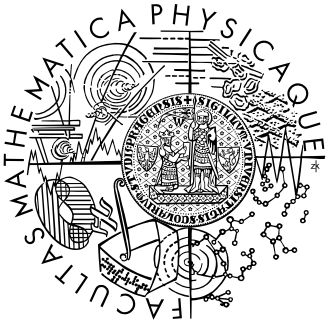
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Magdalena Zvejšková

Statistická chyba při reprezentativních výběrech z populace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Tomáš Hanzák
Studijní program: Matematika, Obecná matematika

2010

Na tomto místě bych chtěla poděkovat vedoucímu své bakalářské práce Mgr. Tomáši Hanzákovi za cenné rady a podněty, trpělivost při konzultacích a intenzivní spolupráci při tvorbě této bakalářské práce. Děkuji také svým rodičům, kteří mě vždy podporovali ve studiu.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 27. 5. 2010 Magdalena Zvejšková

Obsah

Úvod 5

1. Základní pojmy 7

1.1 Populace, výběrové šetření, statistická chyba 7
1.2 Metody výběru 8
1.3 Vážení dat 9

2. Teoretická odvození pro kvótní výběry, simulace pro vážená data 12

2.1 Kvótní výběry 12
2.1.1. Případ dvou kategorií 12
2.1.2. Případ více kategorií 15
2.2 Simulace pro vážení dat 18

3. Odhad statistické chyby pro volební modely 21

3.1 Průzkumy volebních preferencí 21
3.2 Vývoj volebních preferencí v ČR 23
3.3 Vyrovnávání časové řady 26
3.3.1. Výsledky pro agenturu MEDIAN 28
3.3.2. Výsledky pro agenturu STEM 32
3.4 Časové diference 35
3.5 Porovnání výsledků dvou agentur 37
3.6 Shrnutí 40

Závěr 42

Literatura 44

Přílohy 45

Název práce: Statistická chyba při reprezentativních výběrech z populace
Autor: Magdalena Zvejšková
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Tomáš Hanzák
E-mail vedoucího: tomas.hanzak@post.cz
Abstrakt: V práci se zabýváme určováním odhadu statistické chyby při výběrových šetřeních. Naším cílem je provést korekce odhadu této chyby v situacích, kdy se přistupuje k vážení dat nebo kdy data pocházejí z kvótního výběru. V případě kvótního výběru s jednou kvótní proměnnou pomocí teoretických úvah dospějeme ke zpřesněnému odhadu statistické chyby. Pro vážená data testujeme platnost upraveného odhadu simulacemi. Poté na reálná data týkající se volebních preferencí politických stran aplikujeme tři různé postupy, jejichž pomocí sestrojíme empirické odhady statistické chyby. Výsledky těchto metod vzájemně porovnáváme.
Klíčová slova: kvótní výběr, statistická chyba, vážení dat, volební model, výběrová šetření

Title: Statistical error in representative samples from population
Author: Magdalena Zvejšková
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Tomáš Hanzák
Supervisor's e-mail address: tomas.hanzak@post.cz
Abstract: In the thesis we deal with statistical error estimation in sampling surveys. The aim is to find corrections of statistical error estimations in the situations where we approach to data weighing or where the data originate from quota samples. Using theoretical considerations we deduce more accurate statistical error estimation in the case of quota sample with one quota variable. In the case of weighed data we test the validity of adjusted estimate using simulations. After that we apply three different methods to the real poll model data and construct empirical error estimates. The results of all three mentioned methods are being compared.
Keywords: quota sample, statistical error, data weighing, polls on voting intention, sampling surveys

Úvod

„Znám tři druhy lží. Velkou lež, malou lež a statistiku.“

George Bernard Shaw (1856–1950)

Tento citát nám demonstruje, jak o statistice smýšlel G. B. Shaw, držitel Nobelovy ceny za literaturu. I 60 let po jeho smrti se běžně setkáváme s lidmi zastávajícími stejný nebo podobný názor.

Lidé, kteří se nikdy blíže nezabývali studiem tohoto oboru, se pravděpodobně nejčastěji se statistikou setkávají při prezentaci výsledků výběrových šetření v médiích. Uveřejněná data potom považují za „pevná“ čísla, přičemž již nevěnují pozornost tomu, že výsledky průzkumů jsou zatíženy chybou. Nutno dodat, že tato chyba v mnoha případech není v médiích udávána.

Použijme jako příklad průzkumy volebních, resp. stranických preferencí v České republice. Pomineme-li fakt, že různé agentury zabývající se touto problematikou udávají za stejná období různé výsledky (což může být způsobeno statistickou chybou, ale také třeba metodologií šetření), nedůvěra veřejnosti ve výběrová šetření pak může být snadno posílena porovnáním dat z těchto šetření se skutečnými výsledky voleb. Průzkumy volebních preferencí ale vyjadřují pouze aktuální podporu politických stran, a tak se přesnost těchto odhadů odvíjí od časové vzdálenosti data provedení průzkumu a data voleb. Navíc je i zde třeba si uvědomit, že každé statistické šetření je zatíženo chybami, jež dělíme na statistické a systematické (viz kapitola 1.1). A právě odhadem statistické chyby při výběrových šetřeních se budeme v této práci zabývat.

Je třeba zdůraznit, že tato práce si neklade za cíl zvýšit statistickou gramotnost veřejnosti či její povědomí o statistické chybě, nicméně je motivována nedůvěrou veřejnosti ve statistické výsledky, a to zejména u volebních průzkumů, dále podceňováním vlivu statistické chyby, a také stále rostoucím počtem prováděných výběrových šetření, kde by při zpracovávání a interpretaci dat mohly být uplatněny postupy uvedené v následujících kapitolách.

V této práci se pokusíme na výsledky výběrových šetření aplikovat matematické postupy tak, abychom u výběrů, kde byly použity kvóty nebo vážení, dosáhli lepších odhadů statistické chyby, než za použití běžně používaných klasických vzorců známých z matematické statistiky určených pro prostý náhodný výběr. Tyto odvozené metody by pak měly být použitelné v praxi, díky čemuž by agentury mohly prezentovat své výsledky výběrových šetření s vhodnějším, a jak uvidíme, tak i nižším odhadem statistické chyby.

Nejdříve se v první kapitole seznámíme se základními pojmy z oblasti výběrových šetření, jež budeme v práci používat.

Ve druhé kapitole této práce se nejprve zaměříme na teoretické odvození zpřesněného odhadu statistické chyby při kvótních výběrech z populace, jenž bude nižší než odhad spočtený pomocí klasického vzorce. Pomocí simulací pak ukážeme, že u vážených dat lze dosáhnout podobného zpřesnění.

Ve třetí stěžejní kapitole se budeme zabývat výslednými daty výběrových šetření volebních preferencí od dvou různých agentur, na něž budou aplikovány tři různé postupy, jimiž se pokusíme empiricky odhadnout statistickou chybu u těchto šetření a porovnat výsledky se zjištěním z kapitoly 2.

Dále k práci přikládáme CD, kde lze nalézt kopii tohoto textu, program se simulacemi, se kterými budeme pracovat ve druhé kapitole, a jednotlivé výpočty související s odhady ve třetí kapitole.

Kapitola 1

Základní pojmy

V této kapitole vysvětlíme některé základní pojmy, se kterými budeme v dalších částech této práce pracovat. Přitom předpokládáme, že čtenář je seznámen se základními pojmy matematické statistiky.

1.1 Populace, výběrové šetření, statistická chyba

Pod pojmem *populace* neboli *základní soubor* rozumíme skupinu jednotek (např. osob, domácností), jež je vymezena stanovením společných vlastností. Přitom musí být zřejmé, zda jedinec do populace patří, či ne. Populaci tvoří například občané České republiky starší 18-ti let. Je patrné, že desetiletý chlapec do této populace nepatří.

Konečná populace je taková populace, která je tvořena konečným počtem jednotek. Takovou populaci můžeme ztotožnit s množinou $\{1, \ldots, N\}$, $N \in \mathbb{N}$. Jakoukoli podmnožinu populace (základního souboru) nazýváme *výběrový soubor* či krátce *výběr*. Takových různých výběrových souborů lze utvořit 2^N, viz [11].

Znaky populace zkoumáme tzv. *statistickým šetřením*. Jestliže při statistickém šetření studujeme vlastnosti každé jednotky populace (základního souboru), provádíme *úplné šetření*. To však zejména u populací velkých rozsahů (se statisíci nebo miliony jednotek) není z důvodu organizační náročnosti a vysokých finančních nákladů ve většině případů uskutečnitelné. Proto přistupujeme k *výběrovému šetření*, kdy prošetřujeme pouze vlastnosti výběru a výsledky šetření se pak snažíme zobecnit na základní soubor, jak je vysvětleno např. v [13].

Vzhledem k tomu, že se v této práci budeme zabývat pouze takovými výběry, kdy populaci tvoří lidé a výběrové šetření probíhá tak, že vybraní jedinci odpovídají na položené otázky, na základě čehož se pak tvoří odhady pro celou populaci, budeme jednotky výběrového rozsahu nazývat *respondenty*. Podmnožinám základního a výběrového souboru budeme říkat *skupiny* či *kategorie*.

Aby bylo možné provést zobecnění výsledků výběrového šetření na základní soubor, je třeba, aby výběr odrážel poměry v základním souboru, což přesněji vystihuje pojem reprezentativity. Řekneme, data získaná výběrovým šetřením (resp. výběr) jsou

reprezentativní, jestliže se popisné charakteristiky výběru až na náhodou chybu shodují s charakteristikami základního souboru, viz [12].

Poznamenejme, že mnozí sociologové, kteří také často provádí výběrová šetření, většinou považují za reprezentativní výhradně prostý náhodný výběr (viz kapitola 1.2), jak plyne z textu [9]. Ve skutečnosti nám ale naopak např. kvótní výběry pomáhají reprezentativity dosáhnout.

I v případě, že jsou výsledky výběrového šetření reprezentativní, nese jejich zobecnění na základní soubor jistou míru nejistoty ohledně správnosti tohoto zobecnění. Říkáme, že zobecnění je zatíženo *statistickou chybou*. Statistická chyba je chyba náhodná, nepravidelná a její vliv na výsledek statistického měření neumíme odstranit, avšak pomocí postupů matematické statistiky je možné tuto chybu odhadnout.

Statistickou chybu standardně odhadujeme výpočtem *směrodatné odchylky*, tj. odmocniny z rozptylu odhadu, kterou přenásobíme příslušným kvantilem, typicky kvantilem normálního rozdělení. Klasický a běžně používaný vzorec pro výpočet směrodatné odchylky nestranného a konzistentního odhadu $\hat{\theta}$ parametru alternativního rozdělení θ (tj. odhadu procenta výskytu nějakého jevu v populaci) vypadá takto:

$$\hat{\sigma} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}. \tag{1}$$

Odhad statistické chyby pak vypočteme jako $\hat{\sigma}u\left(\frac{\alpha}{2}\right)$, kde $u\left(\frac{\alpha}{2}\right)$ je kvantil normálního rozdělení a $n \in \mathbb{N}$ je počet jedinců ve výběrovém souboru. Typicky volíme $\alpha = 0{,}05$, a tedy $u\left(\frac{\alpha}{2}\right) = 1{,}96$. Tvar vzorce (1) vyplývá z předpokladu prostého náhodného výběru z dostatečně velké populace, centrální limitní věta navíc zajišťuje normalitu rozdělení statistické chyby, (obvykle se požaduje, aby $n\theta(1-\theta) > 10$, viz [2]).

Výběr, který není reprezentativní, je zatížen *systematickou chybou*, čímž rozumíme chybu nenáhodnou, která je během opakovaného měření způsobena stále stejnou příčinou. Vzniká například při použití nesprávné metody dotazování nebo při nevhodné volbě výběrového souboru. Použitím správných metod lze tuto chybu značně omezit či úplně eliminovat, viz [13]. Systematickou chybou se však v této práci přímo zabývat nebudeme. Soustředíme se na druhou složku chyby, statistickou chybu.

1.2 Metody výběru

Poznamenejme nejprve, že výběrem v tomto odstavci chápeme způsob volby výběrového souboru. Ten volíme na základě *pravděpodobnostního výběru*, kdy je každé jednotce základního souboru přiřazena nenulová pravděpodobnost, že bude tato jednotka vybrána do výběrového souboru, a dále kovariance zahrnutí každé dvojice jednotek. Pravděpodobnostní výběry dělíme na výběry náhodné a nenáhodné.

Náhodný výběr je takový výběr, kdy je výběrový soubor volen zcela náhodně a nezávisle na našem úsudku. Takový výběr reprezentuje známé i neznámé vlastnosti populace.

Jedním z druhů náhodného výběru je *prostý náhodný výběr*, čímž rozumíme náhodný výběr bez vracení. To znamená, že provádíme náhodný výběr, a pokud již některá jednotka byla vybrána, nemůže být zvolena znovu. U prostého náhodného výběru má každá *n*-tice stejnou pravděpodobnost, že se stane výběrovým souborem, a to

$$P = \frac{1}{\binom{N}{n}},$$

kde $N \in \mathbb{N}$ je velikost základního souboru a $n \leq N$, $n \in \mathbb{N}$, je rozsah výběrového souboru, viz [11] a [13]. Realizace takového výběru pro průzkumy veřejného mínění je však často nemožná, už jen z toho důvodu, že náhodně vybraný jedinec může odmítnout odpovědět na položenou otázku.

Kromě jiných mezi náhodné výběry dále řadíme *náhodný stratifikovaný výběr*, kdy je populace rozdělena do skupin homogenních podle nějakého kritéria, např. podle regionu. Jedinci jsou potom do výběrového souboru vybíráni náhodně podle těchto skupin.

Při *nenáhodných výběrech* naopak upřednostňujeme výběr některých jedinců před ostatními, což může být naším záměrem, nebo nezamýšleným důsledkem zvoleného způsobu získávání dat.

Jakousi kombinací náhodných a nenáhodných výběrů jsou praxi velmi oblíbené *kvótní výběry*. Pro kvótní výběry je charakteristické, že záměrně kopírují strukturu známých vlastností v základním souboru dle zvolených charakteristik. Tedy pokud máme dostatečné informace o struktuře populace, volíme výběr z této populace tak, aby podíl jednotek dané vlastnosti byl ve výběrovém souboru totožný s podílem v základním souboru.

Při výběrových šetřeních je pak kvótní výběr realizován tak, že tazatel náhodně vybírá tzv. *kvótami* určený počet respondentů s danými znaky, jimiž nejčastěji bývají regionální (kraj, obec) a socio-demografické charakteristiky (pohlaví, věk, vzdělání). Počty respondentů s předepsanými vlastnosti odrážejí strukturu populace a v České republice se nejčastěji určují z veřejně dostupných dat Českého statistického úřadu (ČSÚ), viz [5].

1.3 Vážení dat

Výsledky každého výběrového šetření prováděného dotazováním jsou vždy zatíženy systematickou chybou, neboť kromě toho, že respondenti mohou odpovídat nepravdivě, nelze nikdy zaručit, že žádná z dotazovaných osob neodmítne poskytnout

odpověď. Pro snížení této systematické chyby lze kromě realizace výběru pomocí kvót použít *vážení dat*, tedy přiřazení *vah* w_i, jednotlivcům (označme je $i = 1, 2,\ldots, n$) ve výsledném výběru (o rozsahu $n \in \mathbb{N}$). Vážení pak vyrovná převahu určité skupiny ve výsledném výběru oproti skutečnému poměru v základním souboru, jestliže se zkoumaná charakteristika u této skupiny liší od zbytku populace.

Uvažujme případ, kdy chceme odhadnout relativní četnost jedinců s určitou vlastností v základním souboru o velikosti $N \in \mathbb{N}$, např. poměr kuřáků starších 18-ti let v České republice. Při výběrovém šetření bylo dotázáno n respondentů, $n \leq N$, $n \in \mathbb{N}$, jenž tvoří výběrový soubor, přičemž každý odpověděl ano, nebo ne, podle toho, zda je kuřák, nebo nekuřák. Tuto situaci můžeme popsat pomocí náhodných veličin Y_i, $i = 1, 2,\ldots, n$, které nabývají pouze hodnot 1 (jestliže je i-tý respondent kuřák), nebo 0 (pokud i-tý respondent není kuřák).

Bez použití vah bychom hledaný poměr odhadli výběrovým průměrem:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Za použití vah relativní četnost kuřáků odhadujeme pomocí *váženého (výběrového) průměru*:

$$\bar{Y}_w = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i Y_i. \tag{2}$$

V případě, že jednice rozlišujeme pouze na základě jedné vážící proměnné, vypočteme váhy w_i podle následujícího klíče:

$$\frac{očekávaný\ podíl}{zjištěný\ podíl}, \tag{3}$$

kde očekávaný podíl značí podíl jedinců v základním souboru patřících do téže skupiny (např.) jako i-tý respondent a zjištěný podíl značí podíl jedinců ve výběrovém souboru patřících do této skupiny. Jsou-li všechny váhy rovny jedné, je zřejmě vážený průměr totožný s výběrovým průměrem. Pro určení vah při více vážících proměnných je třeba volit složitější postup, jenž je popsán např. v [3].

