2}{4 - 1} = 220\,000 \text{ Kč},$$

$$\widehat{var}(\bar{y}) = \left(\frac{14500-4}{14500}\right)\frac{220000}{4} = 54\,985 \text{ Kč}.$$

Nyní užijeme k odhadování průměrných výdajů vztah (13). Tento odhad sice není nestranný, ale využívá pomocnou informaci o populaci. Nejdříve musíme vypočítat populační úhrn x-ových hodnot μ_x a výběrový podíl (12):

$$\mu_x = \frac{\tau_x}{N} = \frac{42050}{14500} = 2.9,$$

$$r = \frac{1500+1200+900+2000}{4+3+2+4} = 430.8,$$

$$\hat{\mu}_r = 430.8 \cdot 2.9 = 1\,249.32 \text{ Kč}.$$

Rozptyl tohoto odhadu získáme pomocí vztahu (14), ale nejdříve musíme znát výběrový rozptyl (15),

$$s_r^2 = \frac{(1500-430.8\cdot 4)^2 + \cdots + (2000-430.8\cdot 4)^2}{4-1} = 45\,483 \text{ Kč},$$

potom

$$\widehat{var}(\hat{\mu}_r) = \left(\frac{14500-4}{14500}\right)\frac{45483}{4} = 11\,368 \text{ Kč}$$

nebo můžeme vypočítat i upravený odhad rozptylu (16):

$$\widetilde{var}(\hat{\mu}_r) = \left(\frac{2.9}{3.25}\right)^2 \cdot 11368 = 9\,051 \text{ Kč}.$$

Vidíme, že v druhém případě (při užití podílového odhadu) dostaneme odhady rozptylu menší a tudíž i lepší, i když nejsou nestranné.

3 Složitější metody výběru

V této kapitole jsem nejvíce čerpala z literatury [3], [7], a dále také z literatury [4], [6], [8] a [9].

Výběr jednotek z populace můžeme provést buď přímo, aniž by nám bránila jakákoli omezení, což jsme využívali v kapitole 2, a nebo tak, že si populaci rozdělíme na větší či menší části, ze kterých pak jednotky vybíráme. Zatímco v předchozí kapitole mohla být do výběru zahrnuta jakákoli kombinace jednotek, u složitějších výběrů tomu tak není, některé kombinace vzniknout nemohou. Jde přitom o to, aby odhady ze vzniklých výběrů silně nenadhodnocovaly nebo nepodhodnocovaly skutečnost. Z toho plyne, že požadujeme, aby vznikly takové výběry, které povedou k malým výběrovým chybám (rozptylům odhadů).

V dalších podkapitolách si uvedeme některé z těchto složitějších metod a to stratifikovaný výběr, skupinový a systematický výběr a nakonec vícestupňový výběr. U těchto metod po rozdělení populace na několik dílčích subpopulací (podsouborů) buď prošetřujeme všechny subpopulace nebo jen náhodně vybrané. K prošetření můžeme využívat výběr se stejnými i nestejnými pravděpodobnostmi, popřípadě i jejich kombinaci. Budeme opět odhadovat populační průměry a úhrny.

3.1 Stratifikovaný výběr

Hlavní podstatou stratifikovaného výběru je rozdělit populaci do několika vrstev (podsouborů), které nazýváme strata nebo také oblasti. Výběry jednotek jsou pak dělány samostatně z každé z těchto vrstev pomocí zvolené výběrové metody. Jednotlivé oblasti se nepřekrývají, což znamená, že výběry v jednotlivých vrstvách jsou nezávislé, a proto rozptyly odhadů jednotlivých vrstev mohou být sečteny, abychom získali rozptyl odhadu pro celou populaci. Principem přitom je, aby jednotky uvnitř každé vrstvy si byly co nejvíce podobné. Tím docílíme, že odhad populačního průměru a úhrnu bude přesnější. Naopak vrstvy navzájem mohou být naprosto odlišné.

Důvodů pro stratifikaci může být hned několik. Například rozdělení geografické oblasti na podoblasti na základě nějaké známé hodnoty, kterou může být půdní typ nebo nadmořská výška. Ačkoli by se mohlo často zdát, že studované území je stejnorodé, rozdělení do bloků pomůže zajistit, že přesnost odhadu se v důsledku snížení variability zvýší, protože se v podoblastech většinou vyskytují podobné jednotky z hlediska zvoleného statistického znaku a protože provedeme v každé podoblasti samostatný výběr. Dále stratifikovaného výběru užíváme, pokud známe odhad průměru (úhrnu) pro celou populaci a chceme znát dílčí odhady průměrů (úhrnů) jen některých oblastí, ze kterých se populace skládá, např. při politickém členění na okresy, nebo lidská populace může být rozvrstvena na základě pohlaví, věku, socioekonomických faktorů, atd. Jak vidíme, je ovšem vždy nutné mít nějakou doplňující informaci o populaci, abychom byli schopni jednotlivé jednotky zařadit do vrstev ještě předtím, než provedeme výběr.

Tedy populaci rozdělíme do L vrstev a v každé vrstvě provedeme výběr pomocí zvolené výběrové metody. Jak už jsme řekli, výběry z jednotlivých vrstev jsou nezávislé. Potom y_{hi} je hodnota statistického znaku pro i-tou jednotku v h-té vrstvě. Celkový počet jednotek v h-té vrstvě značíme N_h a počet jednotek ve výběru v této vrstvě je n_h. Celkový počet jednotek v celé populaci je potom $N = \sum_{h=1}^{L} N_h$ a celkový rozsah výběru $n = \sum_{h=1}^{L} n_h$. Úhrn y-ových hodnot ve vrstvě h je $\tau_h = \sum_{i=1}^{N_h} y_{hi}$ a průměr pro tuto vrstvu je $\mu_h = \tau_h/N_h$. Populační úhrn je tedy $\tau = \sum_{h=1}^{L} \tau_h$ a populační průměr $\mu = \tau/N$. Jestliže uvnitř každé vrstvy použijeme k prošetření prostý náhodný výběr, pak tuto metodu budeme nazývat *stratifikovaný náhodný výběr*.

3.1.1 Odhadování populačního úhrnu a průměru s využitím libovolné výběrové metody

Předpokládáme, že uvnitř každé vrstvy h došlo k pořízení výběru s_h o n_h jednotkách s využitím libovolné výběrové metody, a dále pak, že každá vrstva h má odhad $\hat{\tau}_h$, který je nestranným odhadem populačního úhrnu v h-té vrstvě τ_h s ohledem na zvolenou metodu. Potom rozptyl odhadu $\hat{\tau}_h$ označíme $var(\hat{\tau}_h)$ a

jeho nestranný odhad $\widehat{var}(\hat{\tau}_h)$.

