Pokroky matematiky, fyziky a astronomie

Jiří Dvořák

Predikce výsledků voleb a stratifikované náhodné výběry

Pokroky matematiky, fyziky a astronomie, Vol. 68 (2023), No. 2, 69–80

Persistent URL: http://dml.cz/dmlcz/151745

Predikce výsledků voleb a stratifikované náhodné výběry

Jiří Dvořák

Abstrakt. Koncept stratifikovaného náhodného výběru představuje užitečnou alternativu prostého náhodného výběru v situaci, kdy je zkoumaná populace složena z několika částí s různými vlastnostmi. Je tak možné například při stejném rozsahu výběru získat odhad průměru s menším rozptylem. V tomto příspěvku si představíme základy teorie stratifikovaných náhodných výběrů a na příkladu druhého kola prezidentských voleb 2023 si ukážeme jejich praktické použití.

1. Predikce výsledků voleb

Volby představují pro řadu médií vděčné téma, které spolehlivě připoutá řadu lidí k obrazovkám telefonů, televizí či počítačů. Jednotlivé televizní kanály a zpravodajské servery investují měsíce práce do přípravy co nejzajímavějšího volebního studia, zvou respektované i jiné komentátory a experty, ladí všechny detaily grafiky a animací. Divák či čtenář však obvykle nevydrží pozorně sledovat volební studio až do sečtení posledního okrsku. Stalo se proto předmětem jisté prestiže nabídnout co nejdříve co nejpřesnější odhad konečného výsledku voleb.

Ukážeme si, jak v tomto kontextu využít koncept stratifikovaných náhodných výběrů, a na příkladu druhého kola prezidentských voleb 2023 předvedeme, že odhady získané z první malé části sečtených hlasů můžou být velmi přesné. Daří se tedy vyrovnat se skutečností, že mezi prvními bývají sečtené okrsky s malým počtem voličů, typicky v malých obcích, kde jsou obvykle názory voličů odlišné od velkých měst, odkud naopak volební výsledky přicházejí později.

2. Výběry z konečných populací

Představme si, že v dané populaci jedinců sledujeme hodnotu nějakého znaku, například výšku (v centimetrech), měsíční příjem (v korunách), nebo zda budou volit daného prezidentského kandidáta ($1 =$ ano, $0 =$ ne). Předpokládáme, že daná populace obsahuje konečný počet jedinců, ale z časových, ekonomických či jiných důvodů nedovedeme zjistit hodnotu tohoto znaku u každého jedince. Provedeme tedy nějakou formu výběrového šetření a náhodně určíme, které jedince do našeho zjišťování zahrneme a které ne.

Z celé populace N jedinců tedy vybereme vzorek o velikosti n, kde $1 \leqq n < N$. Předpokládáme, že hodnoty sledovaného znaku jsou pro jednotlivé jedince pevně dané. Pokud se hodnoty sledovaného znaku mění v čase, vztahujeme dané zjišťování k určitému časovému okamžiku a v tomto okamžiku je považujeme za nenáhodné. Veškerá

RNDr. Jiří Dvořák, Ph.D., Katedra pravděpodobnosti a matematické statistiky MFF UK, Sokolovská 83, 186 75 Praha 8, e-mail: dvorak@karlin.mff.cuni.cz

náhodnost se tedy týká toho, které jedince do svého zjišťování zahrneme. To se odlišuje od běžnější situace, kdy uvažujeme nezávislé, stejně rozdělené náhodné veličiny $X_1, \ldots, X_n$, reprezentující výběr z hypotetické nekonečné populace.

Ve výkladu se pro přehlednost zaměříme na úlohu odhadu průměru sledovaného znaku přes celou populaci, například odhad průměrného platu nebo podílu lidí, kteří budou volit daného kandidáta, viz začátek kapitoly 3. Je ovšem snadné upravit postup pro odhad úhrnu přes celou populaci, například odhad celkového počtu hlasů, které ve volbách získá daný kandidát. Vzorce uvedené níže přebíráme z knihy [13], byť s mírně upraveným značením.

3. Odhad průměru

Uvažujme konečnou populaci N jedinců, očíslovaných $1, 2, \ldots, N$. Hodnotu, která nás na i-tém jedinci zajímá, označíme y_i. Populační průměr μ udává průměrnou hodnotu sledovaného znaku v celé populaci,

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

V případě voleb mají jednotlivá y_i hodnotu 1 (i-tý jedinec bude volit daného kandidáta) nebo 0 (jinak). Pak μ definované předchozím vzorcem udává podíl lidí, kteří budou daného kandidáta volit.

Rozptyl v konečné populaci definujeme vzorcem

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

V jiných oblastech statistiky je zvykem v definici rozptylu uvažovat dělení hodnotou N místo $N-1$, ale v tomto kontextu je volba $N-1$ z jistých důvodů výhodnější a standardně se používá. Krátce se k této otázce vrátíme na konci odstavce 3.1.

Dále budeme uvažovat úlohu odhadu populačního průměru μ. Zaměříme se na dva různé způsoby vybírání n jedinců, které do svého zjišťování zařadíme: prostý náhodný výběr a stratifikovaný náhodný výběr.