Vážení dat je vhodné použít v případě, že odpovědi respondentů (či jejich ochota odpovídat) jsou silně korelované se socio-demografickými znaky populace. Uvažujme například, že chceme zjistit, zda respondent pěstuje pokojové květiny. Kromě odpovědí na tuto otázku budeme v průběhu šetření zaznamenávat, zda je respondent muž, či žena. Po provedení sběru dat zjistíme, že ženy odpovídali kladně častěji než muži (např. i proto, že ženy odpovídaly ochotněji). Zároveň ale poměr mužů a žen ve výběrovém souboru neodpovídá poměru v základním souboru. Protože poměr kladných odpovědí závisí na pohlaví respondenta, má smysl použít vážení dat, čímž můžeme snížit

systematickou chybu. Později ukážeme, že tak lze jistou měrou snížit i statistickou chybu.

Pokud bychom však použili vážení dat nevhodně, může se chyba naopak zvýšit. Kdybychom v předchozí situaci namísto pohlaví zjišťovali datum narození respondenta a poté vážili data, podle toho, ve kterém měsíci se respondenti narodili, pravděpodobně bychom si tím spíše uškodili. Jistě totiž pěstování květin nesouvisí s datem narození.

Vážení dat je tím korektnější, čím jsou hodnoty w_i blíže k jedné. Představme si, že v kategorii „narozen 8. srpna“ je pouze jeden respondent a my mu přiřadíme váhu velikosti 5. Pak jeho odpověď zastupuje pět hypotetických respondentů, jenž patří do stejné kategorie. Zakládáme tedy poměr kladných odpovědí pro všechny osoby v populaci narozené 8. srpna na odpovědi jednoho respondenta, což není příliš korektní, viz [2].

Jak bylo řečeno, vážení dat (často spolu s kvótami) se používá primárně ke snížení systematické chyby. V následující kapitole ukážeme, že lze využít i ke zpřesnění odhadu statistické chyby.

Kapitola 2

Teoretická odvození pro kvótní výběry, simulace pro vážená data

V této kapitole se budeme zajímat o taková výběrová šetření, kdy chce agentura odhadnout relativní četnost nějakého znaku v populaci, a to tak, že je respondentům položena otázka, zda tuto vlastnost mají, nebo nemají. Na tuto otázku je tedy možná odpověď ano, nebo ne a na základě těchto odpovědí se agentura snaží poměr odhadnout.

Mají-li agentury zabývající se výběrovými šetřeními informace o struktuře populace (např. z dat ČSÚ), mohou přikročit ke kvótním výběrům nebo vážení dat, čímž lze dosáhnout snížení systematické chyby. My zkusíme využít kvót a vážení pro snížení odhadu statistické chyby. Konkrétně se budeme zajímat a odhad poměru kladných odpovědí (označme jej p) pro základní soubor a zejména pak o rozptyl tohoto odhadu (a tudíž o odhad statistické chyby).

2.1 Kvótní výběry

2.1.1. Případ dvou kategorií

Přestavme si, že agentura provádí výběrové šetření na základě kvótního výběru, kde respondenty posuzuje podle jedné kvótní proměnné. Tato proměnná (např. pohlaví respondenta) rozčlení výběrový, resp. základní soubor do dvou (disjunktních) kategorií. Předpokládejme, že agentura přitom dokáže předem dané kvóty splnit tak, že struktura výběrového souboru přesně odpovídá struktuře základního souboru.

Popišme kategorie v základním souboru o rozsahu $N \in \mathbb{N}$: označme je A_0 a A_1 a jejich velikost $|A_0| = N_0$, $|A_1| = N_1$, $N_0 + N_1 = N$. Ve výběrovém souboru také rozlišíme obě kategorie: označíme je a_0 a a_1 a jejich velikost $|a_0| = n_0$, $|a_1| = n_1$, $n_0 + n_1 = n$, kde $n \in \mathbb{N}$ značí velikost výběrového souboru. Platí $a_0 \subset A_0$, $a_1 \subset A_1$.

Celý výběrový soubor můžeme považovat za náhodný výběr, neboť respondenti byli tazateli vybíráni náhodně. Na základě jejich odpovědí můžeme odhadnout poměr kladných odpovědí p výběrovým průměrem.

Uvažujme náhodné veličiny Y_1, Y_2,…, Y_n, jež nabývají pouze hodnot 1, odpoví-li respondent kladně, nebo 0, odpoví-li záporně. Výběrový průměr tedy spočteme takto:

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Dále lze říci, že respondenti v rámci každé kategorie byli také vybíráni náhodně. Odhad poměru kladných odpovědí pro každou kategorii zvlášť tedy můžeme opět odhadnout výběrovým průměrem:

$$\hat{p}_0 := \frac{1}{n_0}\sum_{i\in a_0} Y_i,$$
$$\hat{p}_1 := \frac{1}{n_1}\sum_{i\in a_1} Y_i.$$

Přitom odhady $\hat{p}_0$, $\hat{p}_1$ jsou nezávislé. Nyní můžeme pomocí těchto odhadů upravit zápis výběrového průměru pro celý soubor:

$$\hat{p} = \frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}.$$

Jelikož naším cílem je studovat statistickou chybu tohoto odhadu, zaměřme se nyní na odhad rozptylu $\hat{p}$. Ten lze vyjádřit dvojím způsobem, jednak tak, že uvažujeme $\hat{p}$ jako odhad platný pro celý základní soubor bez ohledu na kategorie:

$$\widehat{\operatorname{var}}\,\hat{p} := \frac{\hat{p}(1-\hat{p})}{n} = \frac{1}{n}\left(\frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}\right)\left(1 - \frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}\right) =$$
$$= \frac{1}{n}\left(\frac{n_0}{n}\hat{p}_0 + \frac{n_1}{n}\hat{p}_1\right)\left[\frac{n_0}{n}(1-\hat{p}_0) - \frac{n_1}{n}(1-\hat{p}_1)\right],$$

jednak tak, že $\hat{p}$ chápeme jako lineární kombinaci odhadů $\hat{p}_0$, $\hat{p}_1$ pro jednotlivé kategorie základního souboru:

$$\widetilde{\operatorname{var}}\,\hat{p} := \operatorname{var}\left(\frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}\right) = \left(\frac{n_0}{n}\right)^2 \operatorname{var}\hat{p}_0 + \left(\frac{n_1}{n}\right)^2 \operatorname{var}\hat{p}_1$$
$$= \left(\frac{n_0}{n}\right)^2 \frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \left(\frac{n_1}{n}\right)^2 \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}.$$

Než přistoupíme k dalšímu kroku, shrňme ještě základní fakta týkající se naší situace v Tabulce 1, kde ZS značí základní soubor a VS výběrový soubor a odhady rozptylu pro každou kategorii jsou vypočteny podle klasického vzorce pro rozptyl odhadu parametru alternativního rozdělení.

Tabulka 1: Struktura základního a výběrového souboru pro dvě kategorie respondentů

	Rozsah – ZS	Rozsah – VS	Odhad p	Odhad rozptylu
A_0	N_0	n_0	$\hat{p}_0$	$\frac{\hat{p}_0(1-\hat{p}_0)}{n_0}$
A_1	N_1	n_1	$\hat{p}_1$	$\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$
Celkem	N	n	$\hat{p}$	$\widehat{\text{var}}\,\hat{p}$, resp. $\widetilde{\text{var}}\,\hat{p}$

Nyní tedy máme dva různé odhady rozptylu téhož výběrového průměru a chceme je mezi sebou porovnat. Budeme tedy zkoumat jejich podíl. Konkrétně dokážeme, že platí:

$$\frac{\widetilde{\text{var}}\,\hat{p}}{\widehat{\text{var}}\,\hat{p}} = 1 - \text{R}^2 \in [0,1]. \tag{4}$$

R^2 nazýváme *koeficient determinace* a vyjadřujeme jím přesnost regresního modelu. V tomto případě se jedná o koeficient determinace lineárního regresního modelu, kde 0-1 vysvětlovanou proměnnou Y vysvětlujeme na základě jedné kategoriální-kvótní proměnné a model odhadujeme metodou nejmenších čtverců.

Koeficient determinace určuje, jaké procento změn vysvětlované proměnné je vysvětleno odhadnutým modelem. $\text{R}^2 \in [0,1]$ a čím je blíže jedné, tím je model vhodnější. Obecný vzorec pro výpočet koeficientu determinace je

$$\text{R}^2 = 1 - \frac{S_e}{S_T} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

kde y_i jsou realizace náhodných veličin Y_i, $\hat{y}_i$ je odhad y_i regresním modelem a $\bar{y}$ značí průměrnou hodnotu y_i, $i = 1, 2, \ldots, n$. Veličinu S_e nazýváme *reziduální součet čtverců*, veličinu S_T *totální součet čtverců*, viz [1].

Dosaďme nejprve do levé strany rovnosti (4):

$$\frac{\widetilde{\text{var}}\,\hat{p}}{\widehat{\text{var}}\,\hat{p}} = \frac{\left(\frac{n_0}{n}\right)\hat{p}_0(1-\hat{p}_0) + \left(\frac{n_1}{n}\right)\hat{p}_1(1-\hat{p}_1)}{\left(\frac{n_0}{n}\hat{p}_0 + \frac{n_1}{n}\hat{p}_1\right)\left[\frac{n_0}{n}(1-\hat{p}_0) - \frac{n_1}{n}(1-\hat{p}_1)\right]}.$$

Pro pravou stranu spočteme nejprve S_e a poté S_T z definice R^2, kde $y_i \in \{0, 1\}$, $\bar{y} = \hat{p}$ a $\hat{y}_i \in \{\hat{p}_0, \hat{p}_1\}$, neboť použitím metody nejmenších čtverců odhadneme Y v závislosti na kategorii právě výběrovými průměry $\hat{p}_0$, $\hat{p}_1$.

$$S_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n_0}(y_i - \hat{p}_0)^2 + \sum_{i=n_0+1}^{n}(y_i - \hat{p}_1)^2 =$$
$$= n_0\hat{p}_0(1-\hat{p}_0)^2 + n_0(1-\hat{p}_0)(-\hat{p}_0)^2 + n_1\hat{p}_1(1-\hat{p}_1)^2 + n_1(1-\hat{p}_1)(-\hat{p}_1)^2 =$$
$$= n\left[\left(\frac{n_0}{n}\right)\hat{p}_0(1-\hat{p}_0) + \left(\frac{n_1}{n}\right)\hat{p}_1(1-\hat{p}_1)\right],$$

$$S_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{p})^2 = n\hat{p}(1-\hat{p})^2 + n(1-\hat{p})(-\hat{p})^2 =$$
$$= n\hat{p}(1-\hat{p}) = n\left(\frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}\right)\left(1 - \frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n}\right) =$$
$$= n\left(\frac{n_0}{n}\hat{p}_0 + \frac{n_1}{n}\hat{p}_1\right)\left[\frac{n_0}{n}(1-\hat{p}_0) - \frac{n_1}{n}(1-\hat{p}_1)\right].$$

Dosazením do pravé strany rovnosti (4) dostaneme požadovanou rovnost:

$$1 - \mathrm{R}^2 = \frac{n\left[\left(\frac{n_0}{n}\right)\hat{p}_0(1-\hat{p}_0) + \left(\frac{n_1}{n}\right)\hat{p}_1(1-\hat{p}_1)\right]}{n\left(\frac{n_0}{n}\hat{p}_0 + \frac{n_1}{n}\hat{p}_1\right)\left[\frac{n_0}{n}(1-\hat{p}_0) - \frac{n_1}{n}(1-\hat{p}_1)\right]} = \frac{\widetilde{\mathrm{var}}\,\hat{p}}{\widehat{\mathrm{var}}\,\hat{p}}.$$

Protože $\mathrm{R}^2 \in [0,1]$, vidíme z právě dokázané identity, že odhad $\widetilde{\mathrm{var}}\,\hat{p}$ rozptylu výběrového průměru je nejvýše tak velký jako odhad $\widehat{\mathrm{var}}\,\hat{p}$ rozptylu téhož průměru, neboli máme: $0 \leq \tilde{\sigma}_{\hat{p}} \leq \hat{\sigma}_{\hat{p}}$, kde $\tilde{\sigma}_{\hat{p}}$ je odmocnina z $\widetilde{\mathrm{var}}\,\hat{p}$ a $\hat{\sigma}_{\hat{p}}$ je rovno odmocnině z $\widehat{\mathrm{var}}\,\hat{p}$. Pro danou situaci jsme tedy našli nižší odhad statistické chyby.

2.1.2. Případ více kategorií

Předchozí případ zobecníme na situaci, kdy agentura provádí výběrové šetření na základě kvótního výběru podle jedné kvótní proměnné, jež na rozdíl od odstavce 2.1.1 rozdělí výběrový i základní soubor na K, $K \in \mathbb{N}$, (disjunktních) kategorií. Tyto kategorie v základním souboru označme $A_1, A_2, \ldots, A_K$ a ve výběrovém souboru $a_1, a_2, \ldots, a_K$, přičemž $|A_i| = N_i \in \mathbb{N}$, $|a_i| = n_i \in \mathbb{N}$, $a_i \subset A_i$, $i = 1, 2, \ldots, K$, a $\sum_{i=1}^{K} N_i = N$, $\sum_{i=1}^{K} n_i = n$.

Takovou kvótní proměnnou může být např. nejvyšší dosažené vzdělání, kdy bychom respondenty řadili do kategorie „základní“, „vyučeni“, „středoškolské s maturitou“, nebo „vysokoškolské“.

Při průzkumu je respondentům opět položena otázka, na kterou mohou odpovědět ano, nebo ne (např. zda v posledních dvou letech strávili dovolenou v zahraničí). Cílem agentury je i zde na základě odpovědí respondentů odhadnout poměr kladných odpovědí

p pro celý základní soubor, což lze opět provést výběrovým průměrem. Postupujme zde obdobně jako v minulém odstavci.

Definujeme-li náhodné veličiny Y_1, Y_2,…, Y_n analogicky jako v odstavci 2.1.1, pak zde bude výběrový průměr téhož tvaru:

$$\hat{p} := \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

Soustřeďme se na odhad poměru kladných opovědí pouze v rámci jednotlivých kategorií. Pro *i*-tou kategorii, $i = 1, 2,\ldots, K$, má výběrový průměr tento tvar:

$$\hat{p}_i := \frac{1}{n_i}\sum_{j \in a_i} Y_j.$$

Protože i zde odpovídá struktura výběru struktuře populace, tj. $\frac{n_i}{n} = \frac{N_i}{N}$, můžeme tedy opět takto vyjádřit $\hat{p}$ pomocí $\hat{p}_i$:

$$\hat{p} = \sum_{i=1}^{K} \frac{n_i}{n}\hat{p}_i.$$

Nyní se zaměřme na odhad rozptylu $\hat{p}$. I zde jej můžeme vyjádřit dvojím způsobem:

$$\widehat{\operatorname{var}}\,\hat{p} := \frac{\hat{p}(1-\hat{p})}{n} = \frac{1}{n}\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\right]\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)(1-\hat{p}_i)\right],$$

$$\widetilde{\operatorname{var}}\,\hat{p} := \sum_{i=1}^{K} \operatorname{var}\frac{n_i}{n}\hat{p}_i = \sum_{j=1}^{K}\left(\frac{n_i}{n}\right)^2 \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}. \qquad (6)$$

Pro přehlednost opět shrňme informace z tohoto odstavce v Tabulce 2:

Tabulka 2: Struktura základního a výběrového souboru pro *K* kategorií respondentů

	Rozsah – ZS	Rozsah – VS	Odhad *p*	Odhad rozptylu
A_i	N_i	n_i	$\hat{p}_i$	$\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$
Celkem	N	n	$\hat{p}$	$\widehat{\operatorname{var}}\,\hat{p}$, resp. $\widetilde{\operatorname{var}}\,\hat{p}$

Dokážeme, že i v tomto obecnějším případě platí:

$$\frac{\widetilde{\text{var}}\,\hat{p}}{\widehat{\text{var}}\,\hat{p}} = 1 - \text{R}^2 \in [0,1].$$

Na levé straně dostáváme:

$$\frac{\widetilde{\text{var}}\,\hat{p}}{\widehat{\text{var}}\,\hat{p}} = \frac{\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i(1-\hat{p}_i)}{\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\right]\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)(1-\hat{p}_i)\right]}.$$

Pro výpočet pravé strany nejprve podle vzorce (5) vyjádříme S_e a S_T, kde $\bar{y} = \hat{p}$, $y_j \in \{0,1\}$ a, analogicky jako v kapitole 2.1, $\hat{y}_j \in \{\hat{p}_i, i = 1, 2, \dots, K\}$:

$$S_e = \sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2 = \sum_{i=1}^{K}\sum_{j\in a_i}\left(y_j - \hat{p}_i\right)^2 =$$
$$= \sum_{i=1}^{K}[n_i\hat{p}_i(1-\hat{p}_i)^2 + n_i(1-\hat{p}_i)(-\hat{p}_i)^2] = n\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\,(1-\hat{p}_i),$$

$$S_T = \sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2 = n\hat{p}(1-\hat{p})^2 + n(1-\hat{p})(-\hat{p})^2 = n\hat{p}(1-\hat{p}) =$$
$$= n\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\right]\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)(1-\hat{p}_i)\right].$$

Dostáváme tak

$$1 - \text{R}^2 = \frac{n\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\,(1-\hat{p}_i)}{n\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\hat{p}_i\right]\left[\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)(1-\hat{p}_i)\right]} = \frac{\widetilde{\text{var}}\,\hat{\text{p}}}{\widehat{\text{var}}\,\hat{p}},$$

tedy pravá strana je totožná s levou, a tudíž rovnost, kterou jsme chtěli ukázat, platí. I v tomto obecnějším případě je odhad rozptylu výběrového průměru $\widetilde{\text{var}}\,\hat{p}$ nejvýše tak velký jako odhad $\widehat{\text{var}}\,\hat{p}$. Po odmocnění $\widetilde{\text{var}}\,\hat{p}$ a vynásobením příslušným kvantilem tedy dostáváme nižší odhad statistické chyby.