Pak nestranný odhad celkového populačního úhrnu τ získáme jako součet odhadů $\hat{\tau}_h$ přes všechny vrstvy,

$$\hat{\tau}_{st} = \sum_{h=1}^{L} \hat{\tau}_h.$$

Díky nezávislosti výběrů v jednotlivých vrstvách je rozptyl stratifikovaného odhadu roven součtu rozptylů odhadů v jednotlivých vrstvách,

$$var(\hat{\tau}_{st}) = \sum_{h=1}^{L} var(\hat{\tau}_h),$$

a nestranný odhad tohoto rozptylu je analogicky součtem všech odhadů rozptylu přes všechny vrstvy,

$$\widehat{var}(\hat{\tau}_{st}) = \sum_{h=1}^{L} \widehat{var}(\hat{\tau}_h).$$

Jelikož platí $\mu = \tau/N$, pro stratifikovaný odhad populačního průměru μ obdržíme

$$\hat{\mu}_{st} = \frac{\hat{\tau}_{st}}{N}.$$

Za předpokladu, že výběry v jednotlivých vrstvách jsou nezávislé, rozptyl odhadu můžeme psát ve tvaru

$$var(\hat{\mu}_{st}) = \frac{1}{N^2} var(\hat{\tau}_{st})$$

a jeho nestranný odhad je

$$\widehat{var}(\hat{\mu}_{st}) = \frac{1}{N^2} \widehat{var}(\hat{\tau}_{st}).$$

3.1.2 Odhadování populačního úhrnu a průměru s využitím stratifikovaného náhodného výběru

Zde předpokládáme, že v každé vrstvě se provádí nezávislý prostý náhodný výběr bez vracení. Teorii tohoto výběru již známe a tedy platí, že

$$\hat{\tau}_h = N_h \bar{y}_h$$

je nestranným odhadem τ_h, kde

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \tag{17}$$

je výběrový průměr pro h-tou vrstvu.

Nestranný odhad pro populační úhrn τ je

$$\hat{\tau}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$$

a má rozptyl ve tvaru

$$var(\hat{\tau}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{\sigma_h^2}{n_h},$$

kde

$$\sigma_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2$$

je konečný populační rozptyl h-té vrstvy.

Nestranný odhad tohoto rozptylu odhadu $\hat{\tau}_{st}$ je ve tvaru

$$\widehat{var}(\hat{\tau}_{st}) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h},$$

kde

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

je výběrový rozptyl pro h-tou vrstvu.

S využitím stratifikovaného náhodného výběru se nestranný odhad populačního průměru μ nazývá stratifikovaný výběrový průměr,

$$\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h, \tag{18}$$

a jeho rozptyl je ve tvaru

$$var(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\sigma_h^2}{n_h}. \tag{19}$$

Nestranný odhad tohoto rozptylu je

$$\widehat{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}. \tag{20}$$

3.1.3 Konfidenční intervaly

Jestliže jsou rozsahy výběrů ve všech vrstvách dostatečně velké, potom přibližný $100(1-\alpha)\%$ konfidenční interval pro populační úhrn je stanovený takto,

$$\hat{\tau}_{st} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\hat{\tau}_{st})},$$

kde u_α je příslušný α-kvantil normálního normovaného rozdělení. Pro populační průměr je konfidenční interval dán vztahem

$$\bar{y}_{st} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\bar{y}_{st})}.$$

Jak již víme z kapitoly 2.2, tak normální normované rozdělení užijeme, pokud je rozsah výběru nejméně 50. Jestliže je rozsah výběru menší, aplikujeme Studentovo t-rozdělení o d stupních volnosti, přičemž stupně volnosti získáme výpočtem z tzv. Satterthwaitovy formule:

$$d = \left(\sum_{h=1}^{L} a_h s_h^2 \right)^2 \Bigg/ \left[\sum_{h=1}^{L} (a_h s_h^2)^2 / (n_h - 1) \right],$$

kde $a_h = N_h(N_h - n_h)/n_h$. Jestliže velikosti všech vrstev N_h jsou stejné, všechny rozsahy výběrů n_h jsou stejné a všechny výběrové rozptyly s_h^2 jsou stejné, potom počet stupňů volnosti je $(n - L)$. Satterthwaitova formule je založena na aproximaci rozdělení součtu výběrových rozptylů majících χ^2-rozdělení.

3.1.4 Náklady a rozvržení výběru do vrstev

Řešíme zde otázku, jak rozdělit celkový rozsah výběru n do jednotlivých vrstev, jestliže to není předem stanoveno. Jediná informace, která je v praxi většinou dána, je pouze celkový počet jednotek ve výběru n. Nejjednodušším případem je předpokládat, že máme populaci složenou z N jednotek a každá vrstva má stejný počet jednotek N_h a tedy i stejný rozsah výběru. Potom rozsah výběru pro h-tou vstvu je

$$n_h = \frac{n}{L}.$$

Jestliže jsou ale vrstvy odlišné velikosti, využijeme tzv. *proporcionální rozvržení*, jehož podstatou je, že rozsahy výběru n_h jsou úměrné velikostem N_h pro $h = 1, 2, \ldots, L$. Pokud tedy vrstva h má N_h jednotek, potom rozsah výběru v této vrstvě bude

$$n_h = \frac{nN_h}{N}.$$

Výhodou tohoto rozvržení je, že dává eficientní odhady populačního průměru (úhrnu), jestliže jsou všechny konečné populační rozptyly σ_h^2 stejné. Proporcionální rozvržení také zjednodušuje vzorce pro odhady i jejich rozptyly.

Schéma, které odhaduje populační průměr a úhrn s minimálním rozptylem pro daný celkový rozsah výběru n pomocí stratifikovaného náhodného výběru, nazýváme *optimální rozvržení*,

$$n_h = \frac{nN_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h}, \tag{21}$$

kde za směrodatné odchylky σ_h pro jednotlivé vrstvy dosazujeme hodnoty známé z dřívějších měření (které jsou v tomto smyslu již konstantami, nikoli odhady).

Jinými slovy je to takové rozvržení výběru, které dává přednost vrstvám, v nichž zkoumaný znak vykazuje značnou variabilitu. Cílem tohoto rozvržení je snížit rozptyl odhadu oproti proporcionálnímu výběru, tj. dosáhnout větší přesnosti odhadu.

V některých výběrových situacích jsou náklady na výběr, prošetření a zpracování výběrové jednotky odlišné v každé vrstvě, ať už se to týká peněz nebo času. Proto lze celkové náklady rozdělit na dílčí a popsat lineárním vztahem

$$c = c_0 + c_1 n_1 + c_2 n_2 + \cdots + c_L n_L,$$

kde c jsou celkové náklady na šetření, c_0 jsou pevné režijní náklady (nezávislé na výsledném rozvržení) a c_h jsou náklady na pozorovanou jednotku v h-té vrstvě. Potom pro celkové náklady c je nejnižšího rozptylu dosaženo pomocí toho, že rozsah výběru v h-té vrstvě je přímo úměrný velikosti vrstvy, variabilitě zkoumaného znaku ve vrstvě a nepřímo uměrný odmocnině z jednotkových nákladů, tedy hodnotě výrazu $N_h\sigma_h/\sqrt{c_h}$. Pak můžeme psát

$$n_h = \frac{(c - c_0)N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{L} N_h\sigma_h\sqrt{c_h}}.$$

Tedy optimální rozvržení znamená, čím větší a různorodější jednotlivé vrstvy chceme, tím větší bude rozsah výběru v nich, ale větší náklady na prošetření budou naopak rozsah výběru v dané vrstvě snižovat.