3.1. Prostý náhodný výběr

Nejjednodušší a nejpřirozenější způsob výběru z konečné populace je takzvaný *prostý náhodný výběr*. Jde o náhodný výběr bez opakování z množiny $\{1, 2, \ldots, N\}$, a tedy každá n-tice různých jedinců má stejnou šanci být vybrána jako všechny ostatní n-tice. Náhodně vybranou n-tici zařazenou do výběru označíme $s \subset \{1, 2, \ldots, N\}$ podle anglického výrazu *sample*. Odhad populačního průměru μ na základě výběru s je pak přirozeně

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i. \tag{2}$$

Rozptyl tohoto odhadu je pak

$$\operatorname{var}(\bar{y}) = \left(\frac{N-n}{N} \right) \frac{\sigma^2}{n}. \tag{3}$$

Člen σ^2/n odpovídá rozptylu odhadu střední hodnoty na základě n pozorování nezávislých, stejně rozdělených náhodných veličin s rozptylem σ^2 (výběr z hypotetické nekonečné populace). Tento rozptyl je přímo úměrný σ^2 a nepřímo úměrný počtu pozorování n. Člen $\frac{N-n}{N}$ označujeme jako korekci na konečnou populaci a znamená, že čím více jedinců do výběru zahrneme, tím menší je míra naší nejistoty o odhadnuté hodnotě $\bar{y}$. To platí i bez korekčního členu, ale použití korekčního členu toto ještě zvýrazňuje. V extrémním případě, kdy $n = N$, máme informaci o celé populaci, korekční člen i celý rozptyl má hodnotu 0 a míra nejistoty je nulová.

Výběrový rozptyl pozorovaných hodnot je definován klasicky jako

$$S_n^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2. \tag{4}$$

Jde o nestranný odhad rozptylu v konečné populaci σ^2, tedy střední hodnotou S_n^2 je σ^2. V tomto místě se na moment vrátíme k definici σ^2 pomocí vzorce (1). Při výpočtu σ^2 máme k dispozici informaci o všech jedincích v populaci, můžeme si tedy představovat, že jsme provedli prostý náhodný výběr o rozsahu N. V takovém případě bychom chtěli, aby $S_N^2 = \sigma^2$ a vzorce (1) a (4) pro σ^2 a S_n^2 byly spolu konzistentní. Výhodnou vlastností je také nestrannost S_n^2 jako odhadu σ^2 zmíněná výše. Konečně, díky použití členu $N-1$ ve vzorci (1) má pak korekce na konečnou populaci jednodušší a z jistého pohledu přirozenější tvar $(N-n)/N = 1 - n/N$. Pokud bychom ve vzorci (1) dělili hodnotou N, museli bychom pak používat korekci na konečnou populaci ve tvaru $(N-n)/(N-1)$.

3.2. Stratifikovaný náhodný výběr

V situacích, kdy není celá populace N jedinců homogenní ve smyslu chování hodnot y_i, může být výhodné při výběru jedinců do našeho vzorku tuto skutečnost zohlednit. Pokud dovedeme populaci rozdělit do několika disjunktních podskupin, označovaných jako *strata*, tak, aby uvnitř jednotlivých strat byly vlastnosti všech jedinců pokud možno homogenní, je výhodné vybírat jedince z jednotlivých strat odděleně, nezávisle na sobě. Cílem je při stejném rozsahu výběru n získat přesnější odhad populačního průměru μ, tj. odhad s menším rozptylem, než by odpovídalo prostému náhodnému výběru z předchozí části. V různých aplikacích může být relevantní dělit populaci například podle pohlaví, dosaženého vzdělání, příslušnosti do nějakého geografického regionu (kraje, okresy), velikosti obce trvalého bydliště a podobně, případně podle nějaké kombinace těchto znaků. Zdůrazněme však ještě jednou, že cílem je vytvořit strata takovým způsobem, aby populace uvnitř jednotlivých strat byly co nejvíce homogenní.

Předpokládejme nyní, že máme populaci rozdělenou do L strat, přičemž l-té stratum zahrnuje N_l jedinců, kde $\sum_{l=1}^{L} N_l = N$. V l-tém stratu provedeme nezávisle na ostatních prostý náhodný výběr o rozsahu n_l, kde $\sum_{l=1}^{L} n_l = n$, a tuto náhodně vybranou n_l-tici označíme s_l. Tento způsob výběru jedinců, které zahrneme do našeho zjišťování, se označuje jako *stratifikovaný náhodný výběr*.

Odhad průměru v l-tém stratu je

$$\bar{y}_l = \frac{1}{n_l} \sum_{i \in s_l} y_i \tag{5}$$

a odhad populačního průměru μ na základě stratifikovaného výběru je

$$\bar{y}_{\mathrm{st}} = \frac{1}{N} \sum_{l=1}^{L} N_l \bar{y}_l. \tag{6}$$

Jde o vážený průměr dílčích odhadů z jednotlivých strat, kde váhy N_l/N určují zastoupení strat v celé populaci. Rozptyl tohoto odhadu je

$$\operatorname{var}(\bar{y}_{\mathrm{st}}) = \sum_{l=1}^{L} \left(\frac{N_l}{N}\right)^2 \left(\frac{N_l - n_l}{N_l}\right) \frac{\sigma_l^2}{n_l}, \tag{7}$$

kde σ_l^2 je rozptyl v konečné populaci v l-tém stratu, daný obdobou vzorce (1). Explicitní vzorec pro σ_l^2 zde neuvádíme, abychom nemuseli zavádět další značení určující, kteří jedinci náleží do kterého strata.