Ukázali jsme, že v uvedených konkrétních situacích lze nalézt lepší odhad rozptylu, a tedy i statistické chyby, využijeme-li faktu, že data pocházejí z kvótního výběru.

Přitom v praxi není nutné počítat výběrové průměry pro každou kategorii – lze postupovat i tak, že se pomocí základního vzorce (5) spočte koeficient determinace. Pro použití tohoto vzorce je však třeba zjistit hodnoty $\hat{y}_j$, a to např. lineární, logistickou či jinou regresí, případně jinými vhodnými metodami, kde je Y vysvětlovanou

proměnnou a kvótní proměnné vysvětlujícími proměnnými. Potom stačí spočíst odhad rozptylu klasickým vzorcem a tento odhad vynásobit hodnotou $1 - R^2$.

Pokud bychom tímto způsobem chtěli upravit statistickou chybu i v případě, že členíme základní soubor podle více kvótních proměnných (tj. znaků, jejichž strukturu v souboru známe), můžeme si jednotlivé kategorie, které jsou jednoznačně určeny skladbou vlastností respondentů, představit např. jako listy regresního stromu. Listy pak reprezentují K, $K \in \mathbb{N}$, disjunktních kategorií. Intuice nám potom napovídá, že by mělo být možné postupovat analogicky jako v této kapitole, tj. v rámci každé kategorie můžeme opět odhadnout poměr kladných odpovědí a na základě těchto odhadů spočítat odhad rozptylu výběrového průměru. Vztah se zde ale pro tuto situaci již nebudeme snažit teoreticky odvodit.

2.2 Simulace pro vážení dat

Kromě kvótních výběrů se v praxi velmi často přistupuje k vážení dat. I v tomto případě se můžeme ptát, zda nelze odhad statistické chyby upravit podobně jako pro data získaná na základě kvótního výběru v kapitole 2.1.

Uvažme tedy situaci, kdy agentura při výběrovém šetření vybírá respondenty náhodně, zároveň ale u výběrového i základního souboru rozlišuje kategorie dané (jednou) vážící proměnnou, což může být opět např. pohlaví, věk, kraj, nejvyšší dosažené vzdělání. Rozdíl oproti předchozí situaci spočívá v tom, že tazatelům nejsou určeny kvóty, kolika respondentů z každé kategorie se mají dotázat. Struktura výběru se tak může značně lišit od struktury základního souboru.

Agentura pak za účelem snížení systematické chyby spočítá odhad poměru kladných odpovědí namísto výběrového průměru váženým výběrovým průměrem.

Zachovejme značení z odstavce 2.1.2. Pouze budeme místo kvótní proměnné používat vážící proměnnou. Pak podle vzorce (2) bude výběrový průměr tohoto tvaru

$$\hat{p}_w := \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i Y_i,$$

kde příslušné váhy spočteme podle vyjádření (3). Je zřejmé, že váhy budou vždy pro všechny respondenty z téže kategorie stejné, a to

$$w_i := \frac{N_i}{N} \frac{n}{n_i},$$

$i = 1, 2, \ldots, K$. Lze tedy psát

$$\hat{p}_w = \frac{\sum_{i=1}^{K} n_i w_i \hat{p}_i}{\sum_{i=1}^{K} n_i w_i} = \frac{\sum_{i=1}^{K} n_i \frac{N_i}{N}\frac{n}{n_i}\hat{p}_i}{\sum_{i=1}^{K} n_i \frac{N_i}{N}\frac{n}{n_i}} = \sum\nolimits_{i=1}^{K} \left(\frac{N_i}{N}\right)\hat{p}_i,$$

$$\widehat{\operatorname{var}}\,\hat{p}_w := \sum\nolimits_{i=1}^{K} \operatorname{var}\frac{N_i}{N}\hat{p}_i = \sum\nolimits_{j=1}^{K} \left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}.$$

Všimněme si, že tento odhad rozptylu se příliš neliší od (6). Pokud by se i zde jednalo o kvótní výběr, budou váhy w_i, $i = 1, 2,\ldots, K$, rovny jedné a oba vzorce se budou shodovat. To nás přivádí k domněnce, že i zde by mohl platit vztah

$$\frac{\widehat{\operatorname{var}}\,\hat{p}_w}{\widehat{\operatorname{var}}\,\hat{p}} = 1 - \mathrm{R}^2.$$

Dokázat tuto rovnost by již zřejmě bylo složitější, a proto budeme testovat platnost tohoto vztahu simulacemi. Soubor s naprogramovanými simulacemi (Simulace pro vážení dat.xls) lze nalézt na přiloženém CD. Program zde nyní stručně popíšeme.

Označme nejprve $C := \frac{\widehat{\operatorname{var}}\,\hat{p}_w}{\widehat{\operatorname{var}}\,\hat{p}}$ a nazvěme ho *redukční koeficient*. Tento koeficient budeme porovnávat s hodnotou $1 - \mathrm{R}^2$.

Soustřeďme se ale v programu nejprve na žlutě vybarvená pole. Hodnoty v těchto polích lze volit. V poli G2 zadáváme počet respondentů, tedy velikost výběrového souboru pro každou simulaci šetření. Poznamenejme, že rychlost provedení simulací programem závisí na počtu respondentů. Pro vyšší počty (tisíce) je třeba brát zřetel na to, že program bude pracovat pomaleji. V polích B9 – F9 volíme podíly jedinců (v %) v pěti kategoriích populace (tedy K = 5), které jsou dány nějakou vážící proměnnou (např. nejvyšší dosažené vzdělání, věkové kategorie). V polích G9 – K9 potom zadáváme poměr kladných odpovědí (opět v %) respondentů v rámci každé kategorie ve výběrovém souboru, tj. $\mathrm{E}\hat{p}_i$, $i = 1,\ldots, 5$.

Program pak (po stisku klávesy F9) provede podle zadaných hodnot simulaci 1000 výběrových šetření (představované jednotlivými řádky). Generuje počty jedinců ve výběru patřících do jednotlivých kategorií (sloupce B – F) podle multinomického rozdělení, jehož parametry jsou celkový počet respondentů a zadané procentuální zastoupení kategorií v populaci. Dále generuje pro každou kategorii počet respondentů (sloupce G – K), kteří na položenou otázku odpověděli kladně, a to podle binomického rozdělení, jehož parametry jsou počet respondentů ve výběru v dané kategorii (sloupce B – F) a zadané procentuální zastoupení kladných odpovědí v populaci v této kategorii.

Ve sloupcích L – P pak program pro jednotlivé situace spočítá váhy podle klíče (5). V následujících dvou sloupcích je vždy spočten vážený i nevážený výběrový průměr, jež odhadují relativní četnost odpovědí „Ano“ pro celou populaci.

V horní části programu (pole G5) je uvedena hodnota R^2, vypočtená podle definice (viz vzorec (5)). Ve sloupci Q je pak podle klasického vzorce (1) vypočtena směrodatná odchylka (pole Q2), a dále výběrová směrodatná odchylka na základě nevážených průměrů (Q3) a výběrová směrodatná odchylka na základě vážených průměrů (Q4).

Následuje výpočet redukčního koeficientu *C* (pole Q5), který můžeme porovnávat se skutečnou hodnotou $1 - R^2$ v následujícím řádku (pole Q6).

Poslední dvě hodnoty jsou pro tuto kapitolu klíčové. Pokud provedeme několik simulací (ať se stejným zadáním, nebo různým), vidíme, že ve většině případů se od sebe tyto hodnoty liší pouze nepatrně, což podporuje pravdivost naší domněnky, že vztah $\frac{\widehat{\mathrm{var}}\ \hat{p}_w}{\widehat{\mathrm{var}}\ \hat{p}} = 1 - R^2$ je platný (alespoň s dostatečnou přesností) i v případě vážení.

Jak již bylo řečeno v kapitole 1.3, v praxi je třeba dát si pozor, abychom vážili data podle vhodné vážící proměnné. Jinak by tato metoda mohla mít na statistickou chybu i negativní vliv. To lze pozorovat i u prováděných simulací – pokud nastavíme zadání tak, že příslušnost ke kategorii nemá na pravděpodobnost kladné odpovědi vliv, tzn. ve sloupcích G – K nastavíme všechny hodnoty stejné, pak již $1 - R^2 = 1$ a redukční koeficient se od této hodnoty příliš neliší. Přitom může být i větší než jedna, což by znamenalo, že po vynásobení klasicky vypočteného odhadu touto hodnotou dojde ke zhoršení statistické chyby. Při volbě příliš malého podílu některé kategorie v populaci se zase může stát, že ve výběrovém souboru nebude žádný respondent z této kategorie (pak v programu nastane problém s dělením nulou).

Poznamenejme, že v praxi se často používá kombinace kvótního výběru a vážení dat. Nejprve se výběrový soubor volí na základě kvót a poté se ještě získaná data převažují (typicky podle stejných proměnných). Pokud by tedy vzorec $C = 1 - R^2$ opravdu platil jak pro vážená data, tak pro více proměnných v případě kvótních výběrů, přičemž intuice a simulace naznačují, že tomu tak opravdu je, nejspíše by bylo možné tohoto vztahu využít i při kombinaci obou postupů.

Kapitola 3

Odhad statistické chyby pro volební modely

Nyní se budeme zabývat jedním odvětvím výběrového šetření, a to volebními modely neboli modely, jež zachycují odhady volebních preferencí jednotlivých politických stran. Jedná se o speciální případ průzkumů veřejného mínění.

U veřejně dostupných reálných modelů z minulých let, resp. měsíců se v této kapitole pokusíme z dat empiricky odhadnout statistickou chybu, a to hned třemi různými způsoby. Číselné hodnoty těchto odhadů pak můžeme přímo porovnávat právě s hodnotami spočtenými klasickým vzorcem z (1). Nejprve se ale podrobněji seznámíme s průzkumy veřejného mínění a volebními modely.

3.1 Průzkumy volebních preferencí

Volební modely mají široké využití. Dávají voličům i samotným politickým stranám informaci, kolik procent hlasů by strany mohly ve volbách získat. Některá politická strana se tak třeba dozví, zdali je natolik podporována, že by se v příštích volbách mohla stát stranou parlamentní, jiná strana zjistí, nakolik ji poškodilo nemorální chování člena této strany. Na základě průzkumů volebních preferencí mohou politické strany upravovat svou volební kampaň. Některé strany si dokonce nechávají vypracovat volební model pouze pro vlastní účely. Kromě toho využívají volební modely např. sázkové kanceláře, které podle volebního modelu a jeho statistické chyby určují sázkové kurzy. Lze si pak vsadit třeba na vítězství strany ve volbách či na to, zda některá strana dostane dostatečný počet hlasů, aby mohla usednout v Parlamentu.

Průzkumy volebních preferencí stály na samotném počátku historie výzkumů veřejného mínění. První výzkumy veřejného mínění se totiž začaly provádět v USA počátkem 19. století za účelem zjištění volebních preferencí, a to zejména před prezidentskými volbami. V této době ještě nebyly dodržovány žádné metodologické principy a prováděly se pouze jednoduché ankety zvané „straw polls“, které se těšily velké oblibě až do 30. let 20. století.

Nicméně už roku 1916 byl výzkum poprvé proveden na základě kvótního výběru a kolem roku 1935 se začaly používat moderní vědecké metodologické postupy – byly standardizovány metody dotazování a zpracování výsledků. Výzkumy veřejného mínění se v tomto období přestávají týkat pouze volebních prognóz a věnují se i průzkumu trhu a žebříčkům popularity významných osobností kulturního života.

Od té doby neustále roste počet prováděných výzkumů veřejného mínění, které jsou často sponzorovány médii, a také průzkumů trhu, jež zase často financují výrobci různých produktů a poskytovatelé služeb. Také se neustále rozvíjí a upravují vědecké postupy, jež jsou potom převáděny do praxe, viz [4].

Každá agentura zabývající se touto problematikou se přirozeně snaží uvádět výsledky, jež by co nejvíce odpovídaly reálnému veřejnému mínění. Data ale samozřejmě i zde budou vždy zatížena statistickou chybou, která je tudíž součástí výsledků každého takového průzkumu a jako taková by měla být vždy prezentována společně s ostatními daty.

V této kapitole se tedy budeme věnovat právě odhadu statistické chyby u volebních modelů vypracovaných v České republice dvěma zvolenými agenturami STEM, s. r. o. a MEDIAN, s. r. o. Tyto agentury provádějí průzkum volebních preferencí pravidelně každý měsíc již několik let, a výsledky prezentují na webových stránkách v podobě tiskových zpráv, viz [8] a [10]. Lze tedy dlouhodobě sledovat vývoj volebních preferencí, čehož budeme dále využívat. Tyto dvě agentury samozřejmě nejsou v České republice jediné, které se kontinuálnímu průzkumu volebních preferencí věnují. Jako další jmenujme např. Centrum pro výzkum veřejného mínění (CVVM). Pro naše potřeby ale postačí pouze dvě agentury.

Seznamme se nejprve s pojmy, jež je třeba rozlišovat, a to stranické a volební preference. *Stranické preference* reprezentují výsledky výzkumu vztažené ke všem respondentům. Dávají nám informace o tom, jak oprávnění voliči deklarují své rozhodnutí ve volbách, a to včetně těch, kteří neví, jakou stranu by volili, či jsou rozhodnuti k volbám nejít, viz [7].

Oproti tomu *volební model* neboli *model volebních preferencí* reprezentuje předpokládané výsledky voleb např. do Poslanecké sněmovny, přičemž do výsledných dat nejsou zahrnuty odpovědi těch respondentů, kteří vyloučili svou účast u voleb do Poslanecké sněmovny, nebo kteří nejsou rozhodnuti, kterou stranu by volili. Nadále budeme pracovat právě s modelem volebních preferencí.

Stranické i volební preference jsou dlouhodobě dány socio-demografickými znaky veřejnosti a celkovým politickým smýšlením v rámci státu. Krátkodobě bývají velmi ovlivňovány každodenním děním na politické scéně, zprávami v médiích i volebními kampaněmi, proto se v meziměsíčním srovnání mohou u některé strany vyskytovat i poměrně velké rozdíly v preferencích. Tyto rozdíly nelze vysvětlovat statistickou chybou, a tudíž se jejich vliv budeme snažit z našich odhadů statistické chyby pokud možno vyloučit.

Je třeba zdůraznit, že volební model, resp. model stranických preferencí vyjadřuje pouze aktuální podporu jednotlivých politických stran, která se v čase často velmi mění. Proto jsou pro předpověď výsledků voleb podstatné především takové průzkumy, jež jsou prováděny v kratší časové vzdálenosti od těchto voleb. Přesnost předpovědi navíc může být ovlivněna i volebním systémem dané země a jinými faktory, viz [6].

Poznamenejme ještě, že zde sice používáme následující postupy pro odhad statistické chyby u průzkumů volebních preferencí, nicméně lze tyto postupy aplikovat na kterékoliv jiné výběrové šetření, jež se provádí kontinuálně, tedy např. při opakovaných průzkumech trhu, které si zadávají některé společnosti. Výsledky těchto šetření ale zpravidla nejsou veřejně dostupné a nelze je mezi sebou porovnávat. Z těchto důvodů se v této práci věnujeme právě volebním modelům.

3.2 Vývoj volebních preferencí v ČR

Jak již bylo uvedeno výše, budeme se zde zabývat konkrétními daty získanými z tiskových zpráv o volebních modelech agentur MEDIAN, s. r. o. a STEM, s. r. o., které se mimo jiné věnují průzkumu stranických i volebních preferencí. Uveďme zde proto základní data prezentovaná v těchto zprávách.

Agentura MEDIAN prezentuje výsledky průzkumů stranických i volebních preferencí kontinuálně každý měsíc od roku 2006, viz [8]. Terénní sběr dat probíhá vždy přibližně během celého kalendářního měsíce, ke kterému jsou pak výsledky vztaženy, a to stratifikovaným adresním náhodným výběrem, což znamená, že respondenti jsou vybíráni podle adresy jejich bydliště a zároveň např. podle kraje ve kterém žijí. Respondenti pak po vážení dat tvoří reprezentativní vzorek populace ČR podle údajů ČSÚ. Sběr dat je uskutečňován osobním rozhovorem mezi tazatelem a respondentem, výsledky jsou zaznamenávány do elektronického dotazníku.

Dále je dle tvrzení samotné agentury ve výsledcích zohledněna i informace, jak respondenti hlasovali v posledních volbách do Poslanecké sněmovny, přičemž není konkrétně uvedeno, jak je tato informace použita při tvorbě volebního modelu. Bylo by to možné provést například tak, že kromě zjišťování, koho by respondent volil v současné době, je položena i otázka, koho respondent volil při minulých volbách. Poté by byl volební model pomocí vážení získaných dat upraven podle známých výsledků posledních voleb tak, aby zohledňoval minulou volbu. K takovým úpravám můžeme přistoupit, např. pokud v našem výběru nezvykle vzrostou preference KSČM, přitom ale zjistíme, že poměr respondentů, kteří volili tuto stranu při minulých volbách, je ve výběrovém souboru podstatně vyšší než u skutečných výsledků voleb. Vážením získaných dat, tak můžeme upravit současný poměr a tím i zpřesnit odhad volebních preferencí.