3.1.5 Poststratifikace

V některých situacích je lepší nejdříve pořídit výběr, i když byl získán např. prostým náhodným výběrem místo stratifikovaného výběru, a poté klasifikovat vybrané jednotky do vrstev a užít stratifikované odhady. Například lidská populace může být po pořízení výběru prostým náhodným výběrem rozvrstvena dle pohlaví. Rozdíl oproti klasickému stratifikovanému výběru je, že u postratifikace jsou rozsahy výběrů v jednotlivých vrstvách $n_1, n_2, \ldots, n_L$ náhodné veličiny.

Například s využitím proporcionálního rozvržení ve stratifikovaném náhodném výběru je rozsah výběru pro h-tou vrstvu pevný, $n_h = nN_h/N$, a rozptyl

odhadu (19) se zjednodušší na tvar

$$var(\bar{y}_{st}) = \frac{N-n}{nN} \sum_{h=1}^{L} \left(\frac{N_h}{N} \right) \sigma_h^2.$$

Při poststratifikaci je vybráno n jednotek prostým náhodným výběrem z celé populace, rozsah výběru n_h v h-té vrstvě má očekávanou hodnotu nN_h/N a tak výsledný výběr směřuje k přibližnému proporcionálnímu rozvržení. Při poststratifikaci je rozptyl stratifikovaného odhadu $\bar{y}_{st}$ (18) přibližně

$$var(\bar{y}_{st}) \approx \frac{N-n}{nN} \sum_{h=1}^{L} \left(\frac{N_h}{N} \right) \sigma_h^2 + \frac{1}{n^2} \left(\frac{N-n}{N-1} \right) \sum_{h=1}^{L} \left(\frac{N-N_h}{N} \right) \sigma_h^2 \qquad (22)$$

a rozptyl odhadu $\hat{\tau}_{st} = N\bar{y}_{st}$ je $var(\hat{\tau}_{st}) = N^2 var(\bar{y}_{st})$. První část výrazu (22) je rozptyl, který získáme užitím stratifikovaného prostého výběru s proporcionálním rozvržením, a druhá část (22) je při poststratifikaci do výrazu přidána kvůli náhodným rozsahům výběru ve vrstvách.

Pro rozptyl odhadu, pomocí něhož konstruujeme konfidenční intervaly pro populační průměr s poststratifikovanými daty získanými prostým náhodným výběrem, se doporučuje užít standardní stratifikovanou metodu (20) spíše než rozptyl odhadu z rovnice (22). Při poststratifikaci určuje standardní vztah (20) podmíněný rozptyl odhadu $\bar{y}_{st}$ (19) za podmínky daných (pevných) rozsahů výběrů $n_1, n_2, \ldots, n_L$. Zatímco vztah (22) je nepodmíněný rozptyl odhadu $\bar{y}_{st}$, protože zde již zahrnujeme náhodnost $n_1, n_2, \ldots, n_L$.

Příklad 5.

V době sklizně pšenice chceme získat přibližný průměr výnosu na m^2. Pšenice je pěstovaná na 4 polích s různou kvalitou půdy o celkové rozloze 60 000 m^2. Odebereme 30 vzorků (v gramech) zralé pšenice z ploch velkých 1 m^2. Vzorky rozdělíme mezi jednotlivá pole dle proporcionálního rozvržení. Pro výpočty máme k dispozici tyto data:

Pole (h)	Rozloha (N_h)	Průměr. hmotnost vzorku ($\bar{y}_h$)	Výběr. rozptyly (s_h^2)
1	20 000 m^2	230 g	300
2	18 000 m^2	200 g	500
3	10 000 m^2	150 g	200
4	12 000 m^2	140 g	400

Nejprve určíme rozsah výběru v jednotlivých vrstvách (počet vzorků) pomocí proporcionálního rozvržení (21):

$$n_h = \frac{nN_h}{N},$$

tedy $n_1 = 10, n_2 = 9, n_3 = 5, n_4 = 6$.

Potom odhadneme průměrný výnos pšenice na všech polích s užitím vztahu (18):

$$\bar{y}_{st} = \frac{20000 \cdot 230 + \cdots + 12000 \cdot 140}{60000} = 189.7 \text{ g/m}^2$$

a směrodatnou odchylku tohoto odhadu pomocí vztahu pro rozptyl (20):

$$\begin{aligned}\widehat{var}(\bar{y}_{st}) &= \left(\frac{20000}{60000}\right)^2 \left(\frac{20000-10}{20000}\right) \frac{300}{10} + \cdots + \left(\frac{12000}{60000}\right)^2 \left(\frac{12000-6}{12000}\right) \frac{400}{6} \\ &= 12.08 \text{ g/m}^2.\end{aligned}$$

Směrodatná odchylka je tedy $\sqrt{12.08} = 3.48$ g/m^2.

3.2 Skupinový a systematický výběr

Ačkoli by se mohlo zdát, že tyto dvě metody výběru jsou naprosto odlišné, protože jedna shlukuje jednotky ve výběru dohromady a druhá je od sebe odděluje, přesto ale mají stejnou strukturu. Podstatou těchto výběrů je, že populaci dělíme na primární jednotky a každá primární jednotka sestává ještě z několika sekundárních jednotek. U těchto výběrů je důležité si uvědomit, že jakmile bude jakákoli sekundární jednotka z primární jednotky zahrnuta do výběru, potom všechny sekundární jednotky v této primární jednotce budou zahrnuty do výběru a prošetřeny. I kdyby šetření probíhalo na sekundární jednotce, jsou to

primární jednotky, které jsou vybírány. Tedy v principu se můžeme obejít bez pojmu sekundární jednotky, pokud na primární jednotky budeme pohlížet jako na vybrané jednotky a budeme prošetřovat všechny jejich y-ové hodnoty. Potom všechny vlastnosti odhadů můžeme dostat na základě metody, pomocí níž byl výběr primárních jednotek získán.

U systematického výběru není neobvyklé mít rozsah výběru 1, potom mluvíme o jediné primární jednotce. Tento systematický výběr s jednou startovací jednotkou využívá hodně průzkumů, je totiž jednodušší z hlediska provedení a z hlediska nákladů levnější. Princip je, že jednotky v populaci seřadíme zcela náhodně (nezávislé na předmětu šetření). Potom stanovíme krok, kterým budeme vybírat jednotky (sekundární) ve stejné vzdálenosti od sebe. Náhodně také vybereme startovací jednotku, u které systematický výběr započne a která bude také do výběru zahrnuta. Například výběr každého čtvrtého bytu v panelovém domě se startovací jednotkou bytem č. 2. Z výběru o rozsahu 1 je možné získat nestranné odhady populačního průměru nebo úhrnu, ale není možné získat nestranný odhad jejich rozptylu. Užití systematičnosti navíc vede k větším rozptylům, než kdybychom užili prostý náhodný výběr.

U skupinového výběru může být celá populace rozdělena do menších či větších skupin (primární jednotky) a vybíráme tedy celé skupiny nikoli jednotlivé prvky, jak již bylo řečeno. Sekundární jednotky uvnitř primární jednotky jsou obvykle v těsné blízkosti mezi sebou a v prostorovém uspořádání se jeví jako dlouhé a úzké pásy navzájem sousedních jednotek. Tohoto uspořádání bude zejména využito u populací s velkým počtem jednotek (tisíce až statisíce). Skupinami jednotek mohou být např. rodiny, školy nebo plochy půdy atd.