Protože jsme jednotlivá strata sestavili tak, aby populace uvnitř strat byla co nejvíce homogenní, bude rozptyl σ_l^2 v jednotlivých stratech menší než rozptyl σ^2 v celé populaci. Tím pádem může být i rozptyl (7) odhadu založeného na stratifikovaném náhodném výběru menší než rozptyl (3) odhadu založeného na prostém náhodném výběru. Takovou situaci ukáže příklad v následujícím odstavci.

3.3. Ukázka výhodnosti použití stratifikovaného náhodného výběru

Představme si nyní modelovou situaci, kdy máme populaci $N = 15$ jedinců, kterou rozdělíme do dvou strat o velikostech $N_1 = 10$ a $N_2 = 5$. V prvním stratu budou hodnoty zkoumaného znaku rovny 8, 9, 10, 11, 12, 8, 10, 10, 10 a 12. Snadným výpočtem zjistíme, že populační průměr v prvním stratu je $\mu_1 = 10$, a platí $\sigma_1^2 = 2$. V druhém stratu pak budou hodnoty zkoumaného znaku rovny 21, 23, 25, 27 a 29. Je tedy $\mu_2 = 25$ a $\sigma_2^2 = 10$. Jedinci uvnitř každého strata tedy tvoří poměrně homogenní populaci, ale napříč straty se jejich vlastnosti liší jak ve smyslu průměru, tak rozptylu. Pokud bychom zkoumali tuto populaci celou, nerozdělenou na strata, zjistili bychom, že populační průměr je $\mu = 15$ a rozptyl v konečné populaci je $\sigma^2 = 808/14 \doteq 57{,}7$.

Prostý náhodný výběr s $n = 6$ vede podle vzorce (3) k rozptylu $\operatorname{var}(\bar{y}) = 808/140 \doteq \doteq 5{,}77$. Stratifikovaný náhodný výběr s $n_1 = 4$, $n_2 = 2$ vede podle vzorce (7) k rozptylu $\operatorname{var}(\bar{y}_{\mathrm{st}}) = 7/15 \doteq 0{,}467$.

Rozptyl odhadu μ je tedy v případě stratifikovaného náhodného výběru přibližně dvanáctkrát menší, přitom ale $n_1 + n_2 = 6 = n$, a tedy nákladnost zjišťování je stejná jako u prostého náhodného výběru. Rozsahy výběrů v jednotlivých stratech jsme v tomto případě zvolili proporční celkové velikosti strat. To je přirozená volba, ale nemusí být optimální ve smyslu minimalizace rozptylu odhadu μ. Otázku optimální alokace diskutuje například kniha [13] v sekci 11.5.

Tabulka 1 uvádí několik příkladů realizace prostého náhodného výběru o rozsahu $n = 6$ v tomto ilustračním příkladu. Jednotlivé řádky odpovídají nezávislým opakováním výběru. Vidíme, že odhad $\bar{y}$ populačního průměru $\mu = 15$ je od správné hodnoty někdy i dosti vzdálený. Naproti tomu tabulka 2 ukazuje několik příkladů realizace stratifikovaného náhodného výběru s parametry $n_1 = 4$ a $n_2 = 2$. Odhad $\bar{y}_{\mathrm{st}}$ populačního průměru μ je správné hodnotě obvykle výrazně blíže, než tomu bylo v případě

Celá populace															$\bar{y}$
8	**9**	10	11	12	8	10	**10**	**10**	12	21	23	**25**	**27**	29	14,83
8	**9**	10	**11**	12	8	10	10	10	**12**	21	23	**25**	**27**	**29**	18,83
8	9	**10**	**11**	12	**8**	10	**10**	10	**12**	**21**	23	25	27	29	12,00
8	9	10	11	**12**	8	**10**	**10**	10	12	**21**	23	**25**	27	**29**	17,83
8	**9**	**10**	11	**12**	8	10	10	**10**	12	**21**	23	25	27	29	11,66

Tab. 1. Ukázka možných realizací prostého náhodného výběru v celé populaci. Jednotlivé řádky odpovídají nezávislým opakováním výběru na stejné populaci. Hodnoty zvýrazněné tučně byly zařazeny do výběru. Poslední sloupec udává odhad populačního průměru μ na základě prostého náhodného výběru.

První stratum										Druhé stratum					$\bar{y}_{\mathrm{st}}$
8	**9**	10	11	12	8	10	**10**	**10**	12	21	23	**25**	27	**29**	15,17
8	9	**10**	**11**	12	**8**	10	10	10	12	21	23	**25**	**27**	29	14,83
8	**9**	10	11	12	8	10	**10**	**10**	**12**	21	23	25	**27**	**29**	16,17
8	9	10	11	**12**	8	**10**	10	**10**	12	21	23	**25**	**27**	29	15,33
8	9	**10**	**11**	**12**	**8**	10	10	10	12	**21**	23	**25**	27	29	14,50

Tab. 2. Ukázka možných realizací stratifikovaného náhodného výběru. Jednotlivé řádky odpovídají nezávislým opakováním výběru na stejné populaci, rozdělené do dvou strat. Hodnoty zvýrazněné tučně byly zařazeny do výběru. Poslední sloupec udává odhad populačního průměru μ na základě stratifikovaného náhodného výběru.

prostého náhodného výběru v tabulce 1. To je dáno faktem, že ve stratifikovaném náhodném výběru je míra zastoupení jednotlivých strat pevně určena a není tedy možné žádnou subpopulaci (stratum) přehlédnout nebo jí věnovat menší pozornost, než si zaslouží. To se však může stát při prostém náhodném výběru, kdy je míra zastoupení jednotlivých subpopulací náhodná, viz například poslední řádek tabulky 1, kde je druhá subpopulace „vysokých hodnot“ zastoupena jen jednou hodnotou.