Další data, jež v tiskových zprávách MEDIAN prezentuje, uvádíme v Tabulce P1, která je, stejně jako následující Tabulka P2, z důvodu velikosti umístěna v závěru této práce v Přílohách. Ve sloupcích jsou v procentech uvedeny volební preference jednotlivých stran za dané období.

SE značí odhad statistické chyby, jak ji uvádí pro daný měsíc ve zprávě sama agentura. Všimněme si, že za měsíce červenec až říjen roku 2006 nebyla statistická chyba vůbec uvedena. Pokud je v dalších obdobích uvedeno ve sloupci SE rozmezí čísel, pak se nižší hodnoty statistické chyby vždy týkají „menších" politických stran, tj. stran s nižšími volebními preferencemi, a vyšší hodnoty naopak „větších" politických stran, tj. stran s vyššími volebními preferencemi.

V posledním sloupci jsou vypsány počty respondentů za jednotlivá období tvořících výběrový soubor, na jehož základě je volební model sestaven. Jak MEDIAN, tak STEM v tiskových zprávách uvádí počty respondentů pro stranické preference, bylo tedy nutné tyto počty na základě prezentovaných stranických preferencí, konkrétně poměru nerozhodnutých voličů a respondentů, kteří nechtějí volit, přepočítat tak, aby odpovídaly skutečným počtům respondentů, na jejichž odpovědích je volební model založen.

Volební preference strany TOP 09 nejsou v Tabulce 3 až do června 2009 uvedeny, protože teprve v této době strana vznikla. Založením této strany lze pravděpodobně vysvětlit výkyv volebních preferencí KDU-ČSL v srpnu 2009 směrem dolů – pokles byl zapříčiněn přesunem sympatií od KDU-ČSL k nově založené straně.

Politická strana Věci veřejné (VV) sice vznikla již v roce 2001, ale až do září 2009 tato strana měla nízké preference, a tak byla řazena mezi „Ostatní", tzn. ostatní, samostatně neuvedené, politické strany s velmi nízkou přízní voličů.

Agentura STEM prezentuje (veřejně dostupné) výsledky průzkumů volebních preferencí až na výjimky každý měsíc od roku 2005 metodou kvótního výběru, přičemž sběr dat probíhá vždy v prvním týdnu v daném měsíci mezi obyvateli ČR staršími 18-ti let, viz [10]. My však použijeme pouze data od roku 2006, neboť mimo jiné budeme chtít porovnávat výsledky obou agentur ve stejném měsíci, a dřívější data bychom tedy nevyužili.

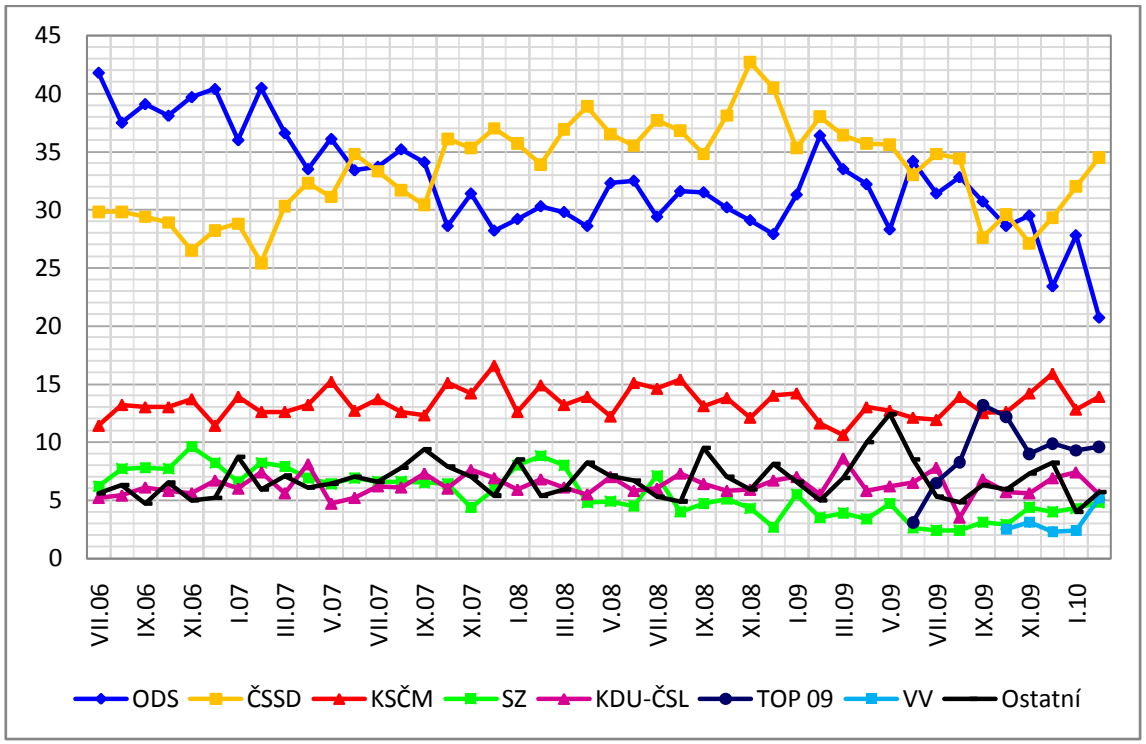
STEM sice ve veřejně dostupných materiálech neuvádí volební preference, ale ze stranických preferencí lze volební model vypočítat klasickou trojčlenkou. Agentura STEM své výsledky stranických preferencí prezentuje společně se statistickou chybou od září roku 2005, a to v rozmezí 1,5–2,5 .

Výše zmíněné výjimky se týkají letních měsíců, kdy tato agentura průzkum neprovádí. Proto nejsou v Tabulce P2, která shrnuje základní data průzkumů agentury STEM, žádné výsledky u těchto měsíců uvedeny. Volební preference jsou i zde uvedeny v procentech.

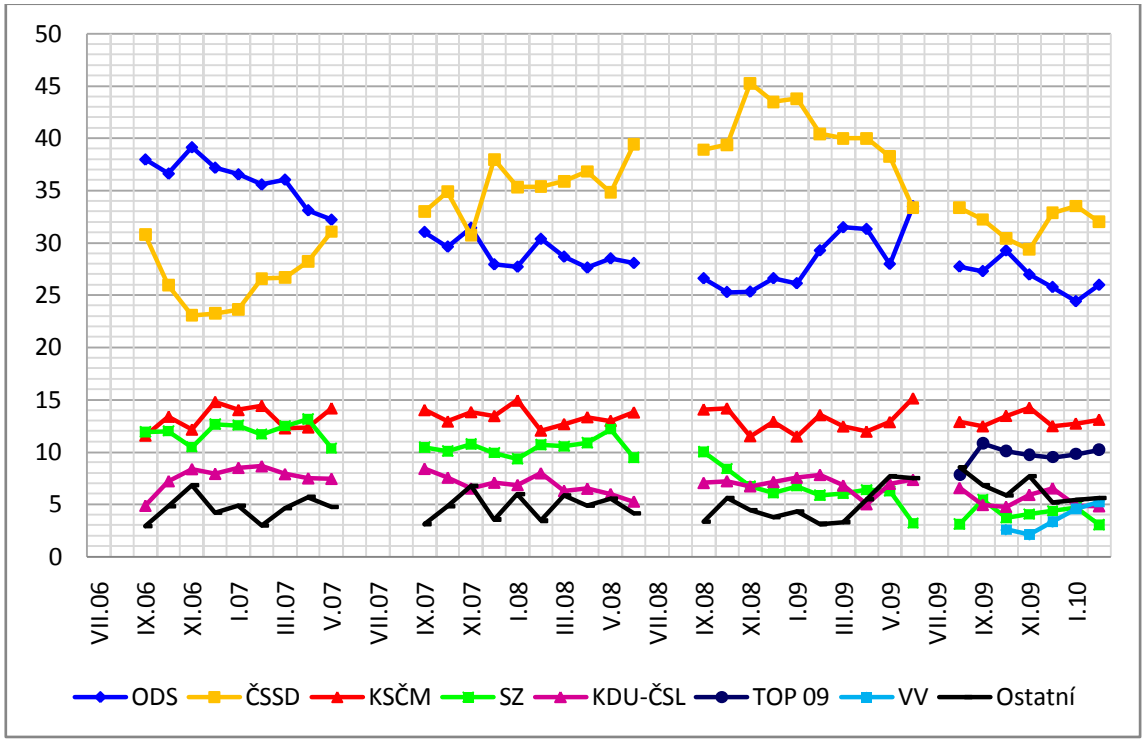
Volební preference politických stran TOP 09 a VV jsou zaznamenány až od srpna 2009, resp. října 2009 z obdobných důvodů, jako tomu bylo u agentury MEDIAN.

Podívejme se na data ještě v grafické podobě (Graf 1 pro případ MEDIANu a Graf 2 pro případ STEMu), kde bude lépe patrný vývoj volebních preferencí pro jednotlivé strany.

Graf 1: Vývoj volebních preferencí (v %) – MEDIAN



Graf 2: Vývoj volebních preferencí (v %) – STEM



Nyní jsme tedy seznámeni s daty, se kterými budeme dále pracovat. Na základě těchto dat budeme v následujícím textu třemi různými způsoby odhadovat statistickou chybu, které se dopouštíme, když výsledky průzkumů volebních preferencí zobecňujeme na základní soubor, což jsou v tomto případě občané ČR starší 18-ti let, kteří chtějí volit a vědí, kterou politickou stranu by volili. Výsledky budeme porovnávat s klasickým vzorcem pro rozptyl, z něhož, jak víme, vyjadřujeme odhad statistické chyby. Přitom rozmezí statistické chyby tak, jak jej uvádějí samotné agentury, až na výjimky pokrývá odhady vypočtené tímto vzorcem (viz Tabulka P3 v Přílohách).

3.3 Vyrovnávání časové řady

Protože se volební preference sledují kontinuálně a volební modely pro každý měsíc jsou veřejně dostupné, lze sledovat vývoj preferencí v čase, čehož zde můžeme velmi dobře využít. Vývoj volebních preferencí, resp. jejich odhady, jak je udávají agentury, pro jednotlivé časy můžeme vyrovnat pomocí nějaké křivky, jinak řečeno můžeme časovou řadu proložit vhodným polynomem. O této křivce budeme předpokládat, že kopíruje skutečný vývoj volebních preferencí. Z rozdílů (reziduí) mezi hodnotami udávanými agenturami a touto křivkou sestrojíme empirický odhad statistické chyby.

Dále budeme postupovat pro každou agenturu a stranu zvlášť, přičemž se budeme věnovat pouze stálým parlamentním stranám, tj. ODS, ČSSD, KSČM a KDU-ČSL.

Označme $t = 1, 2, \ldots, T$ časový okamžik (měsíc), ve kterém byl uskutečněn průzkum, kde T je celkový počet těchto měsíců. Nechť dále p_t je agenturou zjištěná hodnota volebních preferencí pro jednu stranu v čase t, a $\hat{p}_t$ je odhad p_t pomocí vhodného polynomu. Definujme rezidua $r_t := p_t - \hat{p}_t$.

Vývoj volebních preferencí jedné strany proložíme nějakým vhodným polynomem, přičemž vhodnost budeme kontrolovat výběrovým autokorelačním koeficientem 1. řádu. Budeme požadovat, aby byl výraz

$$r := \frac{\sum_{t=1}^{T-1}(r_t - \bar{r}_t)(r_{t+1} - \bar{r}_t)}{\sum_{t=1}^{T}(r_t - \bar{r}_t)^2},$$

kde $\bar{r}_t$ značí průměr reziduí r_t, dostatečně blízký nule. Jinými slovy budeme chtít, aby hypotéza nulovosti této autokorelace nebyla zamítnuta. Tak budeme kontrolovat, zda jsme vývoj volebních preferencí nevyrovnali příliš, což by naznačovala hodnota $r \gg 0$. Zároveň budeme ověřovat normalitu reziduí Shapirovým-Wilkovým testem, neboť při výpočtu statistické chyby používáme kvantil právě normálního rozdělení.

Postupujme nyní podobně jako v kapitole 2. Spočítejme odhad rozptylu $\hat{p}_t$ klasickým vzorcem:

$$\text{var}\, r_t = \frac{\hat{p}_t(1-\hat{p}_t)}{n_t} \tag{7}$$

a položme si otázku, zda by tento odhad nebylo možné snížit pomocí nějakého redukčního koeficientu menšího než jedna, označme jej opět C. Odhad rozptylu tedy vyjádřeme v tomto tvaru:

$$\text{var}\, r_t = C\left[\frac{\hat{p}_t(1-\hat{p}_t)}{n_t}\right],$$

kde C je reálné nezáporné číslo, a n_t je počet respondentů v čase *t*. Naším cílem je nyní odhadnout právě C = konst., která udává, kolikrát lze snížit odhad rozptylu vypočtený pomocí klasického vzorce (7).

Abychom omezili vliv měnícího se počtu respondentů a hodnot $\hat{p}_t$ v čase, definujeme *normovaná rezidua* takto:

$$\tilde{r}_t := \frac{r_t}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{n_t}}}.$$

Pak dostáváme:

$$\text{var}\, \tilde{r}_t = \frac{n_t}{\hat{p}_t(1-\hat{p}_t)}\text{var}\, r_t = C.$$

Přitom C, resp. $\text{var}\, \tilde{r}_t$ budeme odhadovat výběrovým rozptylem normovaných reziduí (kde $\bar{r}_t = 0$ a $\bar{\tilde{r}}_t \doteq 0$):

$$\hat{C} = \widehat{\text{var}\, \tilde{r}_t} = \frac{\sum_{t=1}^{T} \tilde{r}_t^2}{T}.$$

Přitom očekáváme $\hat{C} < 1$ (viz vztah $C = 1 - \text{R}^2$ ve druhé kapitole). Pak totiž dosazením do vyjádření $\text{var}\, r_t$ získáme nižší odhad rozptylu. Pokud dostaneme $\hat{C} = 1$, potvrdí se tím vhodnost klasického vzorce (7), a jestliže $\hat{C} > 1$, pak tímto postupem překvapivě dostáváme vyšší odhad statistické chyby a je zde také vhodnější použít vzorec (7). Zkusme tedy tento postup aplikovat na data od zvolených agentur.

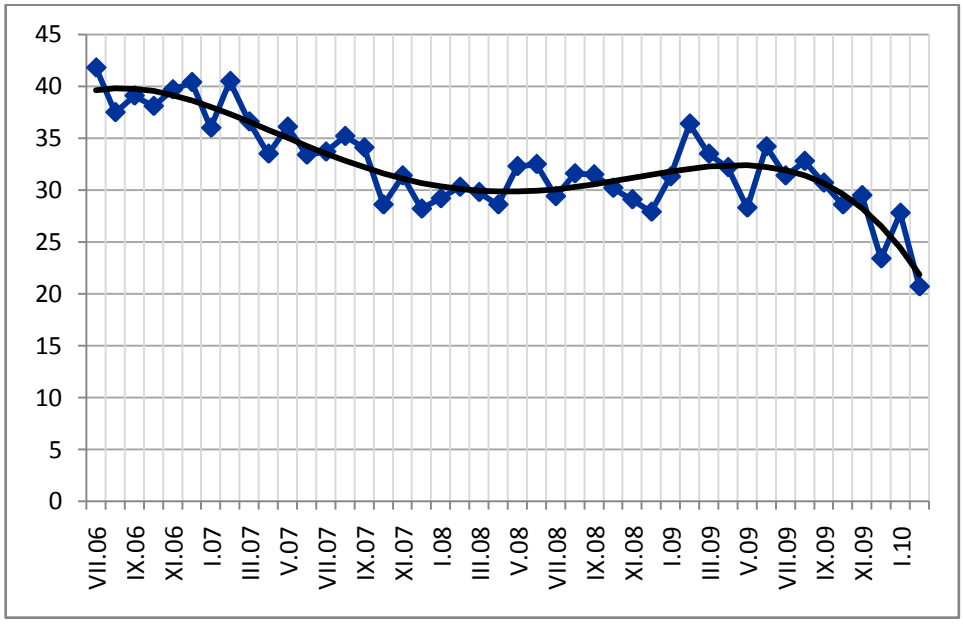
3.3.1. Výsledky pro agenturu MEDIAN

Pro všechny politické strany v tomto odstavci platí $T = 44$. Podívejme se na výsledky, které dostaneme výše uvedeným postupem aplikovaným po řadě na odhady volebních preferencí stran ODS, ČSSD, KSČM a KDU-ČSL.

Začněme u vývoje volebních preferencí strany ODS. Časovou řadu proložíme polynomem čtvrtého stupně: $\hat{p}_t = 39{,}132 + 0{,}593t - 0{,}150t^2 + 0{,}006t^3 - 8 \cdot 10^{-5}t^4$, $t = 1, 2, \ldots, 44$ (viz Graf 3).