Důvodů, proč zvolit skupinový výběr, může být několik, např. kvůli finančním a časovým úsporám. Pokud prošetřujeme na rozlehlém území, kde je rozptýlenost velkého počtu jednotek (např. osob), pak časově a finančně úspornější bude soustředit tyto jednotky do menších skupin (např. obcí) a prošetřit tak celou vybranou skupinu než každou vybranou jednotku zvlášť.

Velikost skupiny může sloužit jako pomocná informace, která může být využita buď při výběru skupin s nestejnými pravděpodobnostmi nebo při vytváření podílového odhadu. Velikost a tvar skupin může ovlivnit eficienci (vydatnost) odhadu. Protože je každá sekundární jednotka uvnitř primární jednotky pozorována, čili variabilita uvnitř primární jednotky nemá vliv na rozptyly odhadů, pak základním principem je získat co nejmenší rozptyl odhadů populačního průměru a úhrnu (co nejpřesnější odhady) tak, že buď populaci rozdělíme do skupin, které si jsou co nejvíce podobné a jednotky uvnitř každé skupiny jsou co nejvíce odlišné, a nebo se snažíme rozdělit populaci do přibližně velkých skupin.

Počet primárních jednotek v populaci označíme N a počet primárních jednotek ve výběru označíme n. M_i je počet sekundárních jednotek v i-té primární jednotce. Celkový počet sekundárních jednotek v populaci pak bude $M = \sum_{i=1}^{N} M_i$. Potom y_{ij} je hodnota j-té sekundární jednotky v i-té primární jednotce, která nás zajímá. Úhrn y-ových hodnot v i-té primární jednotce vypočítáme jako $y_i = \sum_{j=1}^{M_i} y_{ij}$. Populační úhrn získáme jako $\tau = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} y_i$. Populační průměr pro primární jednotky je $\mu_1 = \tau/N$ a pro sekundární jednotky $\mu = \tau/M$.

3.2.1 Primární jednotky vybrané prostým náhodným výběrem

Nestranný odhad

Pokud jsou primární jednotky vybrány prostým náhodným výběrem bez vracení, je nestranný odhad populačního úhrnu roven

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\bar{y}, \tag{23}$$

kde $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$ je výběrový průměr úhrnů primárních jednotek.

Rozptyl tohoto odhadu je

$$var(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n},$$

kde σ_u^2 je konečný populační rozptyl úhrnů primárních jednotek y_i,

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_1)^2.$$

Nestranný odhad rozptylu $\hat{\tau}$ je

$$\widehat{var}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n}, \tag{24}$$

kde s_u^2 je výběrový rozptyl úhrnu primárních jednotek y_i,

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{25}$$

Tyto výpočty jsou známé z teorie prostého náhodného výběru, proto odhady populačního průměru obdržíme analogicky. Nestranný odhad populačního průměru primárních jednotek μ_1 je $\bar{y} = \hat{\tau}/N$ a nestranný odhad populačního průměru sekundárních jednotek μ je $\hat{\mu} = \hat{\tau}/M$. Rozptyl $\bar{y}$ je $var(\bar{y}) = (1/N^2)var(\hat{\tau})$ a rozptyl $\hat{\mu}$ je $var(\hat{\mu}) = (1/M^2)var(\hat{\tau})$. Odhady rozptylů získáme obdobně, a to podělením odhadu rozptylu $\hat{\tau}$ konstantou N^2 nebo M^2.

Podílový odhad

Podílový odhad založený na velikosti skupiny nemusí být nestranný, zvlášť jestliže existuje silná korelace mezi úhrnem primárních jednotek y_i a velikostí té i-té primární jednotky M_i. Podílový odhad populačního úhrnu je

$$\hat{\tau}_r = rM,$$

kde r je výběrový podíl

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}.$$

Podílový odhad $\hat{\tau}_r$ sice není nestranným odhadem, ale jeho směrodatná odchylka bývá pro velký rozsah výběru relativně malá a tudíž střední kvadratická

chyba může být menší než pro nestranný odhad, pokud je vztah mezi y_i a M_i daný přímou úměrností.

Přibližný vztah pro střední kvadratickou chybu nebo také pro rozptyl podílového odhadu je

$$var(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (y_i - M_i\mu)^2. \tag{26}$$

Odhad tohoto rozptylu je pak dán vztahem

$$\widehat{var}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (y_i - rM_i)^2$$

nebo pomocí upraveného odhadu

$$\widetilde{var}(\hat{\tau}_r) = \left(\frac{nM}{N \sum_{i=1}^{n} M_i} \right)^2 \widehat{var}(\hat{\tau}_r).$$

Pro odhadování populačního průměru primárních jednotek μ_1 bude mít podílový odhad tvar $\hat{\mu}_{1r} = \hat{\tau}_r/N$ a rozptyl tohoto odhadu získáme vydělením vztahu (26) konstantou N^2. K odhadování populačního průměru μ pro sekundární jednotky bude podílový odhad ve tvaru $\hat{\mu}_r = \hat{\tau}_r/M = r$, pro který se pak rozptyl vyjádří podělením vztahu (26) konstantou M^2.

3.2.2 Primární jednotky vybrané pomocí výběrů s nestejnými pravděpodobnostmi

Předpokládáme zde, že primární jednotky jsou vybrány pomocí výběrových pravděpodobností úměrných k rozsahu primárních jednotek (skupin), tj. $p_i = M_i/M$. Samozřejmě podmínkou je znalost rozsahu všech skupin ještě před prováděním samotného výběru. Předností výběrů s nestejnými pravděpodobnostmi je, že dávají nestranné odhady, které se snadněji počítají, a že odstraňují nepříznivý vliv nestejné velikosti skupin. Eficience bude obecně vyšší než u předchozích odhadů (konkrétně bude tím větší, čím méně se budou průměry nebo úhrny skupin lišit).

Možným způsobem, jak provést tuto metodu, je vybrat n sekundárních jednotek z jejich celkového počtu M v populaci užitím prostého náhodného výběru s vracením, což znamená, že primární jednotka bude vybrána pokaždé, kdy jsou vybrány její sekundární jednotky.

Hansen–Hurwitzův odhad

Nestranný odhad populačního úhrnu s užitím pravděpodobností úměrných k rozsahu (velikosti) primárních jednotek při výběru s vracením, založený na Hansen–Hurwitzově odhadu, je roven

$$\hat{\tau}_p = \frac{M}{n} \sum_{i=1}^{n} \frac{y_i}{M_i}.$$

Přitom každé pozorování je zahrnuto do součtu tolikrát, kolikrát byla jeho primární jednotka vybrána. Rozptyl tohoto odhadu je dán jako

$$var(\hat{\tau}_p) = \frac{M}{n} \sum_{i=1}^{N} M_i(\bar{y}_i - \mu)^2,$$

kde $\bar{y}_i = y_i/M_i$. Nakonec nestranný odhad tohoto rozptylu je

$$\widehat{var}(\hat{\tau}_p) = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \mu_p)^2,$$

pro $\hat{\mu}_p = \hat{\tau}_p/M$.