4. Predikce výsledků voleb pomocí stratifikace

Představme si, že v probíhajících volbách už byly uzavřeny volební místnosti a bylo zahájeno sčítání hlasů. V jistém okamžiku po začátku sčítání jsou už k dispozici výsledky z části volebních okrsků – Český statistický úřad zveřejňuje průběžně dílčí výsledky sčítání ve strojově čitelném formátu. Chtěli bychom už v tuto chvíli odhadnout, jakého výsledku dosáhne náš oblíbený kandidát na konci sčítání.

Označme N počet voličů, kteří odevzdali platný volební lístek, a symbolem y_i volbu jednotlivých voličů. Budeme psát $y_i = 1$, pokud i-tý volič vybral našeho kandidáta, a $y_i = 0$ jinak. Populační průměr $\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$ pak udává celkový výsledek našeho

kandidáta, resp. hodnota 100μ udává jeho celkový procentuální zisk. Tato hodnota bude známa na konci sčítání, ale můžeme ji zkoušet odhadnout už v průběhu sčítání z dílčích výsledků.

Zkušenost ukazuje, že průběžné výsledky se v průběhu sčítání můžou i výrazně měnit, protože malé okrsky bývají sečteny rychle, nicméně voličské preference se liší ve srovnání s velkými okrsky, které bývají sečteny po delší době. Proto bychom neměli k dílčím výsledkům přistupovat jako k realizaci prostého náhodného výběru a celkový výsledek odhadovat pomocí průběžného výsledku vzorcem (2). Můžeme však využít princip stratifikace.

4.1. Stratifikace pomocí krajů

V nejjednodušší variantě můžeme uvažovat 14 strat určených jednotlivými kraji České republiky a hlavním městem Praha. V takovém případě každý volební okrsek náleží celý do jednoho ze strat. Podle vzorce (6) pak celkový výsledek našeho kandidáta odhadneme váženým průměrem dílčích výsledků v jednotlivých stratech. Použité váhy N_l/N přitom reflektují velikost jednotlivých strat a kompenzují skutečnost, že v některých stratech je sečteno málo okrsků a v jiných hodně.

Tento odhad celkového výsledku by měl být přesnější (ve smyslu menšího rozptylu) než odhad pomocí vzorce (2), protože populace v jednotlivých krajích jsou z hlediska voličských preferencí homogennější než celá populace voličů v České republice. Narážíme však na problém, že potřebné váhy N_l/N v průběhu sčítání neznáme. Hodnoty N_l a N totiž neudávají počty oprávněných voličů v jednotlivých krajích, resp. v celé republice, které jsou pro dané volby známé, ale jde o *počty voličů, kteří odevzdali platný hlas*. Takové voliče budeme v dalších úvahách označovat jako *úspěšné voliče*. Nastává ale problém s lidmi, kteří k volbám nepřišli, případně odevzdali neplatný hlas. Jejich počet neznáme až do ukončení sčítání.

Snadným řešením je použít hodnoty N_l/N zjištěné v předchozích volbách. Relevance takového postupu je založena na předpokladu, že pokud se v rámci daného strata od posledních voleb změnil počet úspěšných voličů, změnil se ve stejném poměru v již sečtené části strata jako ve stratu celém.

4.2. Stratifikace podle výsledků předchozích voleb

Ani v rámci jednotlivých krajů není populace z pohledu voličských preferencí zcela homogenní, například ve velkých městech bývá chování voličů poněkud odlišné od malých obcí. Můžeme se tedy pokusit populaci voličů stratifikovat šikovněji, aby vzniklé subpopulace byly co nejvíce homogenní.

Zajímavou možností je stratifikace na základě výsledků předchozích voleb. Zvolíme si předem počet strat L a budeme předpokládat, že celá populace voličů se dá rozdělit na L subpopulací se shodným chováním voličů. Připouštíme, že v jednotlivých volebních okrscích jsou potenciálně zastoupeni voliči ze všech strat. To je zásadní rozdíl oproti stratifikaci podle krajů diskutované výše. Můžeme si představovat, že voliče v jednom stratu spojuje nějaká společná charakteristika, například „základní vzdělání“, „věk 65+“, ale tyto charakteristiky nezadáváme předem, ani je z vytvořených strat nedovedeme zjistit. Vzhledem k anonymitě hlasování bychom stejně nedovedli jednotlivé odevzdané hlasy roztřídit do skupin podle zadaných charakteristik voličů.

Následující úvahy se týkají předchozích voleb, například při snaze predikovat výsledky druhého kola prezidentské volby teď budeme uvažovat o výsledcích prvního kola. Navíc úvahy vychází z jiných předpokladů, a to, že volba jednotlivých voličů je náhodná a každý volič má pouze jistou pravděpodobnost, že bude volit našeho kandidáta. Připomínáme, že u aktuálních voleb, kde se snažíme predikovat výsledky v průběhu sčítání, předpokládáme, že rozhodnutí jednotlivých voličů jsou daná a náhodně vybraný je pouze vzorek voličů, jejichž hlasy byly dosud sečteny. Úvahy o předchozích volbách a náhodných hlasech slouží jen ke stratifikaci populace voličů.