Vzhledem k tomu, že v následujících případech bude téměř vždy použit týž postup, budeme důležitá data pro přehlednost vypisovat do tabulky. V prvním řádku bude uveden stupeň polynomu, jímž časovou řadu prokládáme, druhý řádek bude obsahovat p-hodnotu testu nulovosti koeficientu u nejvyšší mocniny tohoto polynomu (pokud budeme časovou řadou prokládat polynomu alespoň prvního stupně). Následuje výběrový autokorelační koeficient reziduí a p-hodnota, již dostaneme otestováním nulovosti tohoto autokorelačního koeficientu. V pátém řádku uvádíme p-hodnotu získanou po provedení Shapirova-Wilkova testu normality reziduí a v posledním řádku je uvedena zjištěná hodnota $\hat{C}$. Veškeré hodnoty budeme zaokrouhlovat na tři desetinná místa. Zde bude mít tato tabulka konkrétně podobu Tabulky 3.

Graf 3: Vývoj volebních preferencí ODS a jejich vyrovnání (v %) – MEDIAN



Ačkoli výběrová autokorelace reziduí není nulová, je nule dostatečně blízká, jak naznačuje p-hodnota Spearmanova testu nulovosti autokorelace. Nulovost koeficientu u nejvyšší mocniny prokládaného polynomu zde zamítáme, neboť p-hodnota testu nulovosti je rovna $3 \cdot 10^{-5}$. Podle Shapirova-Wilkova testu nezamítáme hypotézu normality reziduí (p-hodnota = 0,975). Protože tomu tak bude i téměř ve všech ostatních

případech, nebudeme tyto skutečnosti dále zdůrazňovat. Naopak upozorníme pouze na případ, kde tomu bude jinak.

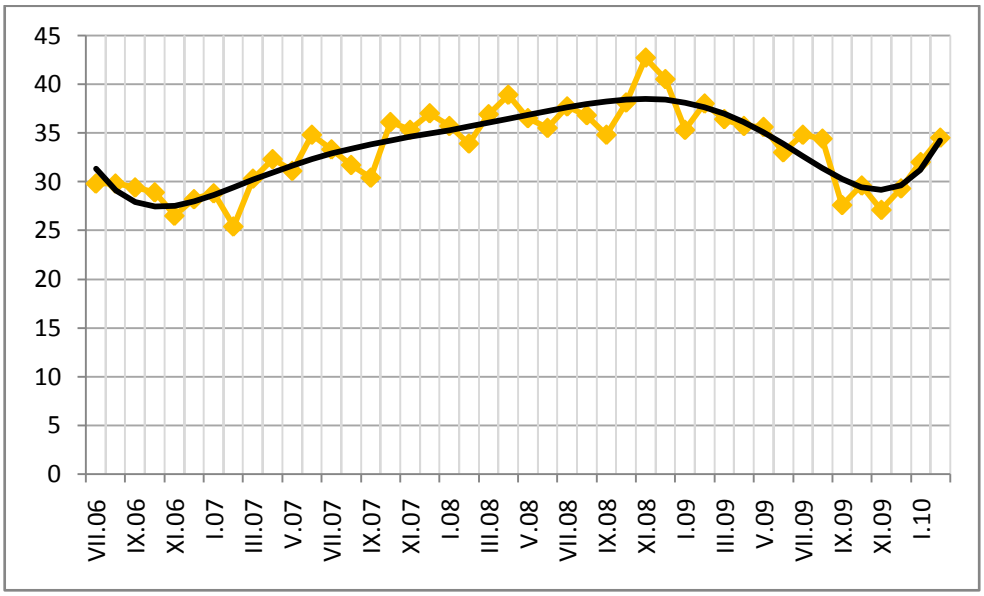
Tabulka 3: Numerické výsledky pro ODS získané vyrovnáváním – MEDIAN

Stupeň polynomu	4
Test nulovosti koeficientu – p-hodnota	$3 \cdot 10^{-5}$
Výběrový autokorelační koeficient	–0,181
Spearmanův test – p-hodnota	0,075
Shapirův-Wilkův test – p-hodnota	0,975
Odhad redukčního koeficient $\hat{C}$	1,006

Zde jsme získali hodnotu redukčního koeficientu (jen velice mírně) přesahující 1. Musíme tedy konstatovat, že lepší odhad statistické chyby zde tímto postupem nezískáváme. Lze říci, že jsme verifikovali použitelnost klasického vzorce (7).

Podívejme se na výsledky další strany, ČSSD. Vývoj volebních proložíme polynomem stupně 6, jenž bude mít tvar $\hat{p}_t = 34{,}959 - 4{,}439t + 0{,}882t^2 - 0{,}072t^3 + 0{,}003t^4 - 5 \cdot 10^{-5}t^5 + 4{,}6 \cdot 10^{-7}t^6$, $t = 1, 2, \ldots, 44$ (viz Graf 4).

Graf 4: Vývoj volebních preferencí ČSSD a jejich vyrovnání (v %) – MEDIAN



Zaměřme se na Tabulku 4. Hodnota $\hat{C}$ je nyní menší než 1. Došlo tedy ke zlepšení odhadu rozptylu, a tedy i statistické chyby.

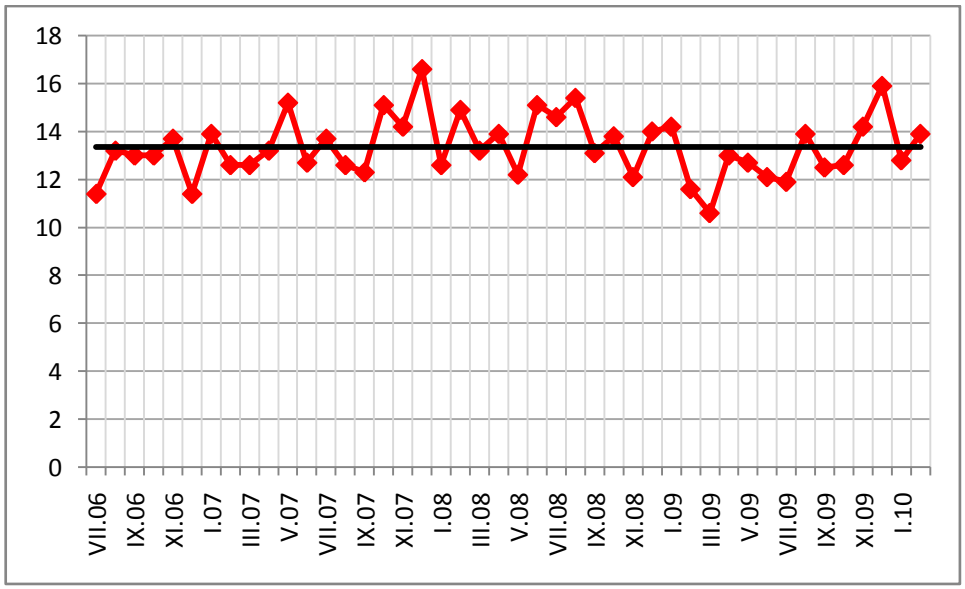
Tabulka 4: Numerické výsledky pro ČSSD získané vyrovnáváním – MEDIAN

Stupeň polynomu	6
Test nulovosti koeficientu – p-hodnota	0,002
Výběrový autokorelační koeficient	0,012
Spearmanův test – p-hodnota	0,743
Shapirův-Wilkův test – p-hodnota	0,878
Odhad redukčního koeficient $\hat{C}$	0,864

V případě KSČM dostáváme ještě lepší výsledek, jak je patrné z Tabulky 5. Redukční koeficient je roven 0,781, tedy nový odhad rozptylu je dokonce přibližně o pětinu menší než původní odhad. Je tomu tak zřejmě proto, že konstanta $\hat{p}_t$ = 13,3455, t = 1, 2,…., 44, (viz Graf 5), kterou jsme zde proložili časovou řadu, velmi dobře aproximuje vývoj volebních preferencí strany KSČM, tak jak jej udává agentura MEDIAN. Volební preference této strany jsou navíc silně korelovány se socio-demografií respondentů (vzdělání, věk, region) a zároveň má strana poměrně stabilní základnu voličů, takže se volební preference v čase příliš nemění a nevyskytují se zde žádné velké výkyvy, ať už směrem dolů, nebo směrem nahoru.

Při vyrovnávání časové řady se totiž může stát, že krátkodobé extrémní výkyvy nebere prokládaný polynom v potaz, takže nám v těchto časových okamžicích vzroste absolutní velikost rezidua, což se pak projeví i na vyšším redukčním koeficientu. To se však u této strany neděje a z toho důvodu je stabilní vývoj preferencí KSČM optimální pro odhad statistické chyby pomocí vyrovnávání časové řady.

Graf 5: Vývoj volebních preferencí KSČM a jejich vyrovnání (v %) – MEDIAN



Tabulka 5: Numerické výsledky pro KSČM získané vyrovnáváním – MEDIAN

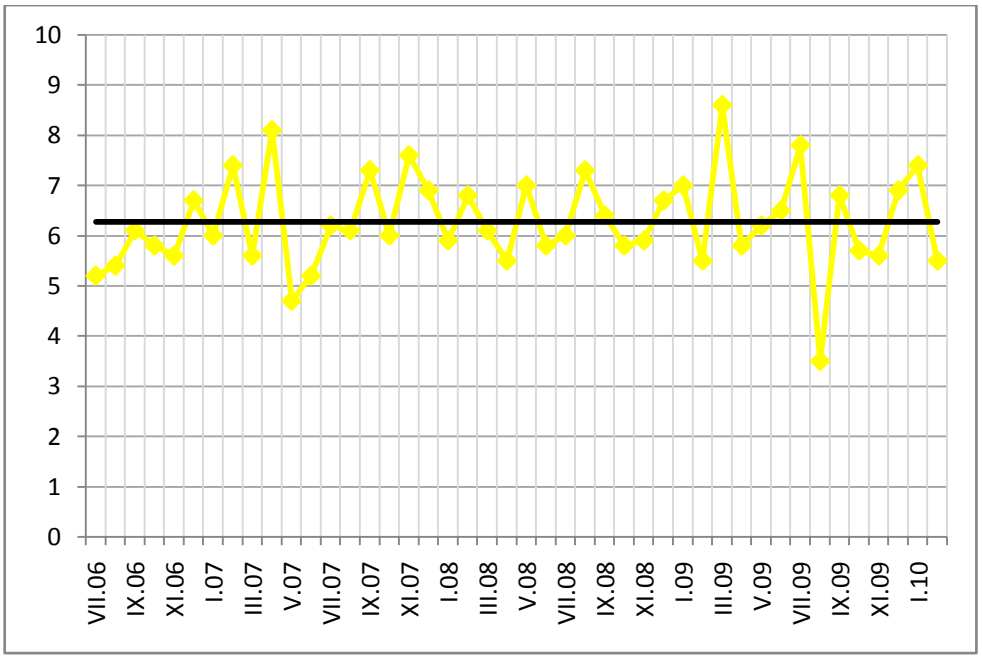
Stupeň polynomu	0
Výběrový autokorelační koeficient	0,047
Spearmanův test – p-hodnota	0,909
Shapirův-Wilkův test – p-hodnota	0,684
Odhad redukčního koeficient $\hat{C}$	0,781

Při tomto postupu získáváme i u strany KDU-ČSL obdobný výsledek jako u předchozích dvou stran. Hodnoty p_t prokládáme tímto polynomem: $\hat{p}_t = 5{,}202 + 0{,}218t - 0{,}013t^2 + 3{,}6\cdot10^{-4}t^3 - 3{,}4\cdot10^{-6}t^4$, $t = 1, 2, \ldots, 44$ (viz Graf 6). V Tabulce 6 vidíme, že hodnota redukčního koeficientu je 0,870, tedy opět dostáváme nižší odhad statistické chyby.

Tabulka 6: Numerické výsledky pro KDU-ČSL získané vyrovnáváním – MEDIAN

Stupeň polynomu	0
Výběrový autokorelační koeficient	–0,336
Spearmanův test – p-hodnota	0,071
Test založený na bodech zvratu – p-hodnota	0,715
Mediánový test – p-hodnota	0,118
Shapirův-Wilkův test – p-hodnota	0,299
Odhad redukčního koeficient $\hat{C}$	0,870

Graf 6: Vývoj volebních preferencí KDU-ČSL a jejich vyrovnání (v %) – MEDIAN



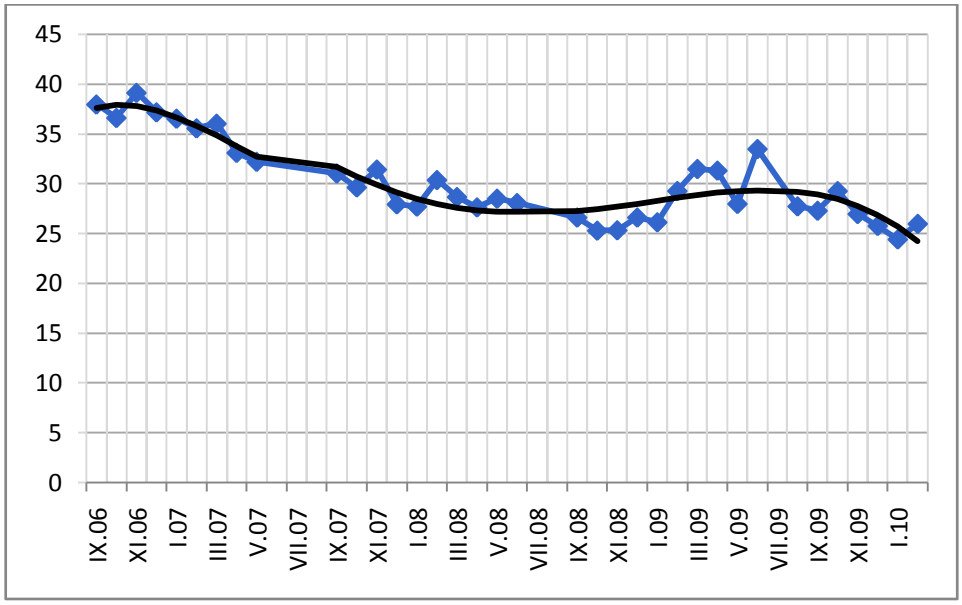
Výběrový autokorelační koeficient je zde záporný a dosti vzdálený od nuly, což lze vysvětlit občasnými výraznými střídavými výkyvy volebních preferencí (např. mezi červencem a srpnem 2009), na něž je Spearmanův korelační koeficient citlivý. Nicméně důležité je, že rezidua nejsou silně kladně korelovaná, což by znamenalo, že jsme data vyrovnali příliš. Ačkoli bychom ani zde nulovost autokorelačního koeficientu nemuseli zamítnout, raději ještě ověříme náhodnost a nezávislost reziduí testem založeném na počtu bodů zvratu a mediánovým testem – p-hodnoty jsou uvedeny v Tabulce 6.

3.3.2. Výsledky pro agenturu STEM

Na data agentury STEM byl aplikován týž postup, opět pro strany ODS, ČSSD, KSČM a KDU-ČSL. Tentokrát máme $T = 36$, neboť v letních měsících agentura průzkum neprovádí. Tyto měsíce jsou tedy při vyrovnávání časové řady vynechávány.

Vývoj volebních preferencí strany ODS jsme proložili polynomem čtvrtého stupně: $\hat{p}_t = 37{,}938 + 0{,}331t - 0{,}173t^2 + 0{,}009t^3 - 1{,}4\cdot10^{-4}t^4$, $t = 1, 2, \ldots, 36$ (viz Graf 7). Shapirovým-Wilkovým testem normality dostáváme poměrně nízkou p-hodnotu. Nicméně pokud bychom zvýšili stupeň prokládaného polynomu, pak podle testu pro koeficient u nejvyšší (tj. páté) mocniny tohoto polynomu nezamítáme jeho nulovost. Vzhledem k tomu, že u agentury MEDIAN jsme časovou řadu volebních preferencí ODS prokládali polynomem čtvrtého stupně, přikláníme se i zde k této variantě. V záznamech v Tabulce 7 vidíme, že hodnota $\hat{C}$ poměrně dosti přesahuje 1, tudíž zde tímto postupem nižší odhad statistické chyby nezískáváme a je tedy vhodnější použít např. vzorec (7).