Odhad $\hat{\mu}_p$ je nestranným pro populační průměr sekundárních jednotek μ při užití pravděpodobností úměrných k rozsahu, zatímco $\hat{\mu}_{1p} = \hat{\tau}_p/N$ je nestranným odhadem pro populační průměr primárních jednotek. Vztahy pro rozptyly těchto odhadů získáme opět tak, že rozptyl pro $\hat{\tau}_p$ podělíme konstantami M^2 nebo N^2.

Horvitz–Thompsonův odhad

Horvitz–Thompsonův odhad může být počítán pro tuto metodu (při výběru s vracením) s užitím pravděpodobností zahrnutí primární jednotky do výběru,

$$\pi_i = 1 - (1 - p_i)^n,$$

a pravděpodobností společného zahrnutí jednotek i a j,

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n],$$

založených na pravděpodobnostech výběru i-té jednotky $p_i = M_i/M$.

Horvitz–Thompsonův odhad pro populační úhrn je potom

$$\hat{\tau}_\pi = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i},$$

kde ν je počet odlišných primárních jednotek ve výběru. Výpočet rozptylu pro tento odhad je uveden v kapitole 2.3.2.

Příklad 6.
Zkoumanou populaci 100 studentů (sekundární jednotky) rozdělíme do 10 skupin (primární jednotky) podle škol, na kterých studují. Prostým náhodným výběrem vybereme 3 skupiny studentů. Bude nás zajímat odhad celkového počtu aktivních SIM karet v této populaci a jeho směrodatná odchylka, přičemž zjištěné hodnoty ve vybraných skupinách jsou: $M_1 = 10$ a $y_1 = 12$, $M_2 = 10$ a $y_2 = 9$ a $M_3 = 10$ a $y_3 = 15$.

Celkový počet SIM karet ve zkoumané populaci odhadneme pomocí vztahu (23):

$$\hat{\tau} = \frac{10}{3}(12 + 9 + 15) = 120.$$

Výběrový rozptyl celkového počtu SIM karet ve skupinách (25) je

$$s_u^2 = \frac{1}{3-1}\left[(12-12)^2 + (9-12)^2 + (15-12)^2\right] = 9,$$

kde $\bar{y} = (1/n)\sum_{i=1}^{n} y_i = 1/3(12 + 9 + 15) = 12$. Pomocí výběrového rozptylu již můžeme odhadnout rozptyl celkového počtu SIM karet užitím vztahu (24),

$$\widehat{var}(\hat{\tau}) = 10(10-3)\frac{9}{3} = 210,$$

a z něj vypočítat směrodatnou odchylku $\sqrt{210} = 14$.

3.3 Vícestupňový výběr

Podstatou tohoto výběru je, že v "prvním stupni" jsou vybrány primární jednotky (skupiny), ve kterých však nejsou prošetřeny všechny jejich sekundární jednotky, jak tomu bylo u předchozího výběru, ale pouze několik náhodně vybraných jednotek. V tomto případě mluvíme o dvoustupňovém výběru. Znamená to, že místo přesně zjištěného úhrnu primárních jednotek (skupin) budeme mít k dispozici pouze jeho odhad sestrojený z jednotek vybraných na "druhém stupni". Ale na druhé straně nám tento výběrový postup dovolí použít i velké skupiny a v nich vybrat poměrně malý počet jednotek. Kdybychom pokračovali dál, vybírali bychom terciární jednotky z každé vybrané sekundární jednotky, což by byl třístupňový výběr. Obecně tak pro vyšší řád mluvíme o vícestupňovém výběru.

Vícestupňový výběr je užíván v praxi u populací složených z velkého počtu jednotek (až třeba z několika miliónů). U takto velkých populací bychom při užití předchozích výběrů mohli narazit na značné problémy. Pokud bychom vzali jako populaci počet obyvatel v České republice a měli z něj vybrat menší počet osob, tak při stratifikaci by zjišťování údajů o vybraných osobách bylo velmi nákladné při jejich rozptýlenosti na tak rozlehlém území. Nebo při užití skupinového uspořádání by mohly např. domácnosti představovat příliš malé skupiny a okresy na druhou stranu příliš velké a variabilní skupiny.

Dvoustupňového výběru můžeme v praxi užít třeba i k získání výběru ulovených ryb v nějaké rybářské oblasti, kde je lepší nejdříve vybrat lodě a poté vybrat ulovené ryby z každé vybrané lodě. Nebo k získání výběru rostlin určitého druhu je vhodné nejdříve vybrat pozemky a potom z každého vybraného pozemku udělat výběr rostlin.

Označíme N jako počet primárních jednotek v populaci a M_i jako počet sekundárních jednotek v i-té primární jednotce. Potom y_{ij} je hodnota j-té sekundární jednotky v i-té primární jednotce. Úhrn y-ových hodnot v i-té primární jednotce je $y_i = \sum_{j=1}^{M_i} y_{ij}$. Průměr pro sekundární jednotky v i-té primární jednotce je $\mu_i = y_i/M_i$. Populační úhrn je $\tau = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$. Populační průměr primárních jednotek je $\mu_1 = \tau/N$, zatímco populační průměr sekundárních jednotek je

$\mu = \tau/M$, kde $M = \sum_{i=1}^{N} M_i$ je celkový počet sekundárních jednotek v populaci.

Dále budeme uvažovat (pokud nebude uvedeno jinak) dvoustupňový výběr. Odvození pro vícestupňový by pak probíhalo analogicky.

3.3.1 Prostý náhodný výběr v každém stupni

Uvažujme dvoustupňový výběr s užitím prostého náhodného výběru v každém stupni. V prvním stupni vybereme n primárních jednotek pomocí prostého náhodného výběru bez vracení. V druhém stupni z i-té primární jednotky získáme opět prostým náhodným výběrem bez vracení m_i sekundárních jednotek, pro $i = 1, 2, \ldots, n$.

Na rozdíl od skupinových výběrů se u dvoustupňových výběrů provádějí téměř vždy jen odhady úhrnu pro každou vybranou primární jednotku a odhady celkového úhrnu.

Nestranný odhad

Protože je ve druhém stupni užit prostý náhodný výběr, pak nestranný odhad úhrnu pro i-tou primární jednotku ve výběru je

$$\hat{y}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} = M_i \bar{y}_i, \tag{27}$$

kde $\bar{y}_i = (1/m_i) \sum_{j=1}^{m_i} y_{ij} = \hat{y}_i / M_i$. A jelikož je prostý náhodný výběr užit i v prvním stupni, je nestranný odhad populačního úhrnu

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} \hat{y}_i. \tag{28}$$

Rozptyl $\hat{\tau}$ je

$$var(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n} \sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}, \tag{29}$$

kde

$$\sigma_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu_1)^2 \tag{30}$$

je populační rozptyl značící variabilitu mezi jednotlivými vybranými primárními jednotkami a

$$\sigma_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2 \tag{31}$$

je populační rozptyl značící variabilitu vybraných sekundárních jednotek uvnitř i-té primární jednotky pro $i = 1, 2, \ldots, N$.