Jednotlivá strata budeme rozlišovat pomocí indexu $l \in \{1, \ldots, L\}$, volební okrsky pak pomocí indexu $k \in \{1, \ldots, K\}$. Počty platných odevzdaných hlasů v jednotlivých okrscích budou $M_1, \ldots, M_K$. Míra podpory našeho kandidáta v l-tém stratu bude označena $p_l \in [0, 1]$, kde $l \in \{1, \ldots, L\}$. Číslo p_l formálně udává pravděpodobnost, že náhodně vybraný volič ze strata l bude volit našeho kandidáta.

Dále symbolem X_k označíme počet platných hlasů pro našeho kandidáta v k-tém okrsku a symbol $\pi_k^{(l)}$ bude značit, jaká část okrsku k náleží do strata l. Platí pak $\pi_k^{(l)} \in [0, 1]$ a $\sum_{l=1}^{L} \pi_k^{(l)} = 1$ pro každé $k \in \{1, \ldots, K\}$, protože každý volič v okrsku musí patřit do některého strata.

Protože veličinu X_k tvoří příspěvky od voličů z různých strat, budeme její rozdělení reprezentovat váženým součtem binomických rozdělení, symbolicky

$$X_k \sim \sum_{l=1}^{L} \pi_k^{(l)} \text{ Bi } (M_k, p_l),$$

kde Bi (n, p) označuje binomické rozdělení s parametry n a p, tedy rozdělení počtu úspěchů v posloupnosti n nezávislých experimentů, kde v každém experimentu nastává úspěch s pravděpodobností p. Hodnoty parametrů $\pi_k^{(l)}$ a p_l pro různá k a l jsou neznámé, dají se však odhadnout pomocí tzv. EM-algoritmu [7] z hodnot $X_1, \ldots, X_K$, které jsou známé, protože jde o výsledky předchozích voleb. Pro další postup samozřejmě nebudeme používat odhadnuté pravděpodobnosti p_l, ty se mezi minulými a současnými volbami mohly změnit, ale využijeme odhadnuté hodnoty $\pi_k^{(l)}$, popisující zastoupení jednotlivých strat v populaci voličů daného okrsku, a také odhadnuté velikosti strat $N_l = \sum_{k=1}^{K} \pi_k^{(l)} M_k$, $l \in \{1, \ldots, L\}$.

Nyní se vraťme k aktuálním volbám, kdy máme k dispozici informace o první sadě sečtených okrsků. Hlasy pro našeho kandidáta z okrsku k teď přerozdělíme do jednotlivých strat v poměru daném hodnotami $\pi_k^{(l)}$, $l \in \{1, \ldots, L\}$. Po přerozdělení hlasů ze všech dosud sečtených okrsků získáme počty hlasů pro našeho kandidáta v jednotlivých stratech. Podobně postupujeme s hlasy proti našemu kandidátovi, což umožní spočítat $\bar{y}_l$ podle vzorce (5). Pak už můžeme odhadnout celkový výsledek našeho kandidáta pomocí vzorce (6).

Poznamenejme nakonec, že uvedený postup se dá snadno zobecnit z binomického rozdělení na multinomické v případě více kandidujících subjektů. Takový postup je v praxi relevantnější, protože i v případě druhého kola prezidentské volby, kde voliči vybírají ze dvou kandidátů, potřebujeme k odhadu $\pi_k^{(l)}$ výsledky předchozích voleb, kde kandidovalo více subjektů (první kolo prezidentské volby, případně předchozí parlamentní volby). Pro jednoduchost jsme se ale v předchozím výkladu drželi přehlednější situace se dvěma kandidáty v předchozích volbách.

5. Ukázka výsledků – druhé kolo prezidentské volby 2023

Doc. Ing. Marek Omelka, Ph.D., Bc. Hedvika Ranošová a Mgr. Ondřej Týbl z katedry pravděpodobnosti a matematické statistiky MFF UK na objednávku Českého rozhlasu připravili predikční model popsaný v odstavci 4.2 pro účely predikce výsledků parlamentních voleb na podzim 2021. Vypočtené predikce byly poměrně úspěšné, například se podařilo predikovat vypadnutí ČSSD z Poslanecké sněmovny už na 20 procentech sečtených okrsků, přestože výsledky sčítání to poprvé ukázaly přibližně na 40 procentech sečtených okrsků.

Tento postup predikce byl pak použit také v případě prezidentských voleb v roce 2023, stratifikace do sedmi strat byla založena právě na výsledcích parlamentních voleb z roku 2021. Pro ukázku uvádíme v tabulce 3 porovnání průběžných výsledků

Sečteno	Čas	PP_{real}	AB_{real}	PP_{pred}	AB_{pred}
5	14.29	52,28	47,71	58,82	41,18
10	14.33	53,52	46,47	59,47	40,53
15	14.36	53,70	46,29	59,16	40,84
20	14.39	53,86	46,13	59,16	40,84
25	14.43	53,91	46,08	59,16	40,84
30	14.46	54,10	45,89	59,16	40,84
35	14.48	54,21	45,78	59,16	40,84
40	14.50	54,40	45,59	58,85	41,15
45	14.53	54,65	45,34	58,85	41,15
50	14.56	54,91	45,08	58,85	41,15
55	14.59	55,26	44,73	58,85	41,15
60	15.02	55,45	44,54	58,78	41,22
65	15.05	55,67	44,32	58,78	41,22
70	15.09	55,87	44,12	58,78	41,22
75	15.13	56,12	43,87	58,76	41,24
80	15.17	56,42	43,57	58,61	41,39
85	15.23	56,72	43,27	58,61	41,39
90	15.31	57,11	42,88	58,42	41,58
95	15.43	57,55	42,44	58,39	41,61
100	**17.23**	**58,32**	**41,67**		