Graf 7: Vývoj volebních preferencí ODS a jejich vyrovnání (v %) – STEM



Tabulka 7: Numerické výsledky pro ODS získané vyrovnáváním – STEM

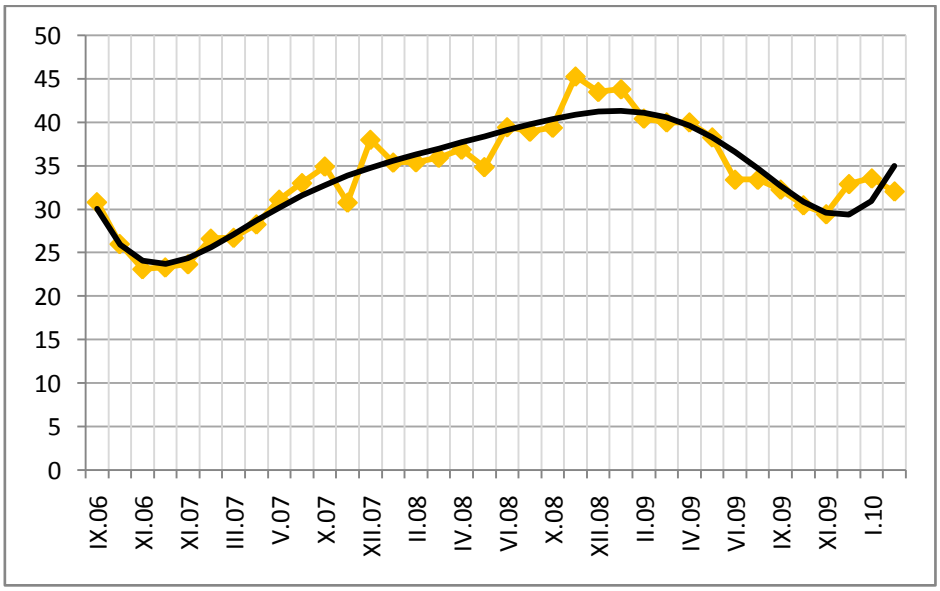
Stupeň polynomu	4
Test nulovosti koeficientu – p-hodnota	$4 \cdot 10^{-4}$
Výběrový autokorelační koeficient	0,026
Spearmanův test – p-hodnota	0,640
Shapirův-Wilkův test – p-hodnota	0,048
Odhad redukčního koeficient $\hat{C}$	1,278

Ani pro volební preference ČSSD nezískáváme nižší odhad, koeficient $\hat{C}$ je dokonce vyšší než u ODS. Polynom, který prokládáme agenturou zjištěnými hodnotami $\hat{p}_t = 37{,}340 - 9{,}156t + 2{,}093t^2 - 0{,}198t^3 + 0{,}010t^4 - 2 \cdot 10^{-4}t^5 + 2{,}1 \cdot 10^{-6}t^6$, $t = 1, 2, \ldots, 36$, sice aproximuje data poměrně přesně (viz Graf 8), přesto dostáváme neuspokojivý výsledek $\hat{C} = 1{,}615$. Výsledky postupu pro stranu ČSSD shrnuje Tabulka 8.

Tabulka 8: Numerické výsledky pro ČSSD získané vyrovnáváním – STEM

Stupeň polynomu	6
Test nulovosti koeficientu – p-hodnota	$5 \cdot 10^{-4}$
Výběrový autokorelační koeficient	0,026
Spearmanův test – p-hodnota	0,467
Shapirův-Wilkův test – p-hodnota	0,088
Odhad redukčního koeficient $\hat{C}$	1,615

Graf 8: Vývoj volebních preferencí ČSSD a jejich vyrovnání (v %) – STEM

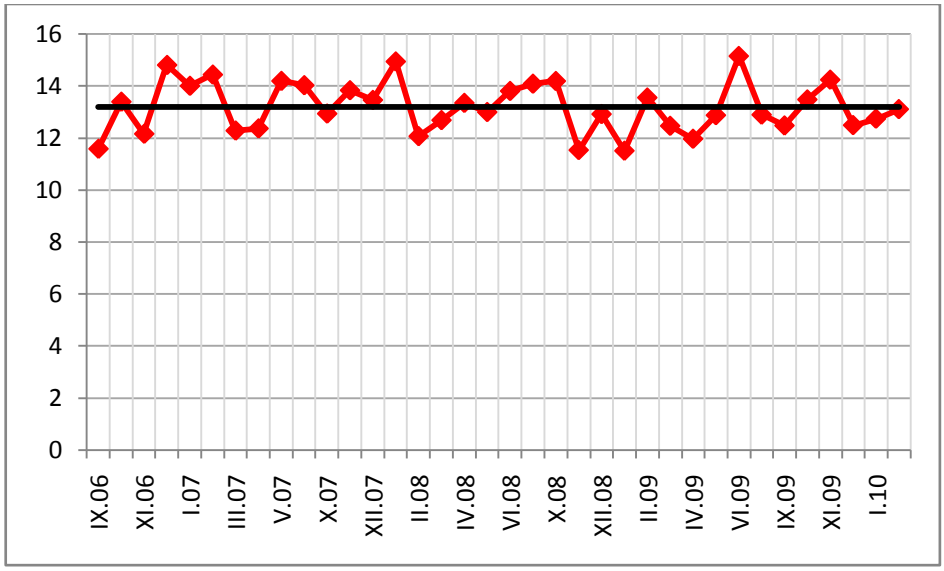


Tabulka 9: Numerické výsledky pro KSČM získané vyrovnáváním – STEM

Stupeň polynomu	0
Výběrový autokorelační koeficient	–0,056
Spearmanův test – p-hodnota	0,793
Shapirův-Wilkův test – p-hodnota	0,622
Odhad redukčního koeficient $\hat{C}$	0,925

U politických stran KSČM a KDU-ČSL už docházíme k lepším závěrům. Data týkající se KSČM jsme opět aproximovali konstantním polynomem (viz Graf 9): $\hat{p}_t$ = 13,1917, t = 1, 2,…., 36. Výsledek pro stranu KSČM je zahrnut v Tabulce 9.

Graf 9: Vývoj volebních preferencí KSČM a jejich vyrovnání (v %) – STEM

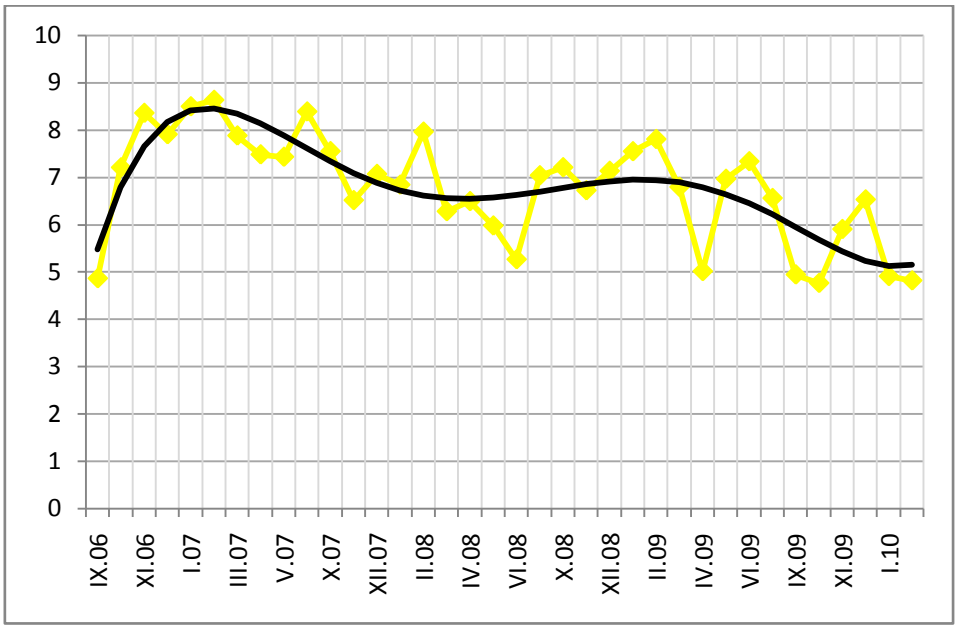


Vývoj volebních preferencí KDU-ČSL jsme proložili polynomem pátého stupně: $\hat{p}_t = 3{,}617 + 2{,}165t - 0{,}329t^2 + 0{,}020t^3 - 5{,}4\cdot10^{-4}t^4 + 5{,}2\cdot10^{-6}t^5$, t = 1, 2,…., 36 (viz Graf 10). Z hlediska agentury STEM jsme u této strany dosáhli největšího snížení rozptylu, jak je uvedeno v Tabulce 10:

Tabulka 10: Numerické výsledky pro KDU-ČSL získané vyrovnáváním – STEM

Stupeň polynomu	5
Test nulovosti koeficientu – p-hodnota	0,006
Výběrový autokorelační koeficient	0,091
Spearmanův test – p-hodnota	0,934
Shapirův-Wilkův test – p-hodnota	0,907
Odhad redukčního koeficient $\hat{C}$	0,877

Graf 10: Vývoj volebních preferencí KDU-ČSL a jejich vyrovnání (v %) – STEM



Po zhlédnutí všech spočítaných hodnot redukčních koeficientů (viz Tabulky 3 až 10 nebo též viz shrnující tabulka 13 v Kapitole 3.6) můžeme konstatovat, že pro data agentury STEM dostáváme ve všech případech horší výsledky než pro data agentury MEDIAN. Můžeme se jen domnívat, že je to způsobeno různou kvalitou vypracování průzkumu, např. technikou dotazování respondentů, způsobem stanovení kvót, sadou vážících proměnných, zohledňováním výsledků posledních voleb apod.

Poznamenejme ještě, že by jistě bylo možné vývoj volebních preferencí jednotlivých stran aproximovat i jiným způsobem než prokládanými polynomy, a to například klouzavými průměry, nicméně vidíme, že polynomy zde plní svůj účel poměrně dobře. V kapitolách 3.4 a 3.5 však namísto hledání optimálního proložení časových řad vyzkoušíme odlišné způsoby odhadu redukčního koeficientu C.

3.4 Časové diference

V předchozím postupu lze poměrně dobře kontrolovat, zdali vývoj preferencí nevyrovnáváme příliš. Můžeme mít ale podezření, že preference vyrovnáváme málo, tedy že prokládaný polynom kopíruje příliš těsně vývoj preferencí. Tím bychom ovšem nevhodným způsobem zmenšili redukční koeficient a náš odhad statistické chyby by byl nesprávný. Proto zde popíšeme poněkud „konzervativnější“ postup založený na časových diferencích, v němž se takového nedostatku nemůžeme dopustit. Jeho nevýhodou ovšem je, že jen zřídka dostaneme lepší výsledek (nižší hodnotu $\hat{C}$) než při předešlém nebo následujícím postupu (viz kapitola 3.5).

I nadále se budeme věnovat pouze čtyřem vybraným stranám, a to opět každé straně a agentuře zvlášť. Označme $d_t = p_t - p_{t-1}$, kde p_t je výše volebních preferencí vybrané politické strany v čase *t*, a p_{t-1} je výše volebních preferencí téže strany, ale v čase $t - 1$, $t = 2, 3, \ldots, T$, kde $T = 44$ v případě agentury MEDIAN a $T = 36$ v případě agentury STEM. Nechť i zde je $r_t = p_t - \hat{p}_t$, kde $\hat{p}_t$ značí odhad hodnoty p_t proloženým polynomem, přičemž u jednotlivých stran využijeme hodnot $\hat{p}_t$ již získaných v průběhu předešlého postupu (viz kapitola 3.3).

Vyjádřeme rozptyl d_t:

$$\begin{aligned} \text{var}\, d_t &= \text{var}\,(p_t - p_{t-1}) = \text{var}\,(\hat{p}_t + r_t - \hat{p}_{t-1} - r_{t-1}) \\ &= \text{var}\,(\hat{p}_t - \hat{p}_{t-1}) + \text{var}\,(r_t - r_{t-1}), \end{aligned}$$

kde jsme využili nezávislosti $\hat{p}_t$ a $\hat{p}_{t-1}$ na r_t a r_{t-1}.

Položme rozptyl $\text{var}\,(\hat{p}_t - \hat{p}_{t-1})$ roven nule, čímž se dopustíme nejvýše zvýšení odhadu redukčního koeficientu. Právě na tomto místě se projevuje „konzervativnost" tohoto přístupu. Rozptyl totiž ve skutečnosti nejspíše nulový nebude a zhoršíme si tím odhad redukčního koeficientu. Přesto v některých případech můžeme dojít k pozitivním výsledkům. Využijme tedy nezávislosti r_t a r_{t-1} a pišme:

$$\begin{aligned} \text{var}\, d_t \approx\ \text{var}\,(r_t - r_{t-1}) =\ \text{var}\, r_t +\ \text{var}\, r_{t-1} = \\ = C\left[\frac{\hat{p}_t(1-\hat{p}_t)}{n_t} + \frac{\hat{p}_{t-1}(1-\hat{p}_{t-1})}{n_{t-1}}\right], \end{aligned}$$

kde jsme písmenem C opět označili redukční koeficient a n_t počet respondentů v čase *t*. Postupujme nyní obdobně jako v kapitole 3.3: Označme normovanou diferenci

$$\tilde{d}_t := \frac{d_t}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{n_t} + \frac{\hat{p}_{t-1}(1-\hat{p}_{t-1})}{n_{t-1}}}}.$$

Potom $\text{var}\,\tilde{d}_t = C$. I u tohoto postupu platí, že nižší odhad rozptylu získáme, pokud $C < 1$. Rozptyl $\text{var}\,\tilde{d}_t$ opět odhadneme výběrovým rozptylem normovaných diferencí:

$$\hat{C} = \widehat{\text{var}\,\tilde{d}_t} = \frac{\sum_{t=2}^{T}\left(\tilde{d}_t - \overline{\tilde{d}_t}\right)^2}{T-1}.$$

kde $\overline{\tilde{d}_t}$ značí průměr normovaných reziduí.

Po aproximaci dat jednotlivých stran polynomy téhož stupně a tvaru jako v odstavcích 3.3.1 a 3.3.2 a aplikaci nového postupu získáváme redukční koeficienty uvedené v Tabulce 11.

Tabulka 11: Odhady redukčních koeficientů získané metodou časových diferencí

	MEDIAN	STEM
ODS	1,245	1,227
ČSSD	0,945	1,890
KSČM	0,734	0,967
KDU-ČSL	1,166	0,987

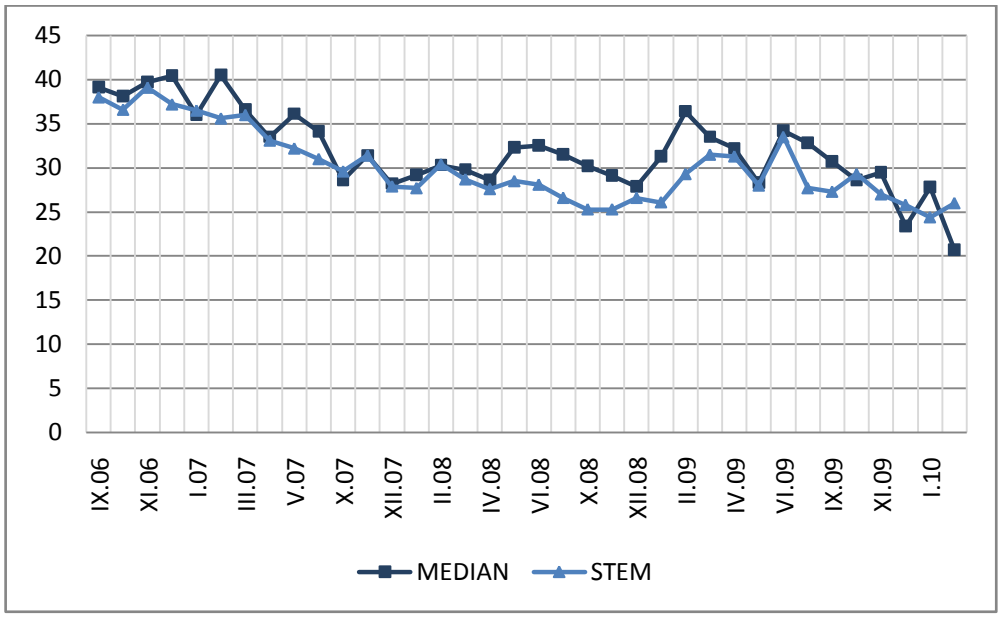
Porovnáme-li tyto odhady koeficientu $\hat{C}$ s hodnotami získanými předešlou metodou, zjistíme, že pouze ve dvou případech dostáváme lepší výsledek, a to v případě KSČM u agentury MEDIAN a v případě ODS u agentury STEM, což může být způsobeno tím, že zde mezi po sobě jdoucími obdobími nedochází k velkým výkyvům ve výši volebních preferencí, a tedy rozptyl $\mathrm{var}\,(\hat{p}_t - \hat{p}_{t-1})$, jež při výpočtech zanedbáváme, je zřejmě nule bližší, než je tomu u jiných stran. Tedy při zanedbání nedojde k zásadnímu zhoršení (zvýšení) odhadu redukčního koeficientu $\hat{C}$.

3.5 Porovnání výsledků dvou agentur

Jak již bylo řečeno, předchozí postupy mají kromě výhod i své nevýhody. V případě metody založené na vyrovnávání časové řady můžeme vývoj volebních preferencí vyrovnat příliš, nebo naopak málo. Při použití časových diferencí přicházíme o lepší výsledky kvůli zanedbání členu $\mathrm{var}\,(\hat{p}_t - \hat{p}_{t-1})$, jenž jistě nulový není. Proto na data zkusíme aplikovat ještě třetí postup, kterým bychom se obou problémům měli vyhnout. Využijeme toho, že pro přibližně stejná období máme data od dvou různých agentur (MEDIAN provádí terénní sběr dat v průběhu celého kalendářního měsíce, STEM během prvního týdnu tohoto měsíce), a tedy můžeme tato data pro každou ze čtyř stran zvlášť porovnávat.