Všimněme si, že rozptyl (29) se skládá ze dvou složek. První je rovna rozptylu, který bychom obdrželi, kdyby byly u každé primární jednotky vybrány všechny jednotky sekundární. Tato složka bude tedy tím větší, čím větší bude variabilita hodnot y_i a čím méně primárních jednotek vybereme. Druhá složka pramení z toho, že se vybrané primární jednotky neprošetřují celé a tudíž se hodnoty y_i odhadují na základě podvýběru sekundárních jednotek. Tato složka bude tím větší, čím větší bude variabilita hodnot y_{ij} a čím méně se sekundárních jednotek vybere.

Celkovou variabilitu vybraných sekundárních jednotek nám tedy tvoří variabilita mezi a uvnitř vybraných primárních jednotek. Tedy rozptyly (30) a (31) se nemohou měnit nezávisle, protože zvětšení jednoho způsobí zmenšení druhého a naopak. To znamená, že pokud vytvoříme vnitřně stejnorodé primární jednotky, pak z nich sice můžeme vybírat poměrně málo sekundárních jednotek, ale jelikož jsou primární jednotky mezi sebou odlišné, tak musíme na druhou stranu těchto primárních jednotek vybrat více.

Nestranný odhad rozptylu $\hat{\tau}$ získáme nahrazením populačních rozptylů výběrovými rozptyly, jinak vše zůstává stejné, tedy

$$\widehat{var}(\hat{\tau}) = N(N - n)\frac{s_u^2}{n} + \frac{N}{n} \sum_{i=1}^{n} M_i(M_i - m_i)\frac{s_i^2}{m_i}, \tag{32}$$

kde

$$s_u^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (\hat{y}_i - \hat{\mu}_1)^2 \tag{33}$$

a

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 \tag{34}$$

a platí $\hat{\mu}_1 = (1/n) \sum_{i=1}^{n} \hat{y}_i$.

Odhady populačních průměrů pak už dopočítáme snadno. Nestranným odhadem populačního průměru primárních jednotek je $\hat{\mu}_1 = \hat{\tau}/N$, jehož rozptyl získáme vydělením vztahu (32) konstantou N^2, a nestranný odhad populačního průměru sekundárních jednotek je $\hat{\mu} = \hat{\tau}/M$, jehož rozptyl získáme obdobně vydělením vztahu (32) konstantou M^2.

Podílový odhad

Podílový odhad populačního úhrnu je opět založen na velikosti skupin čili na rozsahu primárních jednotek a má tvar

$$\hat{\tau}_r = \hat{r} M,$$

kde

$$\hat{r} = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} M_i}.$$

Střední kvadratická chyba nebo také rozptyl tohoto odhadu je dán přibližně vztahem

$$var(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (y_i - M_i \mu)^2 + \frac{N}{n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{\sigma_i^2}{m_i}$$

a odhad tohoto rozptylu je

$$\widehat{var}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} (\hat{y}_i - M_i \hat{r})^2 + \frac{N}{n} \sum_{i=1}^{n} .M_i (M_i - m_i) \frac{s_i^2}{m_i}$$

Odhady pro populační průměry μ_1 a μ jsou $\hat{\mu}_{1r} = \hat{\tau}_r/N$ a $\hat{\mu}_r = \hat{\tau}_r/M = \hat{r}$, jejichž rozptyly získáme opět vydělením rozptylu odhadu $\hat{\tau}_r$ konstantami N^2 nebo M^2.

3.3.2 Výběry primárních jednotek s pravděpodobnostmi úměrnými jejich rozsahu

Výběry se stejnými pravděpodobnostmi na obou stupních nejsou však z hlediska vydatnosti odhadu ve všech případech vhodné, zvláště tehdy, jsou-li vybrané primární jednotky příliš velké a značně variabilní. Proto užíváme v těchto případech výběry primárních jednotek s pravděpodobnostmi úměrnými jejich velikostem (počtu sekundárních jednotek). Budeme předpokládat, že v prvním stupni jsou jednotky vybrány s vracením pomocí pravděpodobností úměrných rozsahu. Pak v každé vybrané primární jednotce se provede výběr stanoveného počtu sekundárních jednotek nezávisle na tom, jestli už byla primární jednotka vybrána dříve, a které sekundární jednotky z ní již byly při té příležitosti vybrány a zahrnuty do výběru. To znamená, že každá sekundární jednotka se může ve výběru objevit více než jedenkrát. Přestože vlastní prošetření několikrát vybrané sekundární jednotky se provede jen jednou, při výpočtu odhadu ji zahrneme do výběru tolikrát, kolikrát byla vybrána.

Pokud použijeme tuto metodu, pak nestranný odhad populačního úhrnu je

$$\hat{\tau}_p = \frac{M}{n} \sum_{i=1}^{n} \frac{\hat{y}_i}{M_i} = \frac{M}{n} \sum_{i=1}^{n} \bar{y}_i,$$

kde $\bar{y}_i = (1/m_i) \sum_{j=1}^{m_i} y_{ij}$ je výběrový průměr uvnitř i-té primární jednotky ve výběru a $\hat{y}_i = M_i \bar{y}_i$.

Rozptyl je roven

$$var(\hat{\tau}_p) = \frac{M}{n} \sum_{i=1}^{N} M_i(\mu_i - \mu)^2 + \frac{M}{n} \sum_{i=1}^{N} \left[\frac{M_i - m_i}{m_i(M_i - 1)} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2 \right]$$

a nestranný odhad tohoto rozptylu

$$\widehat{var}(\hat{\tau}_p) = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu}_p)^2,$$

kde $\hat{\mu}_p = \hat{\tau}_p / M$.

3.3.3 Vícestupňový výběr s vracením

Všimněme si, že při odhadování rozptylu jsme v podkapitole 3.3.2 dostali poněkud jednodušší vzorce, protože jsme v prvím stupni užili výběr s vracením, který je spojen s jednodušší teorií odhadu a nevede k výraznějšímu snížení přesnosti odhadu. Ve skutečnosti je ale odhadování rozptylu stejně jednoduché pro každý vícestupňový výběr, ve kterém jsou primární jednotky taženy postupně s vracením se známými pravděpodobnostmi výběru p_i. Podvýběry mezi jednotlivými primárními jednotkami jsou nezávislé a nestranný odhad úhrnu $\hat{y}_i$ vypočítáme pro každou vybranou primární jednotku i. Pak nestranný odhad populačního úhrnu τ je

$$\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{y}_i}{p_i},$$

kde $\hat{y}_i = M_i \bar{y}_i$, což víme z pokapitoly 3.3.2, a $p_i = M_i/M$.

Nestranný odhad rozptylu tohoto odhadu je potom

$$\widehat{var}(\hat{\tau}_p) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{y}_i}{p_i} - \hat{\tau} \right)^2 .$$

Tento vztah plyne z nezávislosti výběrů s vracením v jednotlivých primárních jednotkách. Jednoduchost tohoto vztahu přitom nezávisí na počtu stupňů ve výběru, ale na tom, jakou metodu jsme v jakém stupni zvolili.