Tab. 3. Vývoj průběžných výsledků sčítání druhého kola prezidentské volby 2023 v závislosti na množství sečtených okrsků. První sloupec udává, kolik procent okrsků bylo v danou chvíli sečteno, druhý sloupec odpovídající čas (sčítání začalo okamžitě po uzavření volebních místností ve 14.00). Další dva sloupce ukazují průběžné výsledky sčítání pro Petra Pavla (PP) a Andreje Babiše (AB). Poslední dva sloupce ukazují predikce celkových výsledků postupem z odstavce 4.2.

sčítání s predikcemi vypočtenými ve stejnou chvíli z aktuálně dostupných dat. Predikce přebíráme z webu Českého rozhlasu [4] tak, jak byly zveřejněny, průběžné výsledky sčítání pak přebíráme ze záznamu volebního studia České televize [3].

Z tabulky je vidět, že predikce byly velmi přesné už ve chvíli, kdy bylo sečteno jen 5 % volebních okrsků. Upozorňujeme však, že ne všechny okrsky jsou stejně velké, a v danou chvíli bylo ve skutečnosti sečteno jen 83 799 platných hlasů z celkových 5 759 197 platných hlasů odevzdaných v druhém kole prezidentské volby, tedy predikce vycházely přibližně z 1,46 procenta sečtených hlasů. Dále vidíme, že se predikce na některých řádcích shodují. To je způsobeno technickými problémy při načítání dat, v některých okamžicích tedy nemohla být aktuální predikce spočítána a byla zveřejněna poslední dostupná predikce.

6. Další přístupy k predikci výsledků během sčítání

Své predikce výsledků druhého kola prezidentských voleb 2023 zveřejnilo také několik soukromých společností. V abecedním pořadí jde o Blindspot Solutions [4], [1], iDNES.cz [6], PAQ research [4], [8], Seznam Zprávy [10] a STEM/MARK [2], [11].

Tyto společnosti jsme oslovili s žádostí o vyjádření k jejich přístupu, abychom mohli čtenářům předložit úplnější přehled možných řešení. Reakce tří společností, které na žádost alespoň krátce odpověděly, shrnujeme níže. Od ostatních společností nemáme k dispozici žádné informace navíc oproti jejich webovým stránkám odkazovaným výše. Jde samozřejmě o interní know-how jednotlivých společností a dá se pochopit, že zveřejnění podrobností mohou vnímat jako nežádoucí.

6.1. iDNES.cz

Společnost iDNES.cz připravila predikci specificky pro druhé kolo prezidentské volby. Predikce byla založena na použití kompenzačních faktorů, které byly určeny předem, mezi prvním a druhým kolem prezidentské volby. Na následujících řádcích shrnujeme postup popsaný v emailové komunikaci se zástupcem iDNES.cz.

Při konstrukci predikce nejprve vezmeme výsledky z prvního kola prezidentské volby 2023 a hlasy od všech kandidátů, kteří nepostoupili do druhého kola, rozdělíme mezi Petra Pavla a Andreje Babiše v poměru 70 : 30 (jedná se o odhad přelévání hlasů). Tím získáme výchozí odhady absolutních výsledků po jednotlivých okrscích, pro jednotlivé kandidáty je označíme PP_k a AB_k, $k \in \{1, \ldots, K\}$.

Dále přirozeným způsobem ztotožníme kandidáty v druhém kole prezidentských voleb 2023 (Petr Pavel, Andrej Babiš) s kandidáty v druhém kole prezidentských voleb 2018 (Jiří Drahoš, Miloš Zeman) a analyzujeme hypotetické dvoukolové volby, zahrnující první kolo voleb 2023 a druhé kolo voleb 2018. Označme $PP(t)$ průběžný procentuální výsledek Petra Pavla po sečtení t procent volebních okrsků v druhém kole této hypotetické volby. Kompenzační faktor používaný dále udává, jakým číslem musíme přenásobit průběžný procentuální výsledek Petra Pavla, abychom získali jeho celkový výsledek, získáme jej tedy jako podíl $PP(100)/PP(t)$.

V těchto hypotetických volbách tedy máme po sečtení t procent okrsků k dispozici pro sečtené okrsky „skutečné počty hlasů“ pro jednotlivé kandidáty, pro nesečtené okrsky pak tyto počty odhadneme součinem $PP_k \cdot PP(100)/PP(t)$ pro Petra Pavla

a podobně pro Andreje Babiše (upozorňujeme, že hodnoty PP_k v tomto odstavci se vztahují k těmto hypotetickým volbám a liší se od hodnot PP_k pro skutečné volby, které jsme zkonstruovali na začátku této kapitoly). Z těchto „skutečných“, resp. predikovaných absolutních čísel pro jednotlivé okrsky pak snadno získáme predikci celkového procentuálního výsledku každého z kandidátů v druhém kole této hypotetické volby.