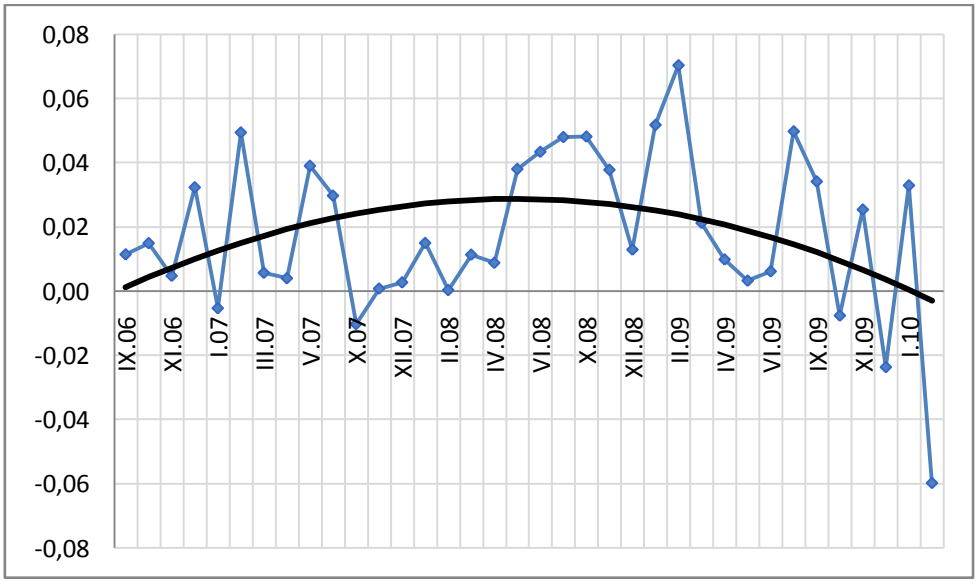
Základem této metody je předpoklad, že obě agentury by za použití totožné metodologie šetření měly dojít k volebním modelům, které jsou až na statistickou chybu stejné. Jak ale bylo řečeno na začátku této kapitoly, agentury nejenže své výsledky opírají o odpovědi různého počtu respondentů, navíc užívají i různých metod při sběru a zpracování dat. Není tedy divu, že agentury přibližně ve stejném čase prezentují rozdílné závěry, a to často i několik období po sobě, což můžeme pozorovat v Grafu 11, jenž zachycuje vývoj volebních preferencí strany ODS jednak podle MEDIANu, jednak podle STEMu.

Graf 11: Porovnání vývoje volebních preferencí ODS (v %) podle MEDIANu a STEMu



Tento problém jsme řešili následovně: Spočetli jsme pro danou stranu rozdíl mezi volebními preferencemi, jak je udává MEDIAN a jak je udává STEM, a to pro každý okamžik $t = 1, 2, \ldots, T$, zde $T = 36$, přičemž jsme vynechali data MEDIANu z těch měsíců, kdy nebyl prováděn průzkum agenturou STEM. Tyto rozdíly jsme v čase pro jednotlivé strany proložili vhodným polynomem, pro ODS viz Graf 12. K hodnotám p_t^S, čímž budeme značit volební preference dané strany v čase t zjištěné agenturou STEM, jsme pak přičetli odhady rozdílů v témže čase získané proložením tohoto polynomu. Tyto nové hodnoty, označme je p_t^{S2}, pak budeme používat namísto původních p_t^S.

Graf 12: Proložení rozdílů polynomem 2. stupně – ODS



Jinak zapsáno: rozdíly spočteme takto $r_t = p_t^M - p_t^S$. Proložením polynomu získáme odhady rozdílů r_t, které budeme značit $\hat{r}_t$. Data agentury STEM pak upravíme následovně: $p_t^{S2} = p_t^S + \hat{r}_t$. Pokud bychom při výpočtu r_t zaměnili p_t^M a p_t^S a poté upravovali data agentury MEDIAN namísto STEMu, dostali bychom při následujícím postupu totožné odhady $\hat{C}$ jako při použití těch r_t a p_t^{S2}, které jsme definovali.

Nyní můžeme přejít k popisu vlastní metody. Předpokládejme tedy, že se volební preference dané politické strany shodují u obou agentur až na náhodnou chybu, tj. $p_t^{S2} = p_t + e_t^{S2}$ a $p_t^M = p_t + e_t^M$, $t = 1, 2, \ldots, 36$, kde p_t^{S2} jsou upravená data pocházející od agentury STEM, p_t^M značíme data pocházející od agentury MEDIAN, p_t je skutečná hodnota volebních preferencí v čase t a e_t^{S2}, e_t^M jsou nezávislé náhodné chyby.

Označme $d_t := p_t^M - p_t^{S2} = p_t + e_t^M - (p_t + e_t^{S2}) = e_t^M - e_t^{S2}$. Dále použijeme odhady volebních preferencí z první použité metody (viz kapitola 3.3) – označme je $\hat{p}_t^S$ pro odhad dat pocházejících z průzkumu STEMu pomocí polynomu, $\hat{p}_t^M$ nechť je odhad pro data prezentovaná MEDIANem. Počítejme:

$$\operatorname{var} d_t = \operatorname{var} e_t^M + \operatorname{var} e_t^S = C\left(\frac{\hat{p}_t^M(1-\hat{p}_t^M)}{n_t^M} + \frac{\hat{p}_t^S(1-\hat{p}_t^S)}{n_t^S}\right), \tag{8}$$

kde n_t^M, n_t^S jsou počty respondentů účastnících se průzkumu MEDIANu, resp. STEMu v čase $t = 1, 2, \ldots, 36$. Poznamenejme, že není úplné jasné, zda bychom ve vzorci (8) měli používat hodnoty $\hat{p}_t^S$, nebo hodnoty $\hat{p}_t^{S2}$, čímž bychom značili odhad p_t^{S2} v čase t získaný proložením vhodného polynomu. My jsme zde zvolili první variantu. Nyní označme normovanou chybu

$$\tilde{d}_t := \frac{d_t}{\sqrt{\frac{\hat{p}_t^M(1-\hat{p}_t^M)}{n_t^M} + \frac{\hat{p}_t^S(1-\hat{p}_t^S)}{n_t^S}}}.$$

Potom $C = \operatorname{var} \tilde{d}_t$ a $\operatorname{var} \tilde{d}_t$ opět odhadneme výběrovým rozptylem normovaných chyb (kde $\overline{\tilde{d}_t} \doteq 0$):

$$\hat{C} = \widehat{\operatorname{var}\, \tilde{d}_t} = \frac{\sum_{t=1}^{T} \tilde{d}_t^2}{T}.$$

Nyní můžeme přejít k vlastním výpočtům pro jednotlivé strany. V Tabulce P4 uvádíme hodnoty p_t^{S2} volebních preferencí zjištěných agenturou STEM pro jednotlivé strany upravené pomocí odhadu polynomem – najdeme ji v Přílohách, neboť jde opět o tabulku větší velikosti. Rozdíly byly v případě stran ODS a ČSSD aproximovány kvadratickým polynomem, v případě KDU-ČSL a KSČM přímkou. Konkrétně byly použity polynomy tohoto tvaru (stupně polynomů jsme volili podle obdobných kritérií jako v kapitole 3.3):

ODS: $\hat{r}_t = -0{,}00232 + 0{,}00346t - 10 \cdot 10^{-5} t^2$,
ČSSD: $\hat{r}_t = 0{,}045020 - 0{,}00487t + 9 \cdot 10^{-5} t^2$,
KSČM: $\hat{r}_t = 0{,}00082 + 6 \cdot 10^{-5} t$,
KDU-ČSL: $\hat{r}_t = -0{,}015 + 5 \cdot 10^{-4} t$, $t = 1, 2, \ldots, 36$.

Ještě jednou připomeňme, že pro odhad volebních preferencí jednotlivých stran u obou agentur používáme polynomy uvedené u první metody (viz kapitola 3.3).

Sledováním výše uvedeného postupu dojdeme k odhadům redukčního koeficientu, které jsou uvedeny v Tabulce 12. U stran ODS a KSČM došlo ke snížení odhadu statistické chyby. U ČSSD a KDU-ČSL dochází naopak ke zvýšení odhadu statistické chyby, i když ne výraznému.

Tabulka 12: Odhady redukčních koeficientů pro jednotlivé strany získané na základě porovnání dat obou agentur.

ODS	0,962
ČSSD	1,113
KSČM	0,853
KDU-ČSL	1,048

3.6 Shrnutí

V Tabulce 13 uvádíme pro shrnutí odhady redukčních koeficientů pro jednotlivé strany a agentury tak, jak jsme je vypočítali použitím uvedených metod (soubory s podrobnými výpočty lze nalézt na přiloženém CD). Vidíme, že velmi často alespoň jedním ze tří postupů dojedeme k takovému odhadu statistické chyby, jenž je nižší než odhad vypočtený klasickým vzorcem. Pro stranu KSČM dokonce dostáváme ve všech případech redukční koeficient menší než jedna. Je tomu tak nejspíše z důvodů, jimiž jsme se zabývali v odstavci 3.3.1.

Naopak u větších stran (ODS, ČSSD) dostáváme mnohdy vyšší odhady redukčního koeficientu, což může být způsobeno např. tím, že tyto strany nemají tak stálé jádro voličů, jako třeba KSČM. Navíc se jedná o vládní strany, u kterých veřejnost více

reaguje na skandály, což se projeví na vývoji volebních preferencí většími a častějšími výkyvy.

Tabulka 13: Redukční koeficienty pro jednotlivé strany, agentury a metody

	MEDIAN		STEM		MEDIAN a STEM
	Vyrovnávání	Časové diference	Vyrovnávání	Časové diference	Porovnání obou agentur
ODS	1,006	1,245	1,278	1,227	0,962
ČSSD	0,864	0,945	1,615	1,890	1,113
KSČM	0,781	0,734	0,925	0,967	0,853
KDU-ČSL	0,870	1,166	0,877	0,987	1,048

Pro srovnání uvádíme v Přílohách Tabulku P5, jež obsahuje nové odhady statistické chyby pro data agentur MEDIAN a STEM, které jsou spočteny pomocí redukčních koeficientů a kvantilu normálního rozdělení $u(0{,}025) = 1{,}96$. Přitom pro každou stranu a agenturu byl zvolen nejmenší redukční koeficient, jaký se nám podařilo získat aplikací všech třech postupů.

Jelikož agentura STEM zakládá své průzkumy na kvótních výběrech a téměř jistě agentura MEDIAN používá při výběrových šetřeních vážení nebo kvóty, můžeme se vrátit ke vztahům, jež jsme odvodili v druhé kapitole, např. (4). Tam jsme dokázali původní odhad rozptylu redukovat $(1 - R^2)$krát. V kapitole 3 jsme odhad rozptylu snížili C-násobně. Bylo by tedy možné dále zkoumat, zdali i zde neplatí vztah $1 - R^2 = C$. Touto záležitostí se však zde již zabývat nebudeme, neboť nemáme k dispozici potřebná data (struktura výběrového souboru, kvótní a vážící proměnné apod.).

Agentury věnující se nejen průzkumům veřejného mínění, ale i např. průzkumy trhu, by pak (namísto sledování postupů uvedených v této kapitole) mohli ze svých dat spočítat koeficient determinace pro dané výběrové šetření např. tak, jak jsme uvedli v druhé kapitole, a pomocí tohoto koeficientu vyjádřit nižší odhad statistické chyby.

Závěr

V této práci jsme se snažili poukazovat na důležitost prezentování statistické chyby společně s výsledky výběrových šetření, a také na důležitost jejího vnímání. Statistická chyba vnáší do výsledků šetření určitou míru nejistoty ohledně správnosti zobecnění těchto výsledků z výběrového souboru na základní soubor.

Podívejme se na tuto záležitost z jiného pohledu. Pokud věnujeme pozornost prezentované statistické chybě, pak tato chyba svým způsobem zpřesňuje náš přehled o situaci v základním souboru. Víme například, že volební preference nějaké politické strany nemusí být přesně 15%, ale že se může lišit až o hodnotu odhadu statistické chyby, tedy např. ± 3%. Tento fakt je důležitý nejen pro vlastní politické strany, ale třeba i pro sázkové kanceláře, jak jsme již zmínili ve třetí kapitole.

Ve druhé kapitole jsme u kvótních výběrů při dotazování na otázku s možnou odpovědí ano, či ne dokázali, že odhad statistické chyby při odhadování poměru kladných odpovědí v populaci výběrovým průměrem lze snížit přímo úměrně hodnotě $1 - R^2$, kde R^2 je koeficient determinace takového vhodného modelu, kterým na základě kvótní proměnné vysvětlujeme 0-1 proměnnou Y, jež reprezentuje odpovědi respondentů. Tím tedy dále zpřesňujeme odhad statistické chyby pro poměr kladných odpovědí.

Také jsme pomocí simulací ukázali, že podobný vztah by mohl platit i v případě, že se po terénním sběru dat přistoupí k vážení dat. Vzhledem k tomu, že kvótní výběry a vážení dat (případně jejich kombinaci) agentury zabývající se výběrovými šetřeními často používají, je možné využít odvozených vztahů i v praxi.

Ve třetí kapitole jsme empiricky odhadovali statistickou chybu přímo pro reálná data, konkrétně pro volební modely. Zde se ne vždy podařilo dosáhnout optimálních výsledků, redukční koeficient byl někdy větší než jedna. Nicméně v některých případech jsme docílili značného snížení odhadu statistické chyby. Nezapomeňme ještě, že jsme v této kapitole pouze odhadovali hodnotu redukčního koeficientu C pomocí více či méně vhodných postupů. Není tedy vyloučeno, že i v případě $\hat{C} > 1$ může ve skutečnosti vztah $C = 1 - R^2$ platit. Způsoby empirického odhadování statistické chyby z této kapitoly přitom nejsou použitelné pouze pro průzkumy volebních preferencí. Lze je aplikovat i na jiná výběrová šetření prováděná kontinuálně.

Ukázali jsme tedy, že existují postupy použitelné v praxi, jimiž lze dosáhnout snížení odhadu statistické chyby u výběrových šetření a tím lépe odhadnout nejistotu ohledně zobecnění výsledků na celou populaci. Přitom snížení (redukce) $C = 1 - R^2$ je tím výraznější, čím je vyšší hodnota koeficientu determinace R^2, tj. čím vyšší je korelace mezi dotazovanou proměnnou a kvótní, resp. vážící proměnnou (proměnnými). Pokud bychom tedy např. vážili data volebního modelu podle výsledků předchozích voleb (a na základě otázky, koho respondenti v těchto volbách volili), lze díky poměrně

vysoké korelaci současné a předchozí volby očekávat značnou redukci odhadu statistické chyby.

Literatura

[1] Anděl, J.: *Základy matematické statistiky*, MATFYZPRESS, 2007.

[2] Anděl, M., Černý, R., Charamza, P., Neustadt, J.: *Přehled metod odhadu statistické chyby ve výběrových šetřeních*, 4. 10. 2009, http://www.quantitative.cz/cz/dokumenty-2/odborne-statisticke-materialy

[3] Deming, W. E., Stephan F. F.: *On a Least Squares Adjustment od Sampled Frequency Table When the Expected Marginal Totals are known*, The Annals of Mathematical Statistics, Vol. 11, No. 40, 1940

[4] Jungová, E.: *Historie výzkumů veřejného mínění*, 8. 10. 2009, http://www.richardjung.cz/index.asp?menu=628

[5] Jungová, E.: *Metody a techniky výzkumu veřejného mínění*, 8. 10. 2009, http://www.richardjung.cz/index.asp?menu=628

[6] Krejčí, J.: *Limity volebních předpovědí*, SDA Info, No. 2, 2004

[7] Lebeda, T., Leontiyeva, Y., Krejčí J.: *Volební preference, jak jim správně porozumět*, 22. 5. 2010 http://www.cvvm.cas.cz/upl/nase_spolecnost/100049s_lebeda-vyzkumy.pdf,

[8] MEDIAN, s. r. o.: *Tisková zpráva VOLEBNÍ PREFERENCE*, 2006 – 2010, 30. 4. 2010, http://www.median.cz

[9] Rabušic, L., Soukup P.: *Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti*, Sociologicky časopis/Czech Sociological Review, Vol. 43, No. 2, 2007

[10] STEM, s. r. o.: *Stranické preference*, 2006 – 2010, 30. 4. 2010, http://www.stem.cz

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

[12] Zvára, K.: *Základy biostatistiky*, 30. 4. 2010, http://www.karlin.mff.cuni.cz/_zvara

[13] Zvárová, J.: *Základy statistiky pro biomedicínské obory*, 30. 4. 2010, http://new.euromise.org/czech/tajne/ucebnice/html/html/statist.html

Přílohy

Tabulka P1: Vývoj volebních preferencí (v %) podle agentury MEDIAN, s. r. o.