3.3.4 Náklady a rozsah výběru

Výhodou dvoustupňového výběru je, že umí vyhovět častým požadavkům z praxe, aby šetření bylo omezeno na co nejmenší počet vybraných primárních jednotek a v nich aby bylo provedeno na stejném počtu vybraných sekundárních jednotek. Budeme tedy uvažovat nestranný odhad $\hat{\tau}$ s prostým náhodným výběrem n primárních jednotek a prostým náhodným výběrem m_i sekundárních jednotek z každé vybrané i-té primární jednotky. Pro zjednodušení budeme dále uvažovat, že všechny primární jednotky jsou stejného rozsahu, tj. $M_i = \bar{M}$ pro

každé i, a že rozsah podvýběru každé vybrané primární jednotky je m sekundárních jednotek.

Předpokládáme, že průměrné náklady výběru jsou popsány touto nákladovou funkcí

$$C = c_0 + c_1 n + c_2 nm,$$

kde C značí celkové náklady na šetření, c_0 pevné režijní náklady, c_1 náklady na výběr primárních jednotek a c_2 náklady na výběr sekundárních jednotek. Počet sekundárních jednotek ve výběru je nm. Pro pevné celkové náklady C získáme nejmenší hodnotu rozptylu odhadu $\hat{\tau}$ s rozsahem podvýběru

$$m_{opt} = \sqrt{\frac{c_1 \sigma_w^2}{c_2(\sigma_b^2 - \sigma_w^2/\bar{M})}},$$

kde σ_b^2 je průměrný rozptyl mezi primárními jednotkami,

$$\sigma_b^2 = \frac{\sum_{i=1}^{N}(\mu_i - \mu)^2}{N-1},$$

a σ_w^2 je průměrný rozptyl uvnitř primárních jednotek,

$$\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2.$$

Pokud σ_b^2 není větší než $\sigma_w^2/\bar{M}$, lze aplikovat přímo $m_{opt} = \bar{M}$. Když užijeme optimální rozsah podvýběru m_{opt}, pak pro n z nákladové rovnice dostaneme

$$n = (C - c_0)/(c_1 + c_2 m_{opt}).$$

Příklad 7.

Chceme odhadnout celkový počet ulovených ryb v určité rybářské oblasti. V této oblasti se pohybuje 10 rybářských lodí (primární jednotky) a každá má 5 sítí (sekundární jednotky). Prostým náhodným výběrem vybereme v prvním stupni

2 lodě a ve druhém stupni 3 sítě z každé vybrané lodi. Počet ulovených ryb na první lodi je 8, 7, 3 a na druhé lodi je 5, 12 a 8.

Nejdříve musíme vypočítat odhady na každé lodi zvlášť pomocí vztahu (27). Na první lodi je tedy odhad celkového počtu ulovených ryb

$$\hat{y}_1 = \frac{5}{3}(8 + 7 + 3) = 30$$

a na druhé lodi je

$$\hat{y}_2 = \frac{5}{3}(5 + 12 + 8) = 42.$$

Poté užitím vztahu (28) odhadneme celkový počet ulovených ryb v celé oblasti:

$$\hat{\tau} = \frac{10}{2}(30 + 42) = 360.$$

Dále zjistíme směrodatnou odchylku tohoto odhadu. Nejdříve vypočítáme variabilitu mezi vybranými loděmi (33):

$$s_u^2 = \frac{1}{2-1}\left[(30 - 36)^2 + (42 - 36)^2\right] = 72,$$

kde $\hat{\mu}_1 = \hat{\tau}/N = 360/10 = 36$. Pak variabilitu vybraných sítí v každé vybrané lodi (34):

$$s_1^2 = \frac{1}{3-1}\left[(8 - 6)^2 + (7 - 6)^2 + (3 - 6)^2\right] = 7,$$

pro $\bar{y}_1 = \hat{y}_1/M_1 = 30/5 = 6$ a

$$s_2^2 = \frac{1}{3-1}\left[(5 - 8.4)^2 + (12 - 8.4)^2 + (8 - 8.4)^2\right] = 12,$$

kde $\bar{y}_2 = \hat{y}_2/M_2 = 30/5 = 8.4$.
Nyní už můžeme odhadnout rozptyl celkového počtu ulovených ryb (32):

$$\widehat{var}(\hat{\tau}) = 10(10 - 2)\frac{72}{2} + \frac{10}{2}\left[5(5 - 3)\frac{7}{3} + 5(5 - 3)\frac{12}{3}\right] = 3195$$

a jeho směrodatná odchylka je $\sqrt{3195} = 57$.

4 Praktická ukázka

Úkolem je zjistit průměrnou nezaměstnanost ve fiktivní populaci pomocí prostého náhodného výběru a stratifikovaného výběru a následně tyto metody porovnat.

Na přiloženém CD máme k dispozici reálná data z Výběrového šetření pracovních sil z Českého statistického úřadu. Jedná se o fiktivní populaci, která se skládá z $N = 58205$ osob z různých obcí. Jednotlivé osoby jsou popsány svým pohlavím, věkem, nejvyšším dosaženým vzděláním a ekonomickou aktivitou. Nás bude zajímat ekonomická aktivita jednotlivých osob.

Jako první si zvolíme rozsah výběru. V praxi se nejčastěji volí 1%-ní výběr z celé populace. Ale samozřejmě záleží na tom, jak přesné výsledky požaduji a kolik peněz na šetření mám k dispozici. My si zvolíme větší výběr pro větší přesnost odhadů a menší výběrové chyby, a to 5 000 osob.

Celý výpočet provádíme pomocí statistického softwaru R (www.r-project.org), jehož zdrojový kód je na přiloženém CD. Nejdříve prostým náhodným výběrem vybereme $n = 5000$ osob z celé populace, na kterých se bude provádět šetření. Z tohoto vzorku pak pomocí vztahu (2) odhadneme průměrnou nezaměstanost pro celou populaci $\bar{y}$, kde je $y_i = 1$, pokud je do výběru zahrnuta nezaměstnaná osoba, jinak $y_i = 0$. To vše opakujeme na padesáti různých výběrech. Výsledky jsou uvedeny v tabulce.

Dále odhadneme opět průměrnou nezaměstanost na vybraném vzorku $n = 5000$ osob, ale pomocí stratifikovaného výběru. U této metody je ovšem postup složitější. Celou populaci rozdělíme do dvou vrstev podle pohlaví, dostaneme celkový počet mužů $N_1 = 27816$ a celkový počet žen $N_2 = 30389$. Rozsah výběru z obou vrstvev určíme proporcionálním rozvržením (21) a získáme počet mužů ve výběru $n_1 = 2389$ a počet žen ve výběru $n_2 = 2611$. Dále už můžeme určit průměrnou nezaměstnanost v jednotlivých vrstvách pomocí vztahu (17), kde je $y_{hi} = 1$, pokud je do výběru zahrnuta nezaměstnaná osoba, jinak $y_{hi} = 0$. A nakonec odhadneme průměrnou nezaměstanost pro celou populaci $\bar{y}_{st}$ vztahem (18). Celý postup opět opakujeme pro padesát různých výběrů. Výsledky jsou uvedeny

v tabulce.