Predikované výsledky vyneseme do grafu jako funkci t a všimneme si, že Pavlův odhad lineárně stoupá s rostoucím t. Proto zkonstruujeme klesající lineární funkci $f_{\mathrm{PP}}(t)$ splňující $f_{\mathrm{PP}}(100) = 1$, kterou použijeme jako multiplikativní korekci a která zajistí, že získané predikce jsou co nejvíce konstantní (tento kandidát potřebuje v průběhu sčítání korekci s hodnotou > 1, ale hodnota korekčního faktoru postupem času klesá k 1). Podobně pak zkonstruujeme pro druhého kandidáta rostoucí lineární funkci $f_{\mathrm{AB}}(t)$ splňující $f_{\mathrm{AB}}(100) = 1$. Tato korekce souvisí opět s faktem, že malé okrsky bývají sečteny dříve, ale voličské preference se v nich obvykle liší od velkých okrsků, které bývají sečteny později.

Z uvažovaných hypotetických voleb jsme získali sadu kompenzačních faktorů a dvě korekční funkce. Při sčítání ve druhém kole skutečných voleb 2023 pak po sečtení t procent volebních okrsků máme k dispozici skutečné počty hlasů získaných jednotlivými kandidáty v sečtených okrscích, pro nesečtené okrsky tyto počty odhadneme součinem $PP_k \cdot PP(100)/PP(t)$ pro Petra Pavla a podobně pro Andreje Babiše. Z těchto skutečných, resp. predikovaných absolutních čísel pro jednotlivé okrsky pak snadno získáme predikci celkového procentuálního zisku každého z kandidátů, kterou nakonec přenásobíme odpovídající korekční funkcí.

Uvedený postup je heuristický, založený na předpokladu, že tábory voličů jednotlivých kandidátů ve druhém kole prezidentské volby 2023 rozumně odpovídají táborům voličů jednotlivých kandidátů v druhém kole prezidentské volby 2018. Získané predikce byly velmi přesné, například již pro 2,41 % sečtených okrsků byl predikovaný celkový výsledek Petra Pavla 58,57 %, Andreje Babiše 41,43 %, viz stránku [6].

6.2. Seznam Zprávy

Mgr. Michal Škop, Ph.D., (Seznam Zprávy, KohoVolit.eu) v naší emailové komunikaci vyzvedl význam otevřených dat poskytovaných Českým statistickým úřadem [9], [5]. Skutečný použitý model popsat nemohl, nicméně poskytl zajímavý vhled do celého procesu:

„Základní trénovací data jsou data za minulé volby. Je dobré, že ČSÚ poskytuje zpětně i ty dávky, jak postupně výsledky zveřejňoval. Takže jsme si třeba mohli pouštět testy těch minulých voleb jakoby v reálném čase (což je dobré i na otestování frontendu a toho, co se bude dít kolem – např. oznámení o výsledcích voleb).“

„Nastavili jsme si i přibližné predikční intervaly u odhadů právě z odhadů minulých voleb – ty jsme použili např. na to vyhlášení výsledků voleb (např. už při druhé dávce bylo v druhém kole jasno, kdy Pavlův interval byl už celý nad 50% hranicí). V prvním kole jsme třeba měli velmi dobře odhadnuté středy, ale intervaly jsme měli trochu „zbytečně“ široké, takže ty jsme ještě předělávali pro 2. kolo. Ale to také vycházelo z toho, že jsme to celé dělali jako servis čtenářům a museli jsme brát třeba v potaz, aby tomu průměrný čtenář rozuměl. Převedeno do statistické řeči nás to tlačilo spíše k širším

intervalům než k užším, protože raději být uvnitř širšího (třeba 99%) intervalu než být u 10 % čísel mimo interval (u 90% intervalů).“

„Model jsme vždy překalibrovali na dané volby (testovací i ostré), takže modely pro 1. a 2. kolo se mírně lišily. Jinak modelů jsme měli několik. Měli jsme i jeden záložní. Pro případ, že by vypadlo napojení na data (což bylo nejrizikovější místo celého procesu), jsme měli záložní model, který vycházel jen z celkově sečtených výsledků a procenta sečtených okrsků v daný čas (protože to se dalo získat i odjinud a přepsat klidně ručně). Ten byl samozřejmě daleko méně přesný, ale v podstatě na druhé kolo stačil také, což jsme ale předem nevěděli. Naštěstí jsme ho nemuseli použít.“

6.3. STEM/MARK

V reakci na žádost o vyjádření zástupce společnosti STEM/MARK odkázal na tiskové zprávy [12], [11] a doplnil: *„Používáme dva nezávislé modely predikce: 1. clustery – pracujeme se skupinami volebních okrsků s podobnou volbou v předchozích několika volbách, při predikci se pak průběžné výsledky v clusterech převažují jejich potenciálem. 2. kalibrace – odhad vah sčítaných okrsků pro celkový výsledek na základě výsledku v předchozích několika volbách a následné převážení těmito vahami.“*

Tisková zpráva po prvním kole prezidentské volby 2023 [12] uvádí, že analytici společnosti upřednostňovali první uvedený postup založený na shlukové analýze, který byl také jako jediný popsán v obou tiskových zprávách. Z toho soudíme, že zveřejněné výsledky byly získány právě tímto postupem. Tisková zpráva po druhém kole prezidentské volby 2023 [11] tento postup shrnuje následovně: *„Jedním ze dvou hlavních použitých algoritmů je shluková analýza. Nejprve byly všechny volební okrsky rozděleny do 12 skupin neboli clusterů (o jeden méně než v 1. kole prezidentské volby před 14 dny) podle podobnosti volebních výsledků ve volbách v 1. kole. V každé skupině byl znám počet obyvatel, respektive počet voličů a s přihlédnutím k dřívější volební účasti tak bylo možné přiřadit dílčím skupinám odlišnou váhu. Posledním krokem predikce byl předpoklad, že okrsky, které patří do stejné skupiny, se budou i v letošních volbách chovat velmi podobně. Na základě průběžných výsledků v jednotlivých skupinách pak algoritmus dopočítal souhrnnou volební predikci na celostátní úrovni.“*