	ODS	ČSSD	KSČM	SZ	KDU-ČSL	TOP 09	VV	Ostatní	SE	# Resp.
VII.06	41,8	29,8	11,4	6,2	5,2	-	-	5,6	-	431
VIII.06	37,5	29,8	13,2	7,7	5,4	-	-	6,3	-	533
IX.06	39,1	29,4	13,0	7,8	6,1	-	-	4,7	-	804
X.06	38,1	28,9	13,0	7,7	5,8	-	-	6,5	-	748
XI.06	39,7	26,5	13,7	9,6	5,6	-	-	5,0	2,9	578
XII.06	40,4	28,2	11,4	8,2	6,7	-	-	5,2	2,4-4,7	369
I.07	36,0	28,8	13,9	6,6	6,0	-	-	8,7	1,6-3,2	448
II.07	40,5	25,4	12,6	8,2	7,4	-	-	5,9	1,6-3,1	508
III.07	36,6	30,3	12,6	7,9	5,6	-	-	7,1	2,4	763
IV.07	33,5	32,3	13,2	6,9	8,1	-	-	6,1	2,0-4,0	494
V.07	36,1	31,1	15,2	6,4	4,7	-	-	6,4	1,6-3,8	568
VI.07	33,4	34,8	12,7	6,9	5,2	-	-	7,0	1,5-3,5	657
VII.07	33,7	33,3	13,7	6,6	6,2	-	-	6,6	1,5-3,5	575
VIII.07	35,2	31,7	12,6	6,6	6,1	-	-	7,8	1,5-3,5	613
IX.07	34,1	30,4	12,3	6,5	7,3	-	-	9,4	1,5-3,5	655
X.07	28,6	36,1	15,1	6,4	6,0	-	-	7,9	1,5-3,5	682
XI.07	31,4	35,3	14,2	4,4	7,6	-	-	7,0	1,5-3,5	611
XII.07	28,2	37,0	16,6	5,9	6,9	-	-	5,4	1,8-4	524
I.08	29,2	35,7	12,6	8,0	5,9	-	-	8,5	1,8-4	570
II.08	30,3	33,9	14,9	8,8	6,8	-	-	5,4	2,0-4,0	556
III.08	29,8	36,9	13,2	8,0	6,1	-	-	5,9	2,0-4,0	552
IV.08	28,6	38,9	13,9	4,8	5,5	-	-	8,2	1,8-3,5	617
V.08	32,3	36,5	12,2	4,9	7,0	-	-	7,1	1,5-3,5	626
VI.08	32,5	35,5	15,1	4,5	5,8	-	-	6,7	1,5-3,5	563
VII.08	29,4	37,7	14,6	7,1	6,0	-	-	5,3	1,5-3,5	559
VIII.08	31,6	36,8	15,4	4,0	7,3	-	-	4,9	1,5-3,5	563
IX.08	31,5	34,8	13,1	4,7	6,4	-	-	9,5	2,0-4,0	597
X.08	30,2	38,1	13,8	5,1	5,8	-	-	7,0	2,0-4,0	672
XI.08	29,1	42,7	12,1	4,3	5,9	-	-	5,9	2,0-4,0	682
XII.08	27,9	40,5	14,0	2,7	6,7	-	-	8,1	2,0-4,0	601
I.09	31,3	35,3	14,2	5,5	7,0	-	-	6,6	2,0-4,0	580
II.09	36,4	38,0	11,6	3,5	5,5	-	-	5,0	2,0-4,0	619
III.09	33,5	36,4	10,6	3,9	8,6	-	-	6,9	2,0-4,0	691
IV.09	32,2	35,7	13,0	3,4	5,8	-	-	10,0	2,0-4,0	568
V.09	28,3	35,6	12,7	4,7	6,2	-	-	12,4	2,0-4,0	658
VI.09	34,2	33,0	12,1	2,6	6,5	3,1	-	8,5	2,0-4,0	567
VII.09	31,4	34,8	11,9	2,4	7,8	6,5	-	5,3	2,0-4,0	596
VIII.09	32,8	34,4	13,9	2,4	3,5	8,3	-	4,8	2,0-4,0	570
IX.09	30,7	27,6	12,5	3,1	6,8	13,2	-	6,3	2,0-4,0	603
X.09	28,6	29,6	12,6	2,9	5,7	12,2	2,5	5,9	2,0-4,0	727
XI.09	29,5	27,1	14,2	4,4	5,6	9,0	3,1	7,3	2,0-4,0	681
XII.09	23,4	29,3	15,9	4,0	6,9	9,9	2,3	8,2	2,0-4,0	492
I.10	27,8	32,0	12,8	4,3	7,4	9,3	2,4	4,0	2,0-4,5	423
II.10	20,7	34,5	13,9	4,8	5,5	9,6	5,2	5,7	1,5-3,5	526

Tabulka P2: Vývoj volebních preferencí (v %) podle agentury STEM, s. r. o.

	ODS	ČSSD	KSČM	SZ	KDU-ČSL	TOP 09	VV	Ostatní	# Resp.
VII.06	-	-	-	-	-	-	-	-	-
VIII.06	-	-	-	-	-	-	-	-	-
IX.06	38,0	30,8	11,6	11,9	4,9	-	-	2,9	1336
X.06	36,6	26,0	13,4	12,0	7,2	-	-	4,8	1398
XI.06	39,1	23,1	12,2	10,5	8,4	-	-	6,8	1394
XII.06	37,2	23,3	14,8	12,7	7,9	-	-	4,2	1394
I.07	36,5	23,6	14,0	12,6	8,5	-	-	4,9	1118
II.07	35,6	26,6	14,4	11,7	8,6	-	-	3,0	1124
III.07	36,0	26,7	12,3	12,5	7,9	-	-	4,6	1075
IV.07	33,1	28,2	12,4	13,2	7,5	-	-	5,7	1050
V.07	32,2	31,1	14,2	10,4	7,4	-	-	4,7	1082
VI.07	-	-	-	-	-	-	-	-	-
VII.07	-	-	-	-	-	-	-	-	-
VIII.07	-	-	-	-	-	-	-	-	-
IX.07	31,0	33,0	14,0	10,5	8,4	-	-	3,1	1233
X.07	29,6	34,9	12,9	10,1	7,6	-	-	4,8	1119
XI.07	31,4	30,7	13,8	10,7	6,5	-	-	6,7	1121
XII.07	27,9	38,0	13,5	9,9	7,1	-	-	3,5	1163
I.08	27,7	35,3	14,9	9,4	6,8	-	-	5,9	1194
II.08	30,4	35,4	12,1	10,7	8,0	-	-	3,4	1251
III.08	28,7	35,9	12,7	10,5	6,3	-	-	5,8	1114
IV.08	27,6	36,8	13,3	10,9	6,5	-	-	4,8	1136
V.08	28,5	34,8	13,0	12,2	6,0	-	-	5,5	1059
VI.08	28,1	39,4	13,8	9,5	5,3	-	-	4,1	1092
VII.08	-	-	-	-	-	-	-	-	-
VIII.08	-	-	-	-	-	-	-	-	-
IX.08	26,6	38,9	14,1	10,0	7,0	-	-	3,3	1135
X.08	25,3	39,4	14,2	8,4	7,2	-	-	5,6	1053
XI.08	25,3	45,3	11,5	6,7	6,7	-	-	4,4	1062
XII.08	26,6	43,5	12,9	6,1	7,1	-	-	3,7	1132
I.09	26,1	43,8	11,5	6,7	7,5	-	-	4,3	1083
II.09	29,3	40,4	13,5	5,9	7,8	-	-	3,1	1100
III.09	31,5	40,0	12,5	6,0	6,8	-	-	3,3	1127
IV.09	31,3	40,0	12,0	6,4	5,0	-	-	5,5	1139
V.09	28,0	38,3	12,9	6,3	7,0	-	-	7,7	1071
VI.09	33,5	33,4	15,1	3,2	7,3	-	-	7,5	1083
VII.09	-	-	-	-	-	-	-	-	-
VIII.09	27,7	33,4	12,9	3,1	6,6	7,8	-	8,5	1077
IX.09	27,3	32,2	12,5	5,4	4,9	10,8	-	6,8	1068
X.09	29,3	30,4	13,5	3,7	4,8	10,1	2,6	5,8	1093
XI.09	27,0	29,4	14,2	4,1	5,9	9,7	2,1	7,6	1104
XII.09	25,8	32,9	12,5	4,4	6,5	9,5	3,3	5,2	1100
I.10	24,4	33,5	12,7	4,7	4,9	9,8	4,6	5,4	1110
II.10	26,0	32,0	13,1	3,0	4,8	10,2	5,3	5,6	1110

Tabulka P3: Odhad statistické chyby klasickým vzorcem (v %):

	MEDIAN				STEM			
	ODS	ČSSD	KSČM	KDU-ČSL	ODS	ČSSD	KSČM	KDU-ČSL
VII.06	4,657	4,318	3,000	2,096	-	-	-	-
VIII.06	4,110	3,883	2,874	1,919	-	-	-	-
IX.06	3,373	3,149	2,325	1,654	2,602	2,475	1,715	1,153
X.06	3,480	3,249	2,410	1,675	2,525	2,299	1,785	1,356
XI.06	3,989	3,598	2,803	1,874	2,562	2,212	1,715	1,453
XII.06	5,007	4,591	3,243	2,551	2,833	2,477	2,082	1,582
I.07	4,445	4,193	3,204	2,199	2,815	2,484	2,030	1,629
II.07	4,269	3,785	2,886	2,276	2,862	2,641	2,101	1,679
III.07	3,418	3,261	2,355	1,631	2,799	2,579	1,913	1,571
IV.07	4,162	4,124	2,985	2,406	2,846	2,723	1,991	1,592
V.07	3,950	3,807	2,953	1,741	2,784	2,758	2,079	1,563
VI.07	3,606	3,642	2,546	1,698	-	-	-	-
VII.07	3,864	3,852	2,811	1,971	-	-	-	-
VIII.07	3,781	3,684	2,627	1,895	-	-	-	-
IX.07	3,630	3,523	2,515	1,992	2,582	2,624	1,938	1,548
X.07	3,392	3,605	2,687	1,782	2,676	2,793	1,966	1,548
XI.07	3,680	3,789	2,768	2,101	2,718	2,701	2,021	1,445
XII.07	3,853	4,134	3,186	2,170	2,579	2,789	1,961	1,473
I.08	3,733	3,933	2,724	1,934	2,539	2,712	2,022	1,432
II.08	3,820	3,935	2,960	2,093	2,548	2,650	1,805	1,500
III.08	3,816	4,025	2,824	1,997	2,656	2,817	1,954	1,425
IV.08	3,566	3,847	2,730	1,799	2,600	2,805	1,977	1,433
V.08	3,663	3,771	2,564	1,999	2,719	2,869	2,025	1,428
VI.08	3,869	3,953	2,958	1,931	2,665	2,898	2,046	1,324
VII.08	3,777	4,018	2,927	1,969	-	-	-	-
VIII.08	3,840	3,984	2,982	2,149	-	-	-	-
IX.08	3,726	3,821	2,707	1,963	2,571	2,836	2,024	1,488
X.08	3,471	3,672	2,608	1,767	2,626	2,951	2,107	1,562
XI.08	3,409	3,712	2,448	1,768	2,616	2,994	1,921	1,506
XII.08	3,586	3,925	2,774	1,999	2,575	2,888	1,953	1,500
I.09	3,774	3,889	2,841	2,077	2,617	2,955	1,900	1,573
II.09	3,790	3,824	2,523	1,796	2,689	2,900	2,022	1,585
III.09	3,519	3,588	2,295	2,090	2,712	2,860	1,928	1,469
IV.09	3,843	3,940	2,766	1,922	2,694	2,845	1,884	1,267
V.09	3,442	3,659	2,544	1,843	2,689	2,911	2,005	1,525
VI.09	3,905	3,870	2,684	2,029	2,811	2,808	2,135	1,553
VII.09	3,726	3,824	2,600	2,153	-	-	-	-
VIII.09	3,854	3,900	2,840	1,509	2,674	2,816	2,001	1,479
IX.09	3,682	3,568	2,640	2,009	2,672	2,803	1,981	1,300
X.09	3,285	3,318	2,412	1,685	2,697	2,728	2,024	1,263
XI.09	3,425	3,338	2,622	1,727	2,618	2,687	2,061	1,390
XII.09	3,741	4,022	3,231	2,240	2,585	2,776	1,953	1,460
I.10	4,269	4,445	3,184	2,495	2,527	2,777	1,961	1,271
II.10	3,462	4,063	2,956	1,948	2,580	2,745	1,985	1,259

Tabulka P4: Upravené hodnoty volebních preferencí – STEM (v %)

	ODS	ČSSD	KSČM	KDU-ČSL
IX.06	38,07	34,81	11,66	6,32
X.06	37,03	29,54	13,48	8,61
XI.06	39,85	26,20	12,25	9,71
XII.06	38,17	25,97	14,91	9,20
I.07	37,80	25,93	14,14	9,72
II.07	37,06	28,49	14,55	9,82
III.07	37,75	28,22	12,40	9,01
IV.07	35,02	29,41	12,49	8,56
V.07	34,31	31,92	14,32	8,45
IX.07	33,30	33,51	14,16	9,35
X.07	32,04	35,12	13,08	8,46
XI.07	33,96	30,68	13,98	7,37
XII.07	30,57	37,64	13,61	7,87
I.08	30,43	34,77	15,10	7,59
II.08	33,16	34,57	12,23	8,65
III.08	31,50	34,88	12,85	6,91
IV.08	30,48	35,62	13,52	7,07
V.08	31,37	33,44	13,18	6,50
VI.08	30,92	37,86	14,00	5,73
IX.08	29,43	37,21	14,28	7,46
X.08	28,07	37,55	14,39	7,57
XI.08	28,03	43,34	11,74	7,03
XII.08	29,23	41,48	13,13	7,38
I.09	28,64	41,71	11,72	7,74
II.09	31,65	38,28	13,78	7,95
III.09	33,71	37,81	12,69	6,88
IV.09	33,38	37,79	12,20	5,04
V.09	29,86	36,07	13,12	6,94
VI.09	35,16	31,21	15,40	7,26
VIII.09	29,18	31,24	13,15	6,42
IX.09	28,50	30,16	12,74	4,75
X.09	30,21	28,42	13,74	4,52
XI.09	27,62	27,48	14,51	5,60
XII.09	26,13	31,06	12,77	6,17
I.10	24,45	31,84	13,02	4,50
II.10	25,67	30,48	13,40	4,35

Tabulka P5: Odhad statistické chyby (v %) za použití redukčního koeficientu

	MEDIAN				STEM			
	ODS	ČSSD	KSČM	KDU-ČSL	ODS	ČSSD	KSČM	KDU-ČSL
C	0,956	0,864	0,666	0,816	0,956	1,129	0,666	0,877
VII.06	4,553	4,014	2,448	1,894	-	-	-	-
VIII.06	4,018	3,609	2,345	1,734	-	-	-	-
IX.06	3,298	2,927	1,897	1,495	2,544	2,630	1,400	1,080
X.06	3,403	3,020	1,966	1,513	2,469	2,442	1,456	1,269
XI.06	3,900	3,344	2,287	1,694	2,505	2,350	1,399	1,360
XII.06	4,895	4,268	2,646	2,305	2,770	2,632	1,698	1,481
I.07	4,346	3,898	2,614	1,987	2,752	2,639	1,656	1,525
II.07	4,174	3,519	2,355	2,057	2,798	2,806	1,714	1,572
III.07	3,342	3,031	1,921	1,474	2,737	2,740	1,561	1,471
IV.07	4,069	3,833	2,435	2,174	2,783	2,892	1,624	1,490
V.07	3,862	3,539	2,409	1,573	2,722	2,930	1,696	1,463
VI.07	3,526	3,386	2,077	1,534	-	-	-	-
VII.07	3,778	3,581	2,293	1,781	-	-	-	-
VIII.07	3,697	3,424	2,143	1,712	-	-	-	-
IX.07	3,549	3,274	2,052	1,800	2,525	2,788	1,581	1,449
X.07	3,316	3,351	2,193	1,610	2,616	2,967	1,604	1,449
XI.07	3,598	3,522	2,258	1,898	2,657	2,870	1,649	1,353
XII.07	3,767	3,843	2,599	1,961	2,521	2,963	1,600	1,379
I.08	3,650	3,656	2,223	1,748	2,482	2,881	1,650	1,341
II.08	3,735	3,657	2,415	1,891	2,492	2,815	1,472	1,405
III.08	3,731	3,742	2,304	1,804	2,597	2,993	1,594	1,334
IV.08	3,486	3,576	2,227	1,625	2,542	2,980	1,613	1,342
V.08	3,582	3,506	2,092	1,806	2,658	3,048	1,652	1,337
VI.08	3,783	3,674	2,413	1,745	2,606	3,079	1,669	1,240
VII.08	3,693	3,734	2,388	1,779	-	-	-	-
VIII.08	3,755	3,703	2,433	1,941	-	-	-	-
IX.08	3,643	3,552	2,208	1,774	2,514	3,013	1,651	1,393
X.08	3,394	3,413	2,128	1,597	2,567	3,135	1,719	1,463
XI.08	3,333	3,451	1,997	1,598	2,557	3,180	1,567	1,410
XII.08	3,506	3,648	2,263	1,806	2,517	3,068	1,594	1,404
I.09	3,690	3,615	2,318	1,876	2,558	3,139	1,550	1,473
II.09	3,706	3,554	2,058	1,623	2,629	3,081	1,650	1,484
III.09	3,441	3,335	1,873	1,889	2,651	3,038	1,573	1,376
IV.09	3,757	3,663	2,257	1,737	2,634	3,022	1,538	1,186
V.09	3,365	3,401	2,076	1,665	2,629	3,092	1,636	1,427
VI.09	3,818	3,598	2,190	1,833	2,748	2,984	1,742	1,454
VII.09	3,643	3,555	2,121	1,945	-	-	-	-
VIII.09	3,768	3,625	2,317	1,363	-	-	-	-
IX.09	3,600	3,316	2,154	1,815	2,612	2,978	1,617	1,217
X.09	3,212	3,084	1,968	1,523	2,637	2,898	1,652	1,182
XI.09	3,349	3,103	2,139	1,560	2,560	2,855	1,682	1,302
XII.09	3,658	3,738	2,636	2,024	2,527	2,949	1,594	1,367
I.10	4,174	4,132	2,598	2,254	2,471	2,951	1,600	1,190
II.10	3,385	3,776	2,412	1,760	2,522	2,916	1,620	1,179