Číslo výběru	Odhad $\bar{y}$ v %	Odhad $\bar{y}_{st}$ v %	Číslo výběru	Odhad $\bar{y}$ v %	Odhad $\bar{y}_{st}$ v %
1	3.28	3.240	26	3.44	3.679
2	2.70	3.600	27	3.24	3.320
3	3.06	3.339	28	3.20	3.499
4	2.80	3.439	29	3.32	3.239
5	3.10	3.099	30	3.18	2.959
6	3.48	3.500	31	3.36	2.999
7	3.48	3.719	32	3.32	3.540
8	3.50	2.979	33	3.12	3.140
9	3.94	3.260	34	3.10	3.219
10	3.30	3.219	35	3.00	3.099
11	3.10	3.460	36	3.62	3.259
12	3.38	3.099	37	3.26	2.859
13	3.14	3.240	38	3.08	3.279
14	3.10	3.559	39	3.02	3.220
15	3.12	3.219	40	3.36	3.280
16	3.62	3.399	41	3.34	3.420
17	2.88	3.420	42	3.12	3.160
18	3.46	3.679	43	3.42	3.099
19	3.16	3.359	44	3.30	3.479
20	3.14	3.479	45	3.46	3.619
21	3.40	2.979	46	2.78	3.360
22	3.42	3.460	47	3.40	2.799
23	3.40	3.700	48	3.18	3.059
24	2.76	3.299	49	3.58	3.160
25	3.14	3.240	50	3.54	3.200

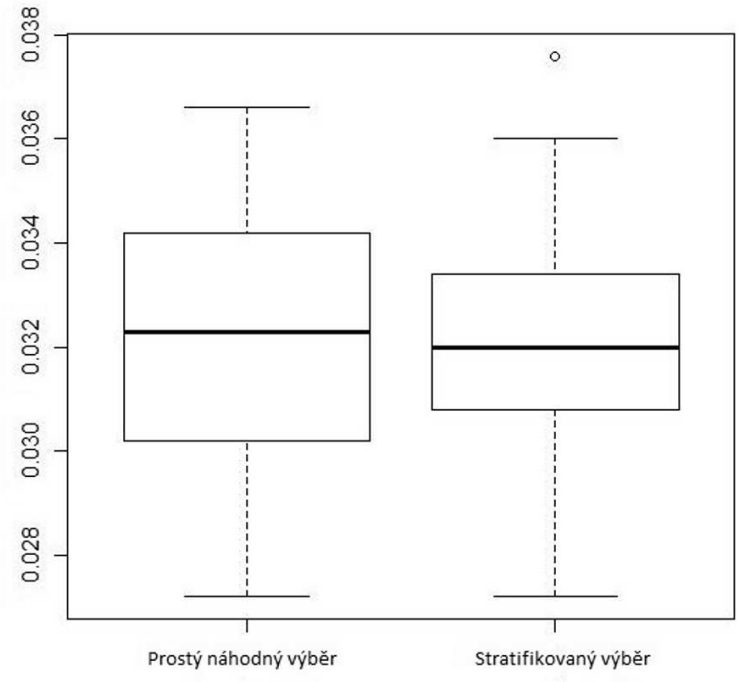
Tab.: Odhadnutá průměrná nezaměstanost pro obě metody výběru

Pro srovnání uvedeme skutečnou hodnotu průměrné nezaměstnanosti v celé populaci pomocí vztahu (1). Protože celkový počet nezaměstnaných v celé populaci je 1 894 osob, činí průměrná nezaměstanost v celé populaci $\mu = 3.254$ %.

Nakonec výsledky z tabulky promítneme do grafu pro porovnání obou metod výběru. Použijeme boxplot neboli krabicový graf.

Boxplot zobrazuje data ve tvaru obdélníku a dvou úseček, které z něj vy-

bíhají. Obdélník obsahuje 50 % dat a uprostřed je rozdělen mediánem $\tilde{x}_{0.5}$ pro symetricky rozložené hodnoty. Jeho dolní hrana je určena dolním kvartilem $\tilde{x}_{0.25}$ a jeho horní hrana horním kvartilem $\tilde{x}_{0.75}$. Délka obdélníku neboli kvartilové rozpětí $(\tilde{x}_{0.75} - \tilde{x}_{0.25})$ nám ukazuje stupeň variability daného souboru. Hodnoty ležící na úsečkách jsou od dolního (horního) kvartilu vzdáleny nejvýše 1,5násobek kvartilového rozpětí, na koncích úseček tedy leží minimální a maximální hodnoty souboru. Hodnoty, které jsou větší než horní kvartil (respektive menší než dolní kvartil) o více jak 1,5násobek kvartilového rozpětí, jsou tzv. odlehlá pozorování a jsou vyznačovány jako izolované body.



Obr. Srovnání obou metod výběru pomocí boxplotu

V našem případě můžeme vidět, že mediány u obou metod jsou si skoro rovny, ale že stratifikovaný výběr je méně variabilní, a tedy dává i přesnější odhady. To se děje, pokud je stratifikace pro řešený problém vhodně zvolena. Tak je tomu opravdu i v tomto případě, neboť nezaměstnanost se chová jinak u mužů než u žen.

Závěr

Psaní této práce mi dalo možnost prohloubit si znalosti v oblasti statistiky, která je využitelná v praxi, což je určitě nejen pro mě, ale i pro čtenáře, velkým přínosem. Popsala jsem, jaké druhy statistického šetření se v praxi používají, a dále pak konkrétně výběrová šetření, kterým je věnována většina diplomové práce. Jsou zde vysvětleny jednotlivé metody výběrových šetření od nejzákladnějších až po ty složitější.

Teď už víme, že podle přesnosti výsledků, jakou zadavatel šetření požaduje, stanovíme rozsah výběru, abychom získali menší výběrové chyby. Ale také záleží na finančních prostředcích, které máme pro šetření k dispozici. Čím větší bude rozsah výběru nebo složitost (obtížnost) šetření, tím samozřejmě nákladnější šetření bude. Dalším úskalím mohou být nekvalitní a málo prověření tazatelé, kteří mohou podávat zkreslené informace o prošetřovaných jednotkách, což vede k nepřesným výsledkům.

Díky této práci jsem se také naučila pracovat se statistickým softwarem R, pomocí něhož můžeme docela snadno řešit konkrétní situace, které se dané problematiky týkají, a to i pokud máme velké množství dat.

Doufám, že tato práce pomůže čtenářům lépe pochopit problematiku výběrových šetření a jejich užití v praxi. A věřím, že i pro mě bude přínosem v budoucím životě.

Literatura

[1] Čermák, V., Vrabec, M.: Teorie výběrových šetření, část 1., VŠE Praha, 1999

[2] Čermák, V., Vrabec, M.: Teorie výběrových šetření, část 2., VŠE Praha, 1998

[3] Čermák, V., Vrabec, M.: Teorie výběrových šetření, část 3., VŠE Praha, 1999

[4] Čermák, V., Vrabec, M.: Teorie výběrových šetření, sbírka úloh, VŠE Praha, 2003

[5] Dupač, V., Hájek, J.: Pravděpodobnost ve vědě a technice, Nakladatelství Československé akademie věd, Praha, 1962

[6] Hájek, J.: Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření, Nakladatelství Československé akademie věd, Praha, 1960

[7] Thompson, S., K.: Sampling, second edition, A Wiley-Intercience Publication, New York, 2002

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

[9] Vytlačil, J.: Výběrová šetření v praxi, SEVT, Praha, 1969