My takový postup můžeme interpretovat jako využití principu stratifikace podobně jako při stratifikaci podle krajů v odstavci 4.1, kde místo krajů uvažujeme jednotlivá strata jako sady volebních okrsků s podobným voličským chováním, získané pomocí shlukové analýzy. Je zde však rozdíl oproti stratům vytvořeným v odstavci 4.2, kde strata vytváříme z jednotlivých voličů a každý volební okrsek v sobě může mít zastoupeny voliče z různých strat. To může působit jako drobný rozdíl, ale koncepčně je velmi důležitý.

Pro úplnost zde doplníme konkrétní výsledky predikce, uvedené v tiskové zprávě [11]. V čase 14.31 při 6,66 % sečtených okrsků byl predikovaný celkový výsledek Petra Pavla 58 %, Andreje Babiše 42 %. Dále, v čase 14.41 při 21,11 % sečtených okrsků byl predikovaný celkový výsledek Petra Pavla 58,2 %, Andreje Babiše 41,8 %. Pohled na celkové výsledky (viz například tabulku 3) ukazuje, že tyto predikce byly velmi přesné. Tisková zpráva [11] pak dokonce uvádí: *„Opakovaná úspěšnost predikce STEM a STEM/MARK potvrdila schopnost statistiků matematickými modely přinést velmi brzy velmi přesné výsledky a odsoudila tzv. exit polly do propadliště dějin.“*

7. Závěr

V krátkosti jsme si představili princip stratifikovaných náhodných výběrů. Stratifikace umožňuje odhadovat (například) průměrnou hodnotu sledovaného znaku v dané konečné populaci přesněji než pomocí prostého náhodného výběru, pokud tuto populaci umíme rozdělit na několik částí, v nichž je chování jedinců homogennější než v celé populaci. Jde tedy o techniku velmi užitečnou v řadě praktických aplikací. Predikce výsledků sčítání voleb pak byla zajímavým uplatněním tohoto postupu, ale nebyla motivací k jeho zavedení. Pro zajímavost a porovnání jsme zmínili i další přístupy k predikci výsledků voleb, i když ne všechny oslovené společnosti nás nechaly nahlédnout pod pokličku.

Literatura

[1] Blindspot Solutions: *Předpověd výsledků 2. kola volby prezidenta ČR pomocí umělé inteligence* (2023). https://volby.blindspot.ai/

[2] CNN Prima NEWS: *Predikce CNN Prima NEWS znovu vyšla. Přesně předpověděla vítězství Petra Pavla* (2023). https://cnn.iprima.cz/predikce-cnn-prima-news-znovu-vysla-presne-predpovedela-vitezstvi-petra-pavla-199274

[3] Česká televize: *Prezidentské volby* (2023). https://www.ceskatelevize.cz/porady/15496675472-prezidentske-volby

[4] Český rozhlas: *Predikce prezidentských voleb* (2023). https://www.irozhlas.cz/volby/prezidentske-volby-2023/predikce

[5] Český statistický úřad: *Otevřená data pro volební výsledky* (2023). https://www.volby.cz/opendata/opendata.htm

[6] iDNES.cz: *Prezidentské volby 2023* (2023). https://www.idnes.cz/volby/prezidentske/2023

[7] McLachlan, G., Krishnan, T.: *The EM algorithm and extensions*. 2. vydání, Wiley, New York, 2008.

[8] PAQ Research: *Kdo bude prezidentem? PAQ predikuje z průběžných výsledků* (2023). https://www.paqresearch.cz/post/predikce-druhe-kolo

[9] Portál otevřených dat: *Výsledky voleb v ČR aneb když otevřená data fungují na 1** (2023). https://data.gov.cz/%C4%8Dl%C3%A1nky/v%C3%BDsledky-voleb-v-%C4%8Dr-aneb-kdy%C5%BE-otev%C5%99en%C3%A1-data-funguj%C3%AD-na-1

[10] Seznam Zprávy: *Kdo se stane prezidentem? Predikce výsledků podle Seznam Zpráv* (2023). https://www.seznamzpravy.cz/p/vysledky-voleb/2023/prezidentske-volby/kolo/2/predikce-odhad

[11] STEM/MARK: *I ve druhém kole prezidentských voleb predikce STEM/MARK a STEM přesně odhadla konečné výsledky* (2023). https://stemmark.cz/i-ve-druhem-kole-prezidentskych-voleb-predikce-stem-mark-a-stem-presne-odhadla-konecne-vysledky/

[12] STEM/MARK: *Predikce výsledků prezidentských voleb STEM/MARK a STEM znovu velmi přesná* (2023). https://stemmark.cz/predikce-vysledku-prezidentskych-voleb-stem-mark-a-stem-znovu-velmi-presna/

[13] Thompson, S. K.: *Sampling*. 2. vydání, Wiley, New York, 2